



Management for Professionals

Sara Dolnicar

Bettina Grün

Friedrich Leisch

Market Segmentation Analysis

Understanding It, Doing It,
and Making It Useful

Management for Professionals

More information about this series at <http://www.springer.com/series/10101>

Sara Dolnicar • Bettina Grün • Friedrich Leisch

Market Segmentation Analysis

Understanding It, Doing It, and Making
It Useful

Published with the support of the Austrian Science Fund (FWF): PUB 580-Z27



Sara Dolnicar
The University of Queensland
Brisbane, Queensland, Australia

Bettina Grün
Johannes Kepler Universität Linz
Linz, Oberösterreich, Austria

Friedrich Leisch
Universität für Bodenkultur Wien
Vienna, Wien, Austria



ISSN 2192-8096
Management for Professionals
ISBN 978-981-10-8817-9
<https://doi.org/10.1007/978-981-10-8818-6>

ISSN 2192-810X (electronic)
ISBN 978-981-10-8818-6 (eBook)

Library of Congress Control Number: 2018936527

© The Editor(s) (if applicable) and The Author(s) 2018. This book is an open access publication.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore

Preface

'Another book on market segmentation' you think. Many outstanding marketing scientists, scholars and consultants have written excellent books on market segmentation. Some books offer practical advice to managers on how to best implement market segmentation in an organisation to ensure that the segmentation strategy is a success. Other books present sophisticated algorithms to extract market segments from consumer data. Our excuse for writing yet another book on market segmentation is to bridge the gap between the managerial and the statistical aspects of market segmentation analysis. We also want to give readers the opportunity to replicate every single calculation and visualisation we discuss in the book. We achieve this by making data sets used in the book available online (<http://www.MarketSegmentationAnalysis.org>) and by accompanying each section with R code. R is an open source environment for statistical computing and graphics, which is freely available for Linux, MacOS and Windows.

Most of the examples used in the book relate to tourism. We have chosen tourism because most people go on vacation and, as a consequence, can relate to the examples, even if professionally they market an entirely different product. Tourism is also very complex compared to other products: a trip consists of many different elements, typically a number of decision makers are involved in the planning process, travel can be motivated by a wide range of motives, and manifests in tourists engaging in an even wider range of activities. Tourists can plan their trip of a lifetime for decades or 'impulse purchase' a city trip a few hours before departure. As a consequence of the complexity of tourism as a product, many alternative market segmentation approaches can be used to break the market down into smaller, more homogeneous consumer groups or market segments. In the case of marketing toothpaste, for example, consumers can be segmented by their willingness to pay or by benefits sought. Tourists can, in addition, be grouped based on their preferences for vacation activities, the people they travel with, how long they travel, whether or not they stay at the same destination or visit a number of destinations, the degree to which they perceive risks to be associated with their trip, their expenditure patterns, their level of variety seeking and so on.

The fact that we use many tourism examples does not mean, however, that this is a book on tourism market segmentation. Market segmentation is a framework that is independent of the nature of the product or service being marketed. Everything we discuss in the book can be used in tourism, but also to market fast moving consumer goods or to try to attract excellent foster carers. The principles and techniques covered in this book can be applied across a variety of industries and geographic markets. This is also reflected by our use of the terms *organisation* and *user* to signal that market segmentation is of value to organisations aimed primarily at generating profits, as well as organisations aimed at achieving other missions.

We have structured the book in a way that makes it possible to use it as a companion throughout the entire journey of market segmentation analysis. In this case, each of the steps can be processed one after the other. Alternatively, it is also possible to just learn more about one specific step of market segmentation analysis. We have broken down the process of market segmentation analysis into ten steps. For each step we discuss the aims, point to potential pitfalls, and offer a range of approaches that can be used. All proposed approaches are accompanied by R code allowing replication of all analyses.

R started in 1992. Over the last two decades, R has developed to become the lingua franca of computational statistics (de Leeuw and Mair 2007, p. 2). It is used for teaching and research in universities all over the world and has been adopted by many non-academic organisations. R is open source software. The source code can be downloaded from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org> for free. The backbone of R's success is that everybody can contribute extension packages. In April 2018 some 12,500 extension packages were available on CRAN. Many more R packages are available on private web pages and in other repositories. Many of these packages can be used for market segmentation, and some will be introduced in this book.

One of the extension packages is called **MSA** (for Market Segmentation Analysis) and contains all data sets used in this book. The package also contains all analyses shown in the book as R demonstrations that can be run directly using commands like `demo ("step-4", package = "MSA")` to run the code from Step 4. For users of other statistical software packages, the data sets are also available at <http://www.MarketSegmentationAnalysis.org>.

This book is not an introduction to R. Readers who are not familiar with R can

1. Ignore the R commands shown throughout the book, and concentrate on the results. Many of the algorithms presented are available in other statistical software packages like SPSS or SAS.
2. Use a companion R package written by Putler and Krider (2012) called **RcmdrPlugin.BCA** that implements a graphical user interface (GUI) through the R Commander (Fox 2017) point-and-click interface to R. Starting with a GUI makes it easier to learn R initially.

3. Learn R from an introductory R textbook. Dalgaard (2008), Hothorn and Everitt (2014) and Kabacoff (2015) offer general introductions to R; Chapman and Feit (2015) and Putler and Krider (2012) discuss marketing and business-related analyses more specifically.

At the end of each of the ten steps of market segmentation analysis, we offer a checklist. These checklists are a starting point for organisations to structure their market segmentation analysis procedure. They can easily be modified, refined and extended to best suit the organisation's needs.

At a practical level, this book is the result of two decades of cross-disciplinary research into market segmentation facilitated by the research agencies of Australia and Austria. We are grateful to the Australian Research Council (ARC) and the Austrian Science Fund (FWF) for supporting our research programme on market segmentation analysis under ARC project numbers DP0557769, DP110101347, LX0559628 and LX0881890 and FWF project numbers P17382-N12, T351-N18 and V170-N18. Computations were partially run on the Vienna Scientific Cluster (VSC) under approval number 70419. We thank our industry partners for making available data sets, including the Austrian National Tourism Organisation (Österreich Werbung) and the Australian National Tourism Organisation (Tourism Australia). We thank Homa Hajibaba, Dominik Ernst and Syma Ahmed for technical support and feedback on earlier versions of the manuscript. We thank the Springer reviewers for recommendations for improvement and Joshua Hartmann for his assistance with illustrations.

Brisbane, Australia
Linz, Austria
Vienna, Austria
April 2018

Sara Dolnicar
Bettina Grün
Friedrich Leisch

References

- Chapman CN, McDonnell Feit E (2015) R for marketing research and analytics. UseR!. Springer International Publishing
- Dalgaard P (2008) Introductory statistics with R, 2nd edn. Springer, New York
- de Leeuw J, Mair P (2007) An introduction to the special volume on “Psychometrics in R”. *J Stat Softw* 20(1):1–5
- Fox J (2017) Using the R Commander: a point-and-click interface for R. Chapman & Hall/CRC Press, Boca Raton
- Hothorn T, Everitt BS (2014) A handbook of statistical analyses using R, 3rd edn. Chapman & Hall/CRC Press, Boca Raton
- Kabacoff RI (2015) R in action: data analysis and graphics with R, 2nd edn. Manning Publications, Shelter Island
- Putler DS, Krider RE (2012) Customer and business analytics: applied data mining for business decision making using R. Chapman & Hall/CRC Press, Boca Raton

Contents

Part I Introduction

1	Market Segmentation	3
1.1	Strategic and Tactical Marketing	3
1.2	Definitions of Market Segmentation	6
1.3	The Benefits of Market Segmentation	7
1.4	The Costs of Market Segmentation	8
References		9
2	Market Segmentation Analysis.....	11
2.1	The Layers of Market Segmentation Analysis	11
2.2	Approaches to Market Segmentation Analysis	13
2.2.1	Based on Organisational Constraints	13
2.2.2	Based on the Choice of (the) Segmentation Variable(s)	14
2.3	Data Structure and Data-Driven Market Segmentation Approaches	16
2.4	Market Segmentation Analysis Step-by-Step	20
References		22

Part II Ten Steps of Market Segmentation Analysis

3	Step 1: Deciding (not) to Segment	25
3.1	Implications of Committing to Market Segmentation	25
3.2	Implementation Barriers	26
3.3	Step 1 Checklist	27
References		29
4	Step 2: Specifying the Ideal Target Segment	31
4.1	Segment Evaluation Criteria	31
4.2	Knock-Out Criteria	33
4.3	Attractiveness Criteria	34

4.4	Implementing a Structured Process	34
4.5	Step 2 Checklist.....	36
	References.....	36
5	Step 3: Collecting Data.....	39
5.1	Segmentation Variables.....	39
5.2	Segmentation Criteria	41
5.2.1	Geographic Segmentation	42
5.2.2	Socio-Demographic Segmentation.....	43
5.2.3	Psychographic Segmentation.....	44
5.2.4	Behavioural Segmentation	44
5.3	Data from Survey Studies	45
5.3.1	Choice of Variables	45
5.3.2	Response Options.....	46
5.3.3	Response Styles.....	47
5.3.4	Sample Size	48
5.4	Data from Internal Sources.....	51
5.5	Data from Experimental Studies	52
5.6	Step 3 Checklist.....	53
	References.....	53
6	Step 4: Exploring Data	57
6.1	A First Glimpse at the Data	57
6.2	Data Cleaning	59
6.3	Descriptive Analysis	61
6.4	Pre-Processing	65
6.4.1	Categorical Variables	65
6.4.2	Numeric Variables	67
6.5	Principal Components Analysis.....	68
6.6	Step 4 Checklist.....	72
	References.....	72
7	Step 5: Extracting Segments.....	75
7.1	Grouping Consumers	75
7.2	Distance-Based Methods	78
7.2.1	Distance Measures	79
7.2.2	Hierarchical Methods.....	83
7.2.3	Partitioning Methods	89
7.2.4	Hybrid Approaches	106
7.3	Model-Based Methods	116
7.3.1	Finite Mixtures of Distributions.....	119
7.3.2	Finite Mixtures of Regressions.....	133
7.3.3	Extensions and Variations	141
7.4	Algorithms with Integrated Variable Selection	142
7.4.1	Biclustering Algorithms	143

7.4.2	Variable Selection Procedure for Clustering Binary Data (VSBD)	148
7.4.3	Variable Reduction: Factor-Cluster Analysis	151
7.5	Data Structure Analysis	153
7.5.1	Cluster Indices	154
7.5.2	Gorge Plots	159
7.5.3	Global Stability Analysis	161
7.5.4	Segment Level Stability Analysis	166
7.6	Step 5 Checklist.....	176
	References.....	176
8	Step 6: Profiling Segments	183
8.1	Identifying Key Characteristics of Market Segments	183
8.2	Traditional Approaches to Profiling Market Segments	184
8.3	Segment Profiling with Visualisations	186
8.3.1	Identifying Defining Characteristics of Market Segments	187
8.3.2	Assessing Segment Separation	190
8.4	Step 6 Checklist.....	196
	References.....	196
9	Step 7: Describing Segments	199
9.1	Developing a Complete Picture of Market Segments	199
9.2	Using Visualisations to Describe Market Segments	200
9.2.1	Nominal and Ordinal Descriptor Variables	200
9.2.2	Metric Descriptor Variables	206
9.3	Testing for Segment Differences in Descriptor Variables.....	210
9.4	Predicting Segments from Descriptor Variables	215
9.4.1	Binary Logistic Regression.....	217
9.4.2	Multinomial Logistic Regression	224
9.4.3	Tree-Based Methods	228
9.5	Step 7 Checklist.....	235
	References.....	235
10	Step 8: Selecting (the) Target Segment(s)	237
10.1	The Targeting Decision	237
10.2	Market Segment Evaluation	238
10.3	Step 8 Checklist.....	242
	References.....	243
11	Step 9: Customising the Marketing Mix	245
11.1	Implications for Marketing Mix Decisions	245
11.2	Product	247
11.3	Price	247
11.4	Place	250
11.5	Promotion	251

11.6 Step 9 Checklist	253
References	254
12 Step 10: Evaluation and Monitoring	255
12.1 Ongoing Tasks in Market Segmentation	255
12.2 Evaluating the Success of the Segmentation Strategy	255
12.3 Stability of Segment Membership and Segment Hopping	256
12.4 Segment Evolution	258
12.5 Step 10 Checklist	266
References	267
A Case Study: Fast Food	269
A.1 Step 1: Deciding (not) to Segment	269
A.2 Step 2: Specifying the Ideal Target Segment	270
A.3 Step 3: Collecting Data	270
A.4 Step 4: Exploring Data	271
A.5 Step 5: Extracting Segments	274
A.5.1 Using <i>k</i> -Means	274
A.5.2 Using Mixtures of Distributions	279
A.5.3 Using Mixtures of Regression Models	282
A.6 Step 6: Profiling Segments	285
A.7 Step 7: Describing Segments	288
A.8 Step 8: Selecting (the) Target Segment(s)	291
A.9 Step 9: Customising the Marketing Mix	294
A.10 Step 10: Evaluation and Monitoring	295
B R and R Packages	297
B.1 What Is R?	297
B.1.1 A Short History of R	297
B.1.2 R Packages	298
B.1.3 Quality Control	299
B.1.4 User Interfaces for R	299
B.2 R Packages Used in the Book	300
B.2.1 MSA	300
B.2.2 flexclust	301
B.2.3 flexmix	301
B.2.4 Other Packages	301
C Data Sets Used in the Book	303
C.1 Tourist Risk Taking	303
C.2 Winter Vacation Activities	304
C.3 Australian Vacation Activities	305
C.4 Australian Travel Motives	308
C.5 Fast Food	309

Contents	xiii
Glossary	311
References	319
Index	321

List of Figures

Fig. 1.1	Strategic and tactical marketing planning. (Modified from McDonald and Morris 1987)	4
Fig. 1.2	The asymmetry of strategic and tactical marketing. (Modified from McDonald and Morris 1987)	5
Fig. 2.1	The layers of market segmentation analysis	12
Fig. 2.2	Ten steps of market segmentation analysis	21
Fig. 5.1	Illustrating the importance of sufficient sample size in market segmentation analysis	48
Fig. 5.2	Effect of sample size on the correctness of segment recovery in artificial data. (Modified from Dolnicar et al. 2014) ..	49
Fig. 5.3	Sample size requirements in dependence of market and data characteristics. (Modified from Dolnicar et al. 2016)	50
Fig. 6.1	Histograms of tourists' age in the Australian travel motives data set	62
Fig. 6.2	Construction principles for box-and-whisker plots (tourists' age distribution)	63
Fig. 6.3	Box-and-whisker plot of tourists' age in the Australian travel motives data set	64
Fig. 6.4	Dot chart of percentages of YES answers in the Australian travel motives data set	65
Fig. 6.5	Principal components 2 and 3 for the Australian travel motives data set	71
Fig. 7.1	k -means and single linkage hierarchical clustering of two spirals	76
Fig. 7.2	A comparison of Euclidean and Manhattan distance	80
Fig. 7.3	A comparison of different linkage methods between two sets of points	84
Fig. 7.4	Single and complete linkage clustering of the tourist data shown in Table 7.2	86

Fig. 7.5	Complete linkage hierarchical cluster analysis of the tourist risk taking data set	88
Fig. 7.6	Bar chart of cluster means from hierarchical clustering for the tourist risk taking data set	89
Fig. 7.7	Simplified visualisation of the k -means clustering algorithm	91
Fig. 7.8	Artificial Gaussian data clustered using squared Euclidean distance (<i>left</i>), Manhattan distance (<i>middle</i>) and angle distance (<i>right</i>)	93
Fig. 7.9	Artificial mobile phone data set	95
Fig. 7.10	Three-segment k -means partition for the artificial mobile phone data set	96
Fig. 7.11	Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the artificial mobile phone data set	98
Fig. 7.12	Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the artificial mobile phone data set including a visualisation of the elbow consisting of two intersecting lines with different slopes	99
Fig. 7.13	Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the tourist risk taking data set	100
Fig. 7.14	Bar chart of cluster means from k -means clustering for the tourist risk taking data set	100
Fig. 7.15	Examples of good (<i>left</i>) and bad (<i>right</i>) starting points for k -means clustering	101
Fig. 7.16	Rectangular (<i>left</i>) and hexagonal (<i>right</i>) grid for self-organising maps	103
Fig. 7.17	5 \times 5 self-organising map of the tourist risk taking data set	104
Fig. 7.18	Schematic representation of an auto-encoding neural network with one hidden layer	105
Fig. 7.19	k -means clustering of the artificial mobile phone data set into 30 clusters	109
Fig. 7.20	Hierarchical clustering of the 30 k -means cluster centres of the artificial mobile phone data set	109
Fig. 7.21	Dendrogram for bagged cluster analysis of the winter vacation activities data set	112
Fig. 7.22	Bar chart of cluster means from bagged cluster analysis of the winter vacation activities data set	114
Fig. 7.23	Boxplot of cluster centres from bagged cluster analysis for segment 5 of the winter vacation activities data set	115
Fig. 7.24	Uncertainty plot of the mixture of normal distributions for the artificial mobile phone data set	121
Fig. 7.25	Visualisation of the 14 covariance models available in package <code>mclust</code>	124
Fig. 7.26	BIC values of the mixtures of normal distributions for the artificial mobile phone data set	124

Fig. 7.27	Scatter plot of the metric variables in the Australian travel motives data set	126
Fig. 7.28	Classification plot of the mixture of normal distributions for the Australian travel motives data set selected using the BIC among all covariance models	127
Fig. 7.29	Classification plot of the mixture of normal distributions for the Australian travel motives data set selected using the BIC among the models with identical covariance matrices across segments	128
Fig. 7.30	AIC, BIC and ICL values of mixtures of binary distributions for the winter vacation activities data set	132
Fig. 7.31	Bar chart of segment-specific probabilities of the mixture of binary distributions fitted to the winter vacation activities data set	134
Fig. 7.32	Artificial theme park data set	135
Fig. 7.33	Scatter plot with observations coloured by segment membership from a two-segment mixture of linear regression models for the artificial theme park data	136
Fig. 7.34	Scatter plot with observations coloured by segment membership together with the segment-specific regression lines from a two-segment mixture of linear regressions fitted to the Australian vacation motives data set	139
Fig. 7.35	Biclustering with constant pattern	144
Fig. 7.36	Bioclustering with constant column pattern	145
Fig. 7.37	Bicluster membership plot for the Australian vacation activities data set	147
Fig. 7.38	Bar chart of cluster means obtained for the Australian travel motives data set after selecting variables with the VSBD algorithm	150
Fig. 7.39	Segment separation plots (<i>left</i>), and gorge plots (<i>middle</i>) for the three-segment solutions, and global stability boxplots (<i>right</i>) for 2–9 segments for the three artificial data sets.	160
Fig. 7.40	Global stability boxplot using k -means clustering for 2–9 segments for the tourist risk taking data set	166
Fig. 7.41	Artificial mobile phone data set with three and six segments extracted	169
Fig. 7.42	Segment level stability within solutions (SLS_W) plot for the artificial mobile phone data set with three and six segments extracted	169
Fig. 7.43	Segment level stability within solutions (SLS_W) plot of the six-segment solution for the Australian travel motives data set....	171
Fig. 7.44	Segment level stability across solutions (SLS_A) plot for the artificial mobile phone data set for three to eight segments	173

Fig. 7.45	Segment level stability across solutions (SLS_A) plot for the Australian travel motives data set for three to eight segments	175
Fig. 8.1	Hierarchical clustering of the segmentation variables of the Australian travel motives data set using Ward's method	188
Fig. 8.2	Segment profile plot for the six-segment solution of the Australian travel motives data set	189
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	191
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)	192
Fig. 8.5	Segment separation plot using principal components 2 and 3 for the Australian travel motives data set	194
Fig. 8.6	Segment separation plot using principal components 2 and 3 for the Australian travel motives data set without observations ..	195
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	202
Fig. 9.2	Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set	203
Fig. 9.3	Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set	204
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	205
Fig. 9.5	Histograms of age by segment for the Australian travel motives data set	207
Fig. 9.6	Histograms of moral obligation to protect the environment by segment for the Australian travel motives data set	208
Fig. 9.7	Parallel box-and-whisker plot of age by segment for the Australian travel motives data set	208
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	209
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	210

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	214
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	220
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 step wise variable selection on the right	223
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.....	226
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	227
Fig. 9.15	Conditional inference tree using membership in segment 3 as dependent variable for the Australian travel motives data set	230
Fig. 9.16	Conditional inference tree using membership in segment 6 as dependent variable for the Australian travel motives data set	232
Fig. 9.17	Conditional inference tree using segment membership as dependent variable for the Australian travel motives data set	234
Fig. 10.1	Segment evaluation plot	241
Fig. 11.1	How the target segment decision affects marketing mix development	246
Fig. 11.2	Total expenditures in Australian dollars (AUD) for the last domestic holiday for tourists in segment 3 and all other tourists	249
Fig. 11.3	Hotel booking avenues used for the last domestic holiday by segment 3 and by the average tourist	250
Fig. 11.4	Information sources used by segment 3 and by the average tourist	252

Fig. 11.5	TV station most frequently watched by segment 3 and all other tourists	253
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	261
Fig. 12.2	Segment profile plot for the six-segment solution of winter vacation activities in 1991/92	262
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	263
Fig. 12.4	Segment profile plot for the six-segment solution of winter vacation activities in 1997/98	265
Fig. A.1	Principal components analysis of the fast food data set	274
Fig. A.2	Scree plot for the fast food data set	275
Fig. A.3	Global stability of k -means segmentation solutions for the fast food data set	276
Fig. A.4	Gorge plot of the four-segment k -means solution for the fast food data set	277
Fig. A.5	Segment level stability across solutions (SLS_A) plot from two to eight segments for the fast food data set	278
Fig. A.6	Segment level stability within solutions (SLS_W) plot for the fast food data set	279
Fig. A.7	Information criteria for the mixture models of binary distributions with 2 to 8 components (segments) for the fast food data set	280
Fig. A.8	Regression coefficients of the two-segment mixture of linear regression models for the fast food data set	285
Fig. A.9	Segment profile plot for the four-segment solution for the fast food data set	286
Fig. A.10	Segment separation plot using principal components 1 and 2 for the fast food data set	287
Fig. A.11	Shaded mosaic plot for cross-tabulation of segment membership and I LIKE IT for the fast food data set	289
Fig. A.12	Shaded mosaic plot for cross-tabulation of segment membership and gender for the fast food data set	290
Fig. A.13	Parallel box-and-whisker plot of age by segment for the fast food data set	290
Fig. A.14	Conditional inference tree using segment 3 membership as dependent variable for the fast food data set	292
Fig. A.15	Example of a simple segment evaluation plot for the fast food data set	293

List of Tables

Table 2.1	Examples of commonly used segmentation variables	15
Table 2.2	Combinations of segmentation approaches based on the nature of segmentation variables used. (Modified from Dolnicar 2004)	17
Table 2.3	Data-driven market segmentation approaches based on data structure. (Modified from Dolnicar and Leisch 2010)	19
Table 4.1	Criteria proposed in the literature for the evaluation of market segments in chronological order. (Modified from: Karlsson 2015)	32
Table 5.1	Gender as a possible segmentation variable in commonsense market segmentation	40
Table 5.2	Segmentation variables in data-driven market segmentation	40
Table 7.1	Data set and segment characteristics informing extraction algorithm selection	77
Table 7.2	Artificial data set on tourist activities: percentage of time spent on three activities	79
Table 7.3	The 14 covariance models available in package <i>mclust</i>	123
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	185
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 <i>p</i> -values	212
Table 10.1	Data underlying the segment evaluation plot	240

Part I

Introduction

Chapter 1

Market Segmentation



1.1 Strategic and Tactical Marketing

The purpose of marketing is to match the genuine needs and desires of consumers with the offers of suppliers particularly suited to satisfy those needs and desires. This matching process benefits consumers and suppliers, and drives an organisation's marketing planning process.

Marketing planning is a logical sequence and a series of activities leading to the setting of marketing objectives and the formulation of plans to achieving them (McDonald and Wilson 2011, p. 24). A marketing plan consists of two components: a strategic and a tactical marketing plan. The strategic plan outlines the long-term direction of an organisation, but does not provide much detail on short-term marketing action required to move in this long-term direction. The tactical marketing plan does the opposite. It translates the long-term strategic plan into detailed instructions for short-term marketing action. The strategic marketing plan states where the organisation wants to go and why. The tactical marketing plan contains instructions on what needs to be done to get there.

This process is much like going on a hiking expedition (Fig. 1.1). Before starting a hike, it is critically important to organise a map, and figure out where exactly one's present location is. Once the present location is known, the next step is to decide which mountain to climb. The choice of the mountain is a strategic decision; it determines all subsequent decisions. As soon as this strategic decision is made, the expedition team can move on to tactical decisions, such as: which shoes to wear for this particular hike, which time of day to depart, and how much food and drink to pack. All these tactical decisions are important to ensure a safe expedition, but they depend entirely on the strategic decision of which mountain to climb.

Preparations for the mountain climbing expedition are similar to the development of an organisational marketing plan. The strategic marketing plan typically identifies consumer needs and desires, strengths and weaknesses internal to the organisation, and external opportunities and threats the organisation may face. A SWOT analysis

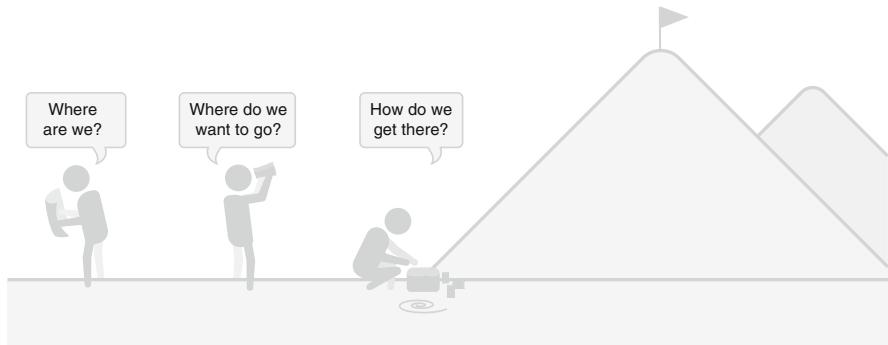


Fig. 1.1 Strategic and tactical marketing planning. (Modified from McDonald and Morris 1987)

explicitly states an organisation's strengths (*S*), weaknesses (*W*), opportunities (*O*), and threats (*T*). As such, the SWOT analysis outlines one side of the matching process: what the supplier is particularly suitable to offer consumers.

The other side of the matching process – consumer needs and desires – is typically investigated using market research. Despite the heavy reliance of market research on survey methodology, a wide range of sources of information are available to explore, and gain detailed insight, into what consumers need or desire, including qualitative research involving focus groups and interviews, observational and experimental research.

Once organisational strengths have been established, potential interference by external factors has been assessed, and consumer needs and desires have been thoroughly investigated, two key decisions have to be made as part of the strategic marketing planning process: which consumers to focus on (segmentation and targeting), and which image of the organisation to create in the market (positioning). These decisions are critical because they determine the long-term direction of the organisation, and cannot easily be reversed.

Only when it has been decided which group of consumers (market segment) the organisation is going to cater for, and how it will present itself to the public to appear most attractive to this target segment, does work on the tactical marketing plan begin. Tactical marketing planning usually covers a period of up to one year. It is traditionally seen to cover four areas: the development and modification of the product in view of needs and desires of the target segment (Product), the determination of the price in view of cost, competition, and the willingness to pay of the target segment (Price), the selection of the most suitable distribution channels to reach the target segment (Place), and the communication and promotion of the offer in a way that is most appealing to the target segment (Promotion).

The tactical marketing plan depends entirely on the strategic marketing plan, but the strategic marketing plan does not depend on the tactical marketing plan. This asymmetry is illustrated in Fig. 1.2 using the mountain expedition analogy. Strategic marketing is responsible for identifying the most suitable mountain to

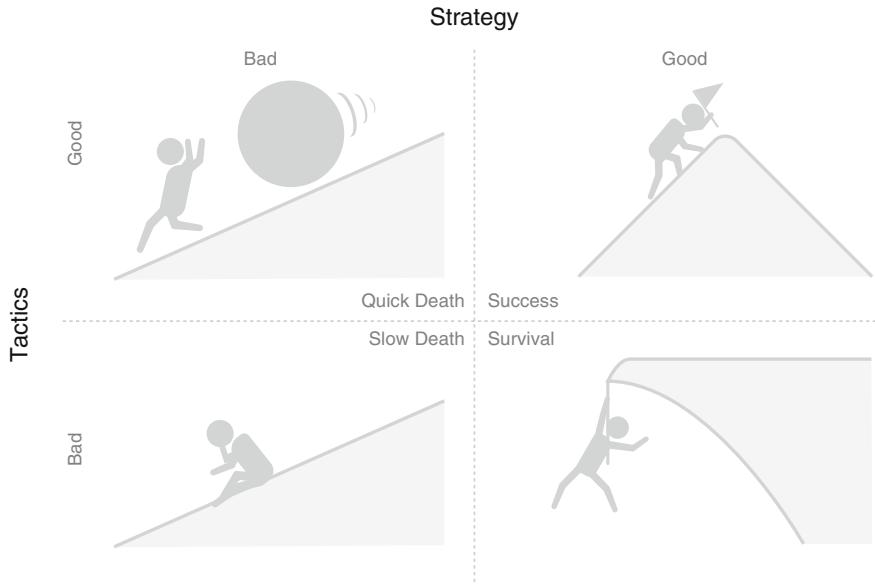


Fig. 1.2 The asymmetry of strategic and tactical marketing. (Modified from McDonald and Morris 1987)

climb. Tactical marketing is responsible for the equipment: the quality of the walking shoes, food, water, a raincoat. As long as the strategic marketing is good, the expedition leads to the right peak. Whether tactical marketing is efficient or not only determines how comfortable (top right hand quadrant in Fig. 1.2) or uncomfortable (bottom right hand quadrant in Fig. 1.2) survival is. If, however, the strategic marketing plan is bad, tactical marketing cannot help. It only affects if the wrong mountain – and with it organisational failure – is reached quickly (top left hand quadrant in Fig. 1.2) or slowly (bottom left hand quadrant in Fig. 1.2).

The combination of good strategic marketing and good tactical marketing leads to the best possible outcome. Bad strategic marketing combined with bad tactical marketing leads to failure, but this failure unfolds slowly. A faster pathway to failure is to have excellent tactical marketing based on bad strategic marketing. This is equivalent to running full speed up to the wrong mountain. Good strategic marketing combined with bad tactical marketing ensures survival, albeit not in a particularly happy place.

To conclude: the importance of strategic and tactical marketing for organisational success is asymmetric. Good tactical marketing can never compensate for bad strategic marketing. Strategic marketing is the foundation of organisational success.

1.2 Definitions of Market Segmentation

Market segmentation is a decision-making tool for the marketing manager in the crucial task of selecting a target market for a given product and designing an appropriate marketing mix (Tynan and Drayton 1987, p. 301). Market segmentation is one of the key building blocks of strategic marketing. Market segmentation is essential for marketing success: the most successful firms drive their businesses based on segmentation (Lilien and Rangaswamy 2003, p. 61). Market segmentation lies at the heart of successful marketing (McDonald 2010), tools such as segmentation [...] have the largest impact on marketing decisions (Roberts et al. 2014, p. 127).

Smith (1956) was the first to propose the use of segmentation as a marketing strategy. Smith defines market segmentation as viewing a heterogeneous market (one characterised by divergent demand) as a number of smaller homogeneous markets (p. 6). Conceptually, market segmentation sits between the two extreme views that (a) all objects are unique and inviolable and (b) the population is homogeneous (Saunders 1980, p. 422). One of the simplest and clearest definitions is that used in a newsletter by Grey Advertising Inc. and cited in Haley (1985, p. 8): market segmentation means cutting markets into slices. Ideally, consumers belonging to the same market segments – or sets of buyers (Tynan and Drayton 1987) – are very similar to one another with respect to the consumer characteristics deemed critical by management. At the same time, optimally, consumers belonging to different market segments are very different from one another with respect to those consumer characteristics. Consumer characteristics deemed critical to market segmentation by management are referred to as segmentation criteria.

The segmentation criterion can be one single consumer characteristic, such as age, gender, country of origin, or stage in the family life cycle. Alternatively, it can contain a larger set of consumer characteristics, such as a number of benefits sought when purchasing a product, a number of activities undertaken when on vacation, values held with respect to the environment, or an expenditure pattern.

An ideal market segmentation situation – for the simplest case of two product features – is illustrated in the left hand panel of Table 2.3 on page 19. The x -axis shows the number of desired features of a mobile telephone, and the y -axis shows the price consumers are willing to pay. Here, three market segments exist: a small segment characterised by wanting many mobile telephone features, and being willing to pay a lot of money for it; a large segment containing consumers who desire the exact opposite (a simple, cheap mobile phone); and another large segment in the middle containing members who want a mid-range phone at a mid-range price. This example illustrates Smith's definition of market segmentation with each of the segments representing one homogeneous market within a larger heterogeneous market.

The example also illustrates why market segmentation is critical to organisational success. A mobile phone company attempting to offer one mobile phone to the entire market is unlikely to satisfy the needs of each of those segments; and unlikely to develop an image in the marketplace that is distinct and reflects an offer

desirable to consumers. Rather, tactical marketing efforts may be wasted because the mobile phone company fails to cater for any of the homogeneous market segments. Selecting one market segment, say the high-end, high-price segment, and offering this segment the exact product it desires, is more likely to lead to both high short-term sales (within this segment), and a long-term positioning as being the best possible provider of high-end, high-price mobile telephones.

Such an approach is referred to as a *concentrated* market strategy (Croft 1994). A concentrated strategy is attractive for organisations who are resource-poor, but are facing fierce competition in the market. Concentrating entirely on satisfying the needs of one market segment can secure the future for such an organisation. It does, however, come at the price of the higher risk associated with depending on one single market segment entirely. An alternative approach, if the capabilities of the organisation permit it, is to pursue a *differentiated* market strategy, and produce three telephones, one for each segment. In such a case, all aspects of the marketing mix would have to be customised for each of the three target segments. A differentiated strategy is suitable in mature markets (Croft 1994) where consumers are capable of differentiating between alternative products. Product variations can thus be customised to meet the needs of a number of market segments. When an organisation decides not to use market segmentation, it is effectively choosing to pursue an *undifferentiated* market strategy, where the same product is marketed using the same marketing mix to the entire market. Examples of undifferentiated marketing include petrol and white bread; they are not particularly targeted at any group within the marketplace. Such an approach may be viable for resource-rich organisations, or in cases where a new product is introduced (Croft 1994), and consumers are not yet able to discriminate between alternative products.

1.3 The Benefits of Market Segmentation

Market segmentation has a number of benefits. At the most general level, market segmentation forces organisations to take stock of where they stand, and where they want to be in future. In so doing, it forces organisations to reflect on what they are particularly good at compared to competitors, and make an effort to gain insights into what consumers want. Market segmentation offers an opportunity to think and rethink, and leads to critical new insights and perspectives.

When implemented well, market segmentation also leads to tangible benefits, including a better understanding of differences between consumers, which improves the match of organisational strengths and consumer needs (McDonald and Dunbar 1995). Such an improved match can, in turn, form the basis of a long-term competitive advantage in the selected target segment(s). The extreme case of long-term competitive advantage is that of market dominance, which results from being best able to cater to the needs of a very specific niche segment (McDonald and Dunbar 1995). Ideal niche segments match the organisational skill set in terms of their needs, are large enough to be profitable, have solid potential for growth,

and are not interesting to competitors (Kotler 1994). Taking market segmentation to the extreme would mean to actually be able to offer a customised product or service to very small groups of consumers. This approach is referred to as micro marketing or hyper-segmentation (Kara and Kaynak 1997). One step further leads to what Kara and Kaynak (1997) refer to as *finer segmentation* where each consumer represents their own market segment. Finer segmentation approaches are becoming more viable with the rise of eCommerce and the use of sophisticated consumer databases enabling providers of products and services to learn from a person's purchase history about what to offer them next.

A marketing mix developed to best reflect the needs of one or more segments is also likely to yield a higher return on investment because less of the effort that goes into the design of the marketing mix is wasted on consumers whose needs the organisation could never satisfy anyway. For small organisations, it may be essential for survival to focus on satisfying very distinct needs of a small group of consumers because they simply lack the financial resources to serve a larger market or multiple market segments (Haley 1985).

Market segmentation has also been shown to be effective in sales management (Maier and Saunders 1990) because it allows direct sales efforts to be targeted at groups of consumers rather than each consumer individually.

At an organisational level, market segmentation can contribute to team building (McDonald and Dunbar 1995) because many of the tasks associated with conducting a market segmentation analysis require representatives from different organisational units to work as a team. If this is achieved successfully, it can also improve communication and information sharing across organisational units.

1.4 The Costs of Market Segmentation

Implementing market segmentation requires a substantial investment by the organisation. A large number of people have to dedicate a substantial amount of time to conduct a thorough market segmentation analysis. If a segmentation strategy is pursued, more human and financial resources are required to develop and implement a customised marketing mix. Finally, the evaluation of the success of the segmentation strategy, and the continuous monitoring of market dynamics (that may point to the need for the segmentation strategy to be modified) imply an ongoing commitment of resources. These resource commitments are made under the assumption that the organisation will benefit from a return on this investment. Yet, the upfront investment is substantial.

In the worst case, if market segmentation is not implemented well, the entire exercise is a waste of resources. Instead of leading to competitive advantage, a failed market segmentation strategy can lead to substantial expenses generating no additional return at all, instead disenfranchising staff involved in the segmentation exercise.

It is for this very reason, that an organisation must make an informed decision about whether or not to embark on the long journey of market segmentation analysis, and the even longer journey of pursuing a market segmentation strategy.

References

- Croft M (1994) Market segmentation. Routledge, London
- Haley RI (1985) Developing effective communications strategy – a benefit segmentation approach. Wiley, New York
- Kara A, Kaynak E (1997) Markets of a single customer: exploiting conceptual developments in market segmentation. *Eur J Mark* 31(11/12):873–895
- Kotler P (1994) Marketing management, 8th edn. Prentice-Hall, Englewood Cliffs
- Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River
- Maier J, Saunders J (1990) The implementation process of segmentation in sales management. *J Pers Sell Sales Manag* 10(1):39–48
- McDonald M (2010) Existentialism – a school of thought based on a conception of the absurdity of the universe. *Int J Market Res* 52(4):427–430
- McDonald M, Dunbar I (1995) Market segmentation: a step-by-step approach to creating profitable market segments. Macmillan, London
- McDonald MHB, Morris P (1987) The marketing plan: a pictorial guide for managers. Butterworth Heineman, London
- McDonald M, Wilson H (2011) Marketing plans: how to prepare them, how to use them. Wiley, Hoboken
- Roberts J, Kayande U, Stremersch S (2014) From academic research to marketing practice: exploring the marketing science value chain. *Int J Res Mark* 31:127–140
- Saunders JA (1980) Cluster analysis for market segmentation. *Eur J Mark* 14(7):422–435
- Smith W (1956) Product differentiation and market segmentation as alternative marketing strategies. *J Mark* 21:3–8
- Tynan AC, Drayton J (1987) Market segmentation. *J Mark Manag* 2(3):301–335

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2

Market Segmentation Analysis



2.1 The Layers of Market Segmentation Analysis

Market segmentation analysis, at its core (see Fig. 2.1), is

the process of grouping consumers into naturally existing or artificially created segments of consumers who share similar product preferences or characteristics.

This process is typically a statistical one. Yet, it is exploratory in nature. Many decisions made by the data analyst in the process of extracting market segments from consumer data affect the final market segmentation solution. For market segmentation analysis to be useful to an organisation, therefore, both a competent data analyst, and a user who understands the broader mission of the organisation (or that of their organisational unit when working in a team) need to be involved when market segments are extracted from consumer data. Throughout this book, we use the term *user* to mean the user of the segmentation analysis; the person or department in the organisation that will use the results from the market segmentation analysis to develop a marketing plan.

To ensure that the grouping of consumers is of the highest quality, a number of additional tasks are required, as illustrated in the second layer in Fig. 2.1. All these tasks are still primarily technical in nature. Collecting good data, for example, is critically important. The statistical segment extraction process at the core of market segmentation analysis cannot compensate for bad data. The grouping of consumers can always only be as good as the data provided to the segment extraction method.

Upon completion of data collection, but before the actual segment extraction takes place, the data needs to be explored to gain preliminary insight into the nature of the market segmentation study that can be conducted using this data. Finally, after consumers have been grouped into market segments, each of these segments needs to be profiled and described in detail. Profiling and describing segments help users

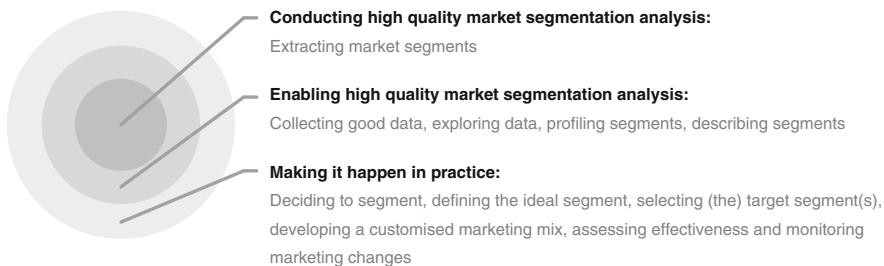


Fig. 2.1 The layers of market segmentation analysis

to understand each of the segments, and select which one(s) to target. When one or more target segments have been chosen, profiling and describing segments inform the development of the customised marketing mix.

If all the tasks in the first (core) and second layer of market segmentation analysis have been implemented well, the result is a theoretically excellent market segmentation solution. But a theoretically excellent market segmentation solution is meaningless unless users can convert such a solution into strategic marketing decisions and tactical marketing action. Therefore, for any market segmentation analysis to be complete, a third layer is required. This third layer includes non-technical tasks. These tasks represent organisational implementation issues, and do not sequentially follow the first and the second layer. As illustrated in Fig. 2.1, the third layer of implementation tasks wraps around technical tasks.

Before any technical tasks are undertaken, an organisation needs to assess whether, in their particular case, implementing a market segmentation strategy will lead to market opportunities otherwise unavailable to them. If the market segmentation analysis points to such opportunities, the organisation must be willing to commit to this long-term strategy. All of these decisions have to be made by the users, and are entirely independent of the technical task of extracting market segments from data.

User input is also critically important at the data collection stage to ensure that relevant information about consumers will be captured. Again, this is not a decision a data analyst can make.

Upon completion of the segment extraction task, users need to assess resulting market segments or market segmentation solutions, and select one or more target segments. Data analysts can provide facts about these segments, but cannot select the most suitable ones. This selection is driven, in part, by the strengths and opportunities of the organisation, and their alignment with the key needs of the market segments. Finally, as soon as one or more target segments have been selected, users need to develop a marketing plan for those market segments, and design a customised marketing mix.

2.2 Approaches to Market Segmentation Analysis

No one single approach is best when conducting market segmentation analysis. Instead, approaches to market segmentation analysis can be systematised in a number of different ways. We present two systematics here, one uses as its basis the extent to which the organisation conducting the market segmentation study is willing or able to make changes to their current approach of targeting the market or a segment of the market and has been proposed by Dibb and Simkin (2008). It is based on the premise that organisations are not in the position to choose any of the available approaches to market segmentation analysis due to organisational constraints. The second systematics is based on the nature of the segmentation variable or variables used in the market segmentation analysis.

2.2.1 Based on Organisational Constraints

Dibb and Simkin (2008) distinguish three approaches to market segmentation: the quantitative survey-based approach, the creation of segments from existing consumer classifications, and the emergence of segments from qualitative research. These three approaches differ in how radical the resulting change is for the organisation. We refer to the approach requiring the most radical change in the organisation as *segment revolution*. It is like jumping on a sandcastle and building a new one. It starts from zero. A less radical approach is that of *segment evolution*, which is like refining an existing sandcastle. As long as the sandcastle is robust, and not too close to the water, this is a perfectly reasonable approach. The least radical approach is not really even a segmentation approach, it is like walking down the beach and seeing a huge pile of sand and thinking: this would make a fantastic sandcastle. It is a random discovery, like a *mutation*, which – if noticed and acted upon – also has the potential of allowing the organisation to harvest the benefits of market segmentation.

Looking at each one of these approaches in more detail, the *segment revolution* or *quantitative survey-based segmentation approach* tends to be seen as the prototypical market segmentation analysis. The key assumption underlying this approach is that the organisation conducting market segmentation analysis is willing and able to start from scratch; to forget entirely about how its marketing was conducted in the past, and commence the segmentation process with a genuinely open mind. If market segmentation analysis reveals a promising niche segment, or a promising set of market segments to target with a differentiated market strategy, the organisation must develop an entirely new marketing plan in view of those findings.

While this approach is indeed a textbook approach in terms of having the highest probability of harvesting all the benefits market segmentation strategy has to offer, it is often not viable in reality. Possible reasons include the unwillingness or inability of an organisation to change sufficiently, or the use of established segments

performing reasonably well. In such cases, market segmentation analysis does not have to be abandoned altogether. Other, less radical approaches are available, including that of *creating segments from currently targeted sectors and segments*. This approach – representing segment evolution rather than revolution – is one of refining and sharpening segment focus. While informed by data and possibly also market research, it is typically achieved by intra-organisational workshopping. Dibb and Simkin (2008) offer a proforma to guide organisations through this process.

The third approach is that of *exploratory research pointing to segments*. Under this approach, market segments are stumbled upon as part of an exploratory research process possibly being undertaken for a very different purpose initially. In times of big data, such segment mutation may well result from data mining of streams of data, rather than from qualitative research. The same holds for segment evolution. The continuous tracking of the nature of market segments in large streams of data flowing in on a continuous basis can be used to check on an ongoing basis whether market structure has changed in ways which make it necessary to adapt the segmentation strategy to ensure organisational survival and prosperity.

2.2.2 *Based on the Choice of (the) Segmentation Variable(s)*

A more technical way of systematising segmentation approaches is to use as a basis the nature of consumer characteristics used to extract market segments. Sometimes one single piece of information about consumers (one segmentation variable) is used. This statistical problem is unidimensional. One example is age. The resulting segments are age groups, and older consumers could be selected as a target segment.

In other cases, multiple pieces of information (multiple segmentation variables) about consumers are important. In this case, the statistical problem becomes multidimensional. One example could be consumers' expenditure patterns. An expenditure pattern underlying a market segmentation analysis could be the total dollars spent on ten different vacation activities, including entrance fees to theme parks, dining out, shopping and so on. Imagine that a tourist destination known for its man-made attractions is trying to identify a suitable target market. Using tourists' expenditure patterns could be useful in this context, helping the destination focus on those tourists who have in the past spent a lot of money on entrance fees for theme parks and zoos. It is reasonable to expect that this past expenditure pattern is predictive of future expenditures. If these tourists can be attracted to the destination, they are likely to make extensive use of the man-made attractions on offer. A few examples of commonly used segmentation variables are provided in Table 2.1.

When one single segmentation variable is used, the segmentation approach is referred to as *a priori* (Mazanec 2000), *convenience-group* (Lilien and Rangaswamy 2003) or *commonsense* market segmentation (Dolnicar 2004). Morritt (2007) describes this approach to market segmentation as one that is created without the benefit of primary market research. Managerial intuition, analysis of secondary data sources, analysis of internal consumer databases, and previously existing segments

Table 2.1 Examples of commonly used segmentation variables

Variable	Dimensions	Sample survey question
Age	Unidimensional	How old are you?
Gender	Unidimensional	Are you female or male?
Country of origin	Unidimensional	Where do you live?
Prior purchase	Unidimensional	Have you booked a cruise trip before?
Benefits sought	Multidimensional	When booking flights online, do you care about <ul style="list-style-type: none"> • convenience • value for money • speed • ability to compare fares
Motives	Multidimensional	When choosing a vacation, do you want to <ul style="list-style-type: none"> • rest and relax • explore new things • meet new people • learn about other cultures • get away from everyday routine

are used to group consumers into different segments (p. 9). The term *a priori segmentation* indicates that the decision about what characterises each segment is made in advance, before any data analysis is conducted. The term *commonsense segmentation* implies that users apply their common sense to choose their target segment. The term *convenience-group segmentation* indicates that the market segments are chosen for the convenience of serving them. When commonsense segmentation is conducted, the provider of the product usually has a reasonably good idea of the nature of the appropriate segment or segments to target. The aim of the segmentation analysis therefore is not to identify the key defining characteristic of the segment, but to gain deeper insight into the nature of the segments.

An example of commonsense segmentation is brand segmentation. Hammond et al. (1996) show in their study that consumers who purchase specific brands do not have distinct profiles with respect to descriptor variables. Of course, this does not hold for all commonsense segmentations. On the contrary: if a powerful segmentation variable is identified, which is reflective of some aspect of purchase behaviour, commonsense segmentation represents a very efficient approach because it is simpler and fewer mistakes can occur in the process of a commonsense market segmentation analysis. Lilien and Rangaswamy (2003) view this kind of market segmentation approach as reactive.

The proactive approach, which exploits multiple segmentation variables, is referred to as *a posteriori* (Mazanec 2000), *cluster based* (Wind 1978; Green 1977) or *post hoc* segmentation (Myers and Tauber 1977). These terms indicate that the nature of the resulting market segments is not known until after the data analysis has been conducted. An alternative term used is that of *data-driven segmentation* (Dolnicar 2004). This term implies that the segmentation solution is

determined through data analysis, that data analysis creates the solution. Morritt (2007) identifies the key characteristic of this approach as being based on primary (original) research into the preferences and purchase behaviour of your target market (p. 9).

When data-driven segmentation is conducted, the organisation has certain assumptions about the consumer characteristics that are critical to identifying a suitable market segment to target, but does not know the exact profiles of suitable target segments. The aim of data-driven segmentation, therefore, is twofold: first, to explore different market segments that can be extracted using the segmentation variables chosen, and, second, to develop a detailed profile and description of the segment(s) selected for targeting.

Commonsense and data-driven segmentation are two extremes, the two pure forms of segmentation approaches based on the nature of the segmentation criterion. In reality, market segmentation studies rarely fall into one of those clear-cut categories. Rather, various combinations of those approaches are used either sequentially or simultaneously, as can be seen in Table 2.2.

Commonsense/commonsense segmentation results from splitting consumers up into groups using one segmentation variable first. Then, one of the resulting segments is selected and split up further using a second segmentation variable. At the other extreme, data-driven/data-driven segmentation is the result of combining two sets of segmentation variables. Table 2.2 provides a few examples.

Morritt (2007) recommends the use of such combinations of segmentation variables in market segmentation analysis, which he refers to as *two-stage, or multi-stage segmentation*. An example of such a multi-stage segmentation is provided by Boksberger and Laesser (2009) who use a set of travel motives as the segmentation variables for data-driven segmentation after having pre-selected senior travellers using a commonsense segmentation approach.

2.3 Data Structure and Data-Driven Market Segmentation Approaches

When conducting data-driven market segmentation, data analysts and users of market segmentation solutions often assume that market segments naturally exist in the data. Such naturally occurring segments, it is assumed, need to merely be revealed and described. In real consumer data, naturally existing, distinct and well-separated market segments rarely exist.

This leads to the question: should market segments be extracted if they do not naturally exist in the data? Dubes and Jain (1979, p. 242) answer this question in the context of cluster validation: it is certainly foolish to impose a clustering structure on data known to be random. Their view was largely shared by the pioneers of market segmentation (Frank et al. 1972; Myers and Tauber 1977) who worked on the assumption that taxonomic procedures describe natural groups present in

Table 2.2 Combinations of segmentation approaches based on the nature of segmentation variables used. (Modified from Dolnicar 2004)

	Commonsense/ commonsense segmentation	Commonsense/ data-driven segmentation	Data-driven/ commonsense segmentation	Data-driven/ data-driven segmentation
Primary segmentation variable(s)	Commonsense (e.g. age, country of origin)	Commonsense (e.g. age, country of origin)	Data-driven (e.g. expenditures, vacation activities)	Data-driven (e.g. travel motives, expenditures)
Secondary segmentation variable(s)	Commonsense (e.g. gender, seeking adventure or not)	Data-driven (e.g. travel motives, vacation activities)	Commonsense (e.g. gender, family status)	Data-driven (e.g. vacation activities, information sources used)
Example 1	Young female tourists	Mature aged tourists who play golf, enjoy wine-tastings and fine dining	Tourists who engage in a large number of activities that attract an entrance fee, such as visiting theme parks and zoos, and who travel with their family	Tourists who want to learn about culture and local people, and who attend local cultural events and food festivals
Example 2	Adventure travellers from Australia	Older tourists who take a holiday to relax, have a change of usual surroundings, and enjoy health / beauty treatments	Tourists who surf and enjoy the night life of the destination, and who are male	Tourists who have high expenditures in a wide range of expenditure categories at the destination, and use airline loyalty program mail outs as their key travel information source

empirical data. Myers and Tauber (1977, p. 71) explicitly state that the aim of market segmentation is to search for ‘natural groupings’ of objects and define market segments as clearly defined natural groupings of people.

More recently, however, acceptance of the fact that empirical data sets typically used for the purpose of market segmentation do not display much cluster structure, has led to a modified view: Mazanec (1997) and Wedel and Kamakura (2000) argue that market segmentation is in fact the process of creating artificial segments that can help users develop more effective marketing strategies. The value of this position has been acknowledged in the early works on market segmentation, despite the fact that the authors of those early studies still aimed at identifying natural segments. Myers and Tauber (1977, p. 74), for example, show an empirical data set which

does not contain natural market segments and ask: Does this mean that there are no actionable segments? Myers and Tauber (1977) then proceed by answering that this is not necessarily the case. Rather, as long as market segments can be created from the empirical data in a way that makes members of the segment similar, while at the same time being distinctly different from other consumers, they may well be of value to an organisation.

Dolnicar and Leisch (2010) distinguish three possible conceptual approaches to data-driven market segmentation: natural, reproducible or constructive segmentation (Table 2.3).

The term *natural segmentation* reflects the traditional view that distinct market segments exist in the data, and that the aim of market segmentation analysis is to find them. This traditional view is reflected well in the statement that the initial premise in segmenting a market is that segments actually do exist (Beane and Ennis 1987, p. 20).

The term *reproducible segmentation* refers to the case where natural market segments do not exist in the data. But the data are not entirely unstructured either. Rather, the data contain some structure – other than cluster structure – making it possible to generate the same segmentation solution repeatedly. The ability to repeatedly reveal the same or very similar market segments, makes results of data-driven segmentation studies less random and more reliable. Reliable results represent a stronger basis for long-term strategic segmentation decisions.

Finally, the term *constructive segmentation* refers to the case where neither cluster structure nor any other data structure exists, which would enable the data analyst to reproduce similar segmentation solutions repeatedly across replications. At first the question arises: should such data be segmented at all? Are segments resulting from such data sets managerially useful? After all they are merely random creations of the data analyst. The answer is: yes. It does make sense to conduct constructive market segmentation because, even if consumer preferences are spread evenly across all possible combinations of attributes, it is still more promising to target subgroups of these consumers (for example, those who like to have many functions on the mobile phone despite a higher price) than to attempt to satisfy the entire range of consumer needs.

The problem is: at the beginning of a market segmentation analysis it is not known whether the empirical data permits natural segmentation, or whether it requires constructive segmentation. Ernst and Dolnicar (2018) provide a rough estimate of the frequency of occurrence of each one of those concepts by classifying 32 empirical tourism survey data sets. These data sets varied greatly in sample size, response formats offered to survey participants, and the nature of the constructs. Results suggest that natural segmentation is extremely rare. Only two data sets (6% of the data sets investigated) contained natural market segments. This finding has major implications: it points to the fact that it is absolutely essential to conduct data structure analysis (see Sect. 7.5) before extracting segments. Results also suggest

Table 2.3 Data-driven market segmentation approaches based on data structure. (Modified from Dolnicar and Leisch 2010)

Natural segmentation	Reproducible segmentation	Constructive segmentation
Cluster structure in the data	Structure in the data, but not cluster structure	No structure in the data
Market segments are revealed	Market segments are created	Market segments are created
Repeated calculations lead to very similar results (very stable across repeated calculations)	Repeated calculations lead to similar results (reasonably stable across repeated calculations)	Repeated calculations lead to different results (unstable across repeated calculations)
Example data:	Example data:	Example data:

that the worst case scenario – the entire lack of data structure – occurs in only 22% of cases. Nearly three quarters of data sets analysed contain some structure – other than cluster structure – which can be exploited to extract market segments re-occurring across repeated calculations.

The proposed conceptualisation, as well as previous empirical estimates of the frequency of occurrence of each of those concepts, indicate that conducting data structure analysis in advance of the actual data-driven market segmentation analysis is a good idea. This is comparable to driving a car in a new city following a navigation system or looking at the map first, to get a feeling for the lay of the land, then planning the route and driving. Data structure analysis achieves a similar aim: it provides an overall picture of the data, which helps to avoid bad methodological decisions and misinterpretations when segmenting the data. A simple way of getting a feeling for the structure of the data, is to repeatedly segment it with different numbers of segments and different algorithms. An automated approach – using stability of repeated segmentation solutions as a criterion – is proposed by Dolnicar and Leisch (2010) and will be discussed in detail in Sect. 7.5. Whichever approach the data analyst chooses, it will provide insight as to the concept of market segmentation study that can be implemented. In the case of natural clustering, the data analyst needs little input from users because the solution is obvious. At the other extreme, when data are entirely unstructured, the data analyst must work hand in hand with users of the market segmentation solution to construct the most strategically useful market segments.

2.4 Market Segmentation Analysis Step-by-Step

We recommend a ten-step approach to market segmentation analysis. Figure 2.2 illustrates the ten steps. The basic structure is the same for both commonsense and data-driven market segmentation: an organisation needs to weigh up the advantages and disadvantages of pursuing a segmentation strategy, and decide whether or not to go ahead (Step 1). Next, the organisation needs to specify characteristics of their ideal market segment (Step 2). Only after this preliminary and predominantly conceptual work is finalised, is empirical data collected or compiled from existing sources (Step 3). These data need to be explored (Step 4) before market segments are extracted (Step 5). The resulting market segments are profiled (Step 6), and described (Step 7) in detail. Step 8 is the point of no return where the organisation carefully selects one or a small number of market segments to target. Based on this choice, a customised marketing mix is developed (Step 9). Upon completion of the market segmentation analysis, the success of implementing a market segmentation strategy needs to be evaluated, and segments need to be continuously monitored (Step 10) for possible changes in size or in characteristics. Such changes may require modifications to the market segmentation strategy.

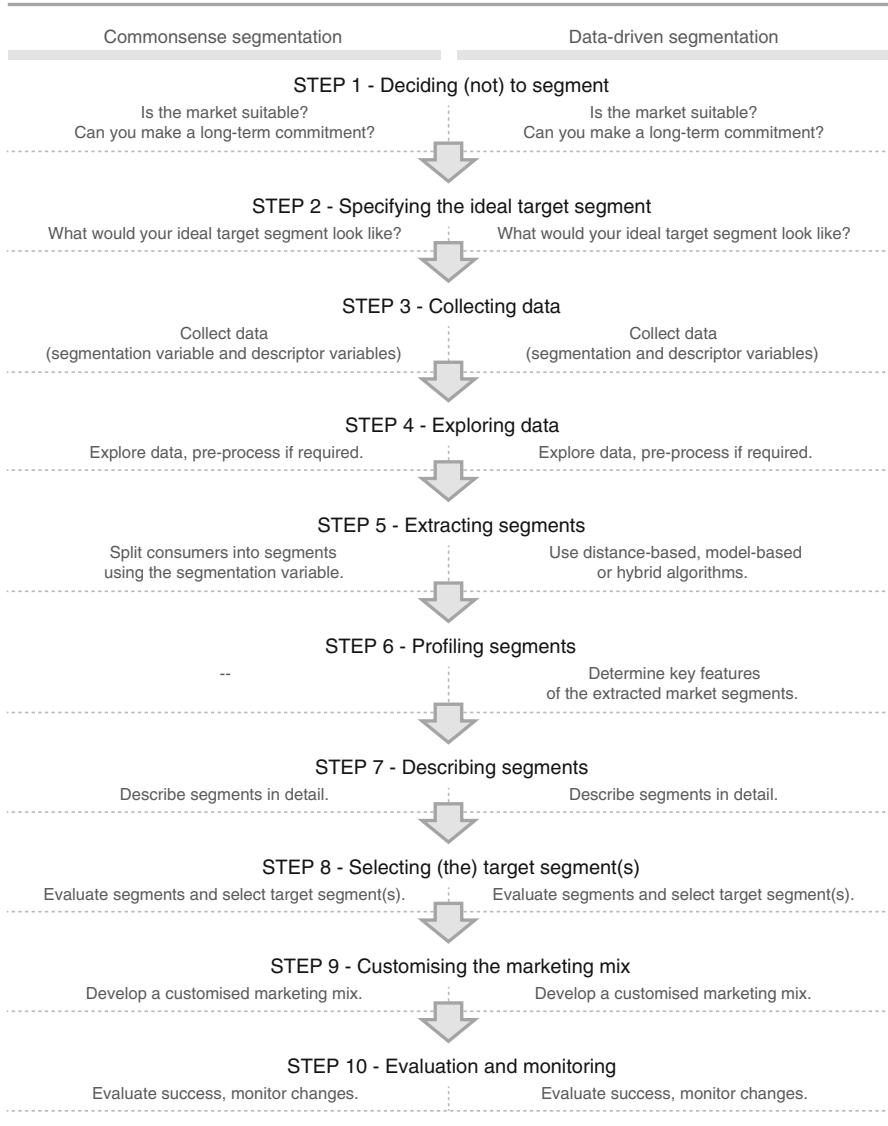


Fig. 2.2 Ten steps of market segmentation analysis

Although the ten steps of market segmentation analysis are the same for commonsense and data-driven segmentation, different tasks need to be completed for each one of those approaches. Typically, data-driven segmentation requires additional decisions to be made. The following chapters discuss each of these steps in detail, and provide tools that can be used to implement each step in practice.

References

- Beane TP, Ennis DM (1987) Market segmentation: a review. *Eur J Mark* 21(5):20–42
- Boksberger PE, Laesser C (2009) Segmentation of the senior travel market by the means of travel motivations. *J Vacat Mark* 15(4):311–322
- Dibb S, Simkin L (2008) Market segmentation success: making it happen! Routledge, New York
- Dolnicar S (2004) Beyond commonsense segmentation: a systematics of segmentation approaches in tourism. *J Travel Res* 42(3):244–250
- Dolnicar S, Leisch F (2010) Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Mark Lett* 21:83–101
- Dubes R, Jain AK (1979) Validity studies in clustering methodologies. *Pattern Recogn* 11(4): 235–254
- Ernst D, Dolnicar S (2018) How to avoid random market segmentation solutions. *J Travel Res.* 57(1): 69–82
- Frank RE, Massy WF, Wind Y (1972) Market segmentation. Prentice-Hall, Englewood Cliffs
- Green PE (1977) A new approach to market segmentation. *Bus Horiz* 20(1):61–73
- Hammond K, Ehrenberg ASC, Goodhardt GJ (1996) Market segmentation for competitive brands. *Eur J Mark* 30(12):39–49
- Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River
- Mazanec J (1997) International city tourism: analysis and strategy. Pinter/Cassell, London
- Mazanec JA (2000) Market segmentation. In: Jafari J (ed) Encyclopedia of tourism. Routledge, London
- Morritt R (2007) Segmentation strategies for hospitality managers: target marketing for competitive advantage. The Haworth Press, New York
- Myers JH, Tauber E (1977) Market structure analysis. American Marketing Association, Chicago
- Wedel M, Kamakura W (2000) Market segmentation – conceptual and methodological foundations. Kluwer Academic, Boston
- Wind Y (1978) Issues and advances in segmentation research. *J Mark Res* 15(3):317–337

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II

Ten Steps of Market Segmentation Analysis

Chapter 3

Step 1: Deciding (not) to Segment



3.1 Implications of Committing to Market Segmentation

Although market segmentation has developed to be a key marketing strategy applied in many organisations, it is not always the best decision to pursue such a strategy. Before investing time and resources in a market segmentation analysis, it is important to understand the implications of pursuing a market segmentation strategy.

The key implication is that the organisation needs to commit to the segmentation strategy on the long term. Market segmentation is a marriage, not a date. The commitment to market segmentation goes hand in hand with the willingness and ability of the organisation to make substantial changes (McDonald and Dunbar 1995) and investments. As Cahill (2006) puts it: Segmenting a market is not free. There are costs of performing the research, fielding surveys, and focus groups, designing multiple packages, and designing multiple advertisements and communication messages (p. 158). Cahill recommends not to segment unless the expected increase in sales is sufficient to justify implementing a segmentation strategy, stating (p. 77) that One of the truisms of segmentation strategy is that using the scheme has to be more profitable than marketing without it, net of the expense of developing and using the scheme itself.

Potentially required changes include the development of new products, the modification of existing products, changes in pricing and distribution channels used to sell the product, as well as all communications with the market. These changes, in turn, are likely to influence the internal structure of the organisation, which may need to be adjusted in view of, for example, targeting a handful of different market segments. Croft (1994) recommends that – to maximise the benefits of market segmentation – organisations need to organise around (p. 66) market

segments, rather than organising around products. Strategic business units in charge of segments offer a suitable organisational structure to ensure ongoing focus on the (changing) needs of market segments.

Because of the major implications of such a long-term organisational commitment, the decision to investigate the potential of a market segmentation strategy must be made at the highest executive level, and must be systematically and continuously communicated and reinforced at all organisational levels and across all organisational units.

3.2 Implementation Barriers

A number of books on market segmentation focus specifically on how market segmentation can be successfully implemented in organisations. These books (among them, Dibb and Simkin 2008; Croft 1994 and McDonald and Dunbar 1995) highlight barriers that can impede the successful roll-out of a market segmentation strategy.

The first group of barriers relates to senior management. Lack of leadership, pro-active championing, commitment and involvement in the market segmentation process by senior leadership undermines the success of market segmentation. As McDonald and Dunbar (1995, p. 158) state: There can be no doubt that unless the chief executive sees the need for a segmentation review, understands the process and shows an active interest in it, it is virtually impossible for a senior marketing executive to implement the conclusions in a meaningful way.

Senior management can also prevent market segmentation to be successfully implemented by not making enough resources available, either for the initial market segmentation analysis itself, or for the long-term implementation of a market segmentation strategy.

A second group of barriers relates to organisational culture. Lack of market or consumer orientation, resistance to change and new ideas, lack of creative thinking, bad communication and lack of sharing of information and insights across organisational units, short-term thinking, unwillingness to make changes and office politics have been identified as preventing the successful implementation of market segmentation (Dibb and Simkin 2008). Croft (1994) developed a short questionnaire to assess the extent to which a lack of market orientation in the organisational culture may represent a barrier to the successful implementation of market segmentation.

Another potential problem is lack of training. If senior management and the team tasked with segmentation do not understand the very foundations of market segmentation, or if they are unaware of the consequences of pursuing such a strategy, the attempt of introducing market segmentation is likely to fail.

Closely linked to these barriers is the lack of a formal marketing function or at least a qualified marketing expert in the organisation. The higher the market diversity and the larger the organisations, the more important is a high degree of formalisation (McDonald and Dunbar 1995, p. 158). The lack of a qualified data manager and analyst in the organisation can also represent major stumbling blocks (Dibb and Simkin 2008).

Another obstacle may be objective restrictions faced by the organisation, including lack of financial resources, or the inability to make the structural changes required. As Beane and Ennis (1987) put it (p. 20): A company with limited resources needs to pick only the best opportunities to pursue. Process-related barriers include not having clarified the objectives of the market segmentation exercise, lack of planning or bad planning, a lack of structured processes to guide the team through all steps of the market segmentation process, a lack of allocation of responsibilities, and time pressure that stands in the way of trying to find the best possible segmentation outcome (Dibb and Simkin 2008; McDonald and Dunbar 1995).

At a more operational level, Doyle and Saunders (1985) note that management science has had a disappointing level of acceptance in industry because management will not use techniques it does not understand (p. 26). One way of counteracting this challenge is to make market segmentation analysis easy to understand, and to present results in a way that facilitates interpretation by managers. This can be achieved by using graphical visualisations (see Steps 6 and 7).

Most of these barriers can be identified from the outset of a market segmentation study, and then proactively removed. If barriers cannot be removed, the option of abandoning the attempt of exploring market segmentation as a potential future strategy should be seriously considered.

If going ahead with the market segmentation analysis, McDonald and Dunbar (1995, p. 164) recommend: Above all, a resolute sense of purpose and dedication is required, tempered by patience and a willingness to appreciate the inevitable problems which will be encountered in implementing the conclusions.

3.3 Step 1 Checklist

This first checklist includes not only tasks, but also a series of questions which, if not answered in the affirmative, serve as knock-out criteria. For example: if an organisation is not market-oriented, even the finest of market segmentation analyses cannot be successfully implemented.

Task	Who is responsible?	Completed?
Ask if the organisation's culture is market-oriented. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if the organisation is genuinely willing to change. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if the organisation takes a long-term perspective. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if the organisation is open to new ideas. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if communication across organisational units is good. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if the organisation is in the position to make significant (structural) changes. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Ask if the organisation has sufficient financial resources to support a market segmentation strategy. If yes, proceed. If no, seriously consider not to proceed.		<input type="checkbox"/>
Secure visible commitment to market segmentation from senior management.		<input type="checkbox"/>
Secure active involvement of senior management in the market segmentation analysis.		<input type="checkbox"/>
Secure required financial commitment from senior management.		<input type="checkbox"/>
Ensure that the market segmentation concept is fully understood. If it is not: conduct training until the market segmentation concept is fully understood.		<input type="checkbox"/>
Ensure that the implications of pursuing a market segmentation strategy are fully understood. If they are not: conduct training until the implications of pursuing a market segmentation strategy are fully understood.		<input type="checkbox"/>
Put together a team of 2-3 people (segmentation team) to conduct the market segmentation analysis.		<input type="checkbox"/>

Task	Who is responsible?	Completed?
Ensure that a marketing expert is on the team.		<input type="checkbox"/>
Ensure that a data expert is on the team.		<input type="checkbox"/>
Ensure that a data analysis expert is on the team.		<input type="checkbox"/>
Set up an advisory committee representing all affected organisational units.		<input type="checkbox"/>
Ensure that the objectives of the market segmentation analysis are clear.		<input type="checkbox"/>
Develop a structured process to follow during market segmentation analysis.		<input type="checkbox"/>
Assign responsibilities to segmentation team members using the structured process.		<input type="checkbox"/>
Ensure that there is enough time to conduct the market segmentation analysis without time pressure.		<input type="checkbox"/>

References

- Beane TP, Ennis DM (1987) Market segmentation: a review. *Eur J Mark* 21(5):20–42
- Cahill DJ (2006) Lifestyle market segmentation. Haworth Press, New York
- Croft M (1994) Market segmentation. Routledge, London
- Dibb S, Simkin L (2008) Market segmentation success: making it happen! Routledge, New York
- Doyle P, Saunders J (1985) Market segmentation and positioning in specialized industrial markets. *J Mark* 49(2):24–32
- McDonald M, Dunbar I (1995) Market segmentation: a step-by-step approach to creating profitable market segments. Macmillan, London

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

Step 2: Specifying the Ideal Target Segment



4.1 Segment Evaluation Criteria

The third layer of market segmentation analysis (illustrated in Fig. 2.1) depends primarily on user input. It is important to understand that – for a market segmentation analysis to produce results that are useful to an organisation – user input cannot be limited to either a briefing at the start of the process, or the development of a marketing mix at the end. Rather, the user needs to be involved in most stages, literally wrapping around the technical aspects of market segmentation analysis.

After having committed to investigating the value of a segmentation strategy in Step 1, the organisation has to make a major contribution to market segmentation analysis in Step 2. While this contribution is conceptual in nature, it guides many of the following steps, most critically Step 3 (data collection) and Step 8 (selecting one or more target segments). In Step 2 the organisation must determine two sets of segment evaluation criteria. One set of evaluation criteria can be referred to as *knock-out criteria*. These criteria are the essential, non-negotiable features of segments that the organisation would consider targeting. The second set of evaluation criteria can be referred to as *attractiveness criteria*. These criteria are used to evaluate the relative attractiveness of the remaining market segments – those in compliance with the knock-out criteria.

The literature does not generally distinguish between these two kinds of criteria. Instead, the literature proposes a wide array of possible segment evaluation criteria and describes them at different levels of detail. Table 4.1 contains a selection of proposed criteria.

Table 4.1 Criteria proposed in the literature for the evaluation of market segments in chronological order. (Modified from Karlsson 2015)

Source	Evaluation criteria
Day (1984)	Measurable, Substantial, Accessible, Sufficiently different, At suitable life-cycle stage
Croft (1994)	Large enough, Growing, Competitively advantageous, Profitable, Likely technological changes, Sensitivity to price, Barriers to entry, Buyer or supplier bargaining power, Socio-political considerations, Cyclical and seasonality, Life-cycle position
Myers (1996)	Large enough, Distinguishable, Accessible, Compatible with company
Wedel and Kamakura (2000)	Identifiable, Substantial, Accessible, Responsive, Stable, Actionable
Perreault Jr and McCarthy (2002)	Substantial, Operational, Heterogeneous between, Homogeneous within
Lilien and Rangaswamy (2003)	Large enough (market potential, current market penetration), Growing (past growth forecasts of technology change), Competitively advantageous (barriers to entry, barriers to exit, position of competitors), Segment saturation (gaps in marketing), Protectable (patentable products, barriers to entry), Environmentally risky (economic, political, and technological change), Fit (coherence with company's strengths and image), Relationships with other segments (synergy, cost interactions, image transfers, cannibalisation), Profitable (entry costs, margin levels, return on investment)
McDonald and Dunbar (2004)	Segment factors (size, growth rate per year, sensitivity to price, service features and external factors, cyclical, seasonality, bargaining power of upstream suppliers), Competition (types of competition, degree of concentration, changes in type and mix, entries and exits, changes in share, substitution by new technology, degrees and type of integration), Financial and economic factors (contribution margins, capacity utilisation, leveraging factors, such as experience and economies of scale, barriers to entry, or exit), Technological factors (maturity and volatility, complexity, differentiation, patents and copyrights, manufacturing processes), Socio-political factors (social attitudes and trends, laws and government agency regulations, influence with pressure groups and government representatives, human factors, such as unionisation and community acceptance)
Dibb and Simkin (2008)	Homogeneous, Large enough, Profitable, Stable, Accessible, Compatible, Actionable
Sternthal and Tybout (2001)	Influence of company's current position in the market on growth opportunities, Competitor's ability and motivation to retaliate, Competence and resources, Segments that will prefer the value that can be created by the firm over current market offerings, Consumer motivation and goals indicating gaps in marketplace offerings when launching a new company
West et al. (2010)	Large enough, Sufficient purchasing power, Characteristics of the segment, Reachable, Able to serve segment effectively, Distinct, Targetable with marketing programs
Solomon et al. (2011)	Differentiable, Measurable, Substantial, Accessible, Actionable

(continued)

Table 4.1 (continued)

Winer and Dhar (2011)	Parsimonious, Large enough, Growing, Competitively advantageous
Jain (2012)	Measurable, Accessible, Substantial, Develops maximum differential in competitive strategy, Preserves competitive advantage, Valid even though imitated
Kotler and Keller (2012)	Measurable, Substantial, Accessible, Differentiable, Actionable, Segment rivalry (competition), Potential entrants, Substitutes, Power of buyers, Power of suppliers, Compatible with company
Pride et al. (2012)	Sales estimates (potential sales for product item, product line, geographical area in the short, medium or long term), Competitive assessment, Cost estimates, Long-term profit opportunities, Financial resources, Managerial skills, Employee expertise, Facilities to compete effectively, Fit with corporate objectives, Legal issues, Conflicts with stakeholders, Technological advances
Sharp (2013)	Measurable, Targetable, Large enough, Profitable

In Sects. 4.2 and 4.3, these criteria are discussed under two separate headings to reflect the difference in nature. The shorter set of knock-out criteria is *essential*. It is not up to the segmentation team to negotiate the extent to which they matter in target segment selection. The second, much longer and much more diverse set of attractiveness criteria represents a shopping list for the segmentation team. Members of the segmentation team need to select which of these criteria they want to use to determine how attractive potential target segments are. The segmentation team also needs to assess the relative importance of each attractiveness criterion to the organisation. Where knock-out criteria automatically eliminate some of the available market segments, attractiveness criteria are first negotiated by the team, and then applied to determine the overall relative attractiveness of each market segment in Step 8.

4.2 Knock-Out Criteria

Knock-out criteria are used to determine if market segments resulting from the market segmentation analysis qualify to be assessed using segment attractiveness criteria. The first set of such criteria was suggested by Kotler (1994) and includes substantiality, measurability and accessibility (Tynan and Drayton 1987). Kotler himself and a number of other authors have since recommended additional criteria that fall into the knock-out criterion category (Wedel and Kamakura 2000; Lilien and Rangaswamy 2003; McDonald and Dunbar 2012):

- The segment must be **homogeneous**; members of the segment must be similar to one another.
- The segment must be **distinct**; members of the segment must be distinctly different from members of other segments.

- The segment must be **large enough**; the segment must contain enough consumers to make it worthwhile to spend extra money on customising the marketing mix for them.
- The segment must be **matching** the strengths of the organisation; the organisation must have the capability to satisfy segment members' needs.
- Members of the segment must be **identifiable**; it must be possible to spot them in the marketplace.
- The segment must be **reachable**; there has to be a way to get in touch with members of the segment in order to make the customised marketing mix accessible to them.

Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee. Most of them do not require further specification, but some do. For example, while size is non-negotiable, the exact minimum viable target segment size needs to be specified.

4.3 Attractiveness Criteria

In addition to the knock-out criteria, Table 4.1 also lists a wide range of segment attractiveness criteria available to the segmentation team to consider when deciding which attractiveness criteria are most useful to their specific situation.

Attractiveness criteria are not binary in nature. Segments are not assessed as either complying or not complying with attractiveness criteria. Rather, each market segment is rated; it can be more or less attractive with respect to a specific criterion. The attractiveness across all criteria determines whether a market segment is selected as a target segment in Step 8 of market segmentation analysis.

4.4 Implementing a Structured Process

There is general agreement in the segmentation literature, that following a structured process when assessing market segments is beneficial (Lilien and Rangaswamy 2003; McDonald and Dunbar 2012).

The most popular structured approach for evaluating market segments in view of selecting them as target markets is the use of a segment evaluation plot (Lilien and Rangaswamy 2003; McDonald and Dunbar 2012) showing segment attractiveness along one axis, and organisational competitiveness on the other axis (for an example

see Fig. 10.1). The segment attractiveness and organisational competitiveness values are determined by the segmentation team. This is necessary because there is no standard set of criteria that could be used by all organisations.

Factors which constitute both segment attractiveness and organisational competitiveness need to be negotiated and agreed upon. To achieve this, a large number of possible criteria has to be investigated before agreement is reached on which criteria are most important for the organisation. McDonald and Dunbar (2012) recommend to use no more than six factors as the basis for calculating these criteria.

Optimally, this task should be completed by a team of people (McDonald and Dunbar 1995; Karlsson 2015). If a core team of two to three people is primarily in charge of market segmentation analysis, this team could propose an initial solution and report their choices to the advisory committee – which consists of representatives of all organisational units – for discussion and possible modification. There are at least two good reasons to include in this process representatives from a wide range of organisational units. First, each organisational unit has a different perspective on the business of the organisation. As a consequence, members of these units bring different positions to the deliberations. Secondly, if the segmentation strategy is implemented, it will affect every single unit of the organisation. Consequently, all units are key stakeholders of market segmentation analysis.

Back to the segment evaluation plot. Obviously the segment evaluation plot cannot be completed in Step 2 of the market segmentation analysis because – at this point – no segments are available to assess yet. But there is a huge benefit in selecting the attractiveness criteria for market segments at this early stage in the process: knowing precisely what it is about market segments that matters to the organisation ensures that all of this information is captured when collecting data (Step 3). It also makes the task of selecting a target segment in Step 8 much easier because the groundwork is laid before the actual segments are on the table.

At the end of this step, the market segmentation team should have a list of approximately six segment attractiveness criteria. Each of these criteria should have a weight attached to it to indicate how important it is to the organisation compared to the other criteria. The typical approach to weighting (Lilien and Rangaswamy 2003; McDonald and Dunbar 2012) is to ask all team members to distribute 100 points across the segmentation criteria. These allocations then have to be negotiated until agreement is reached. Optimally, approval by the advisory committee should be sought because the advisory committee contains representatives from multiple organisational units bringing a range of different perspectives to the challenge of specifying segment attractiveness criteria.

4.5 Step 2 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		<input type="checkbox"/>
Discuss and agree on the knock-out criteria of homogeneity, distinctness, size, match, identifiability and reachability. These knock-out criteria will lead to the automatic elimination of market segments which do not comply (in Step 8 at the latest).		<input type="checkbox"/>
Present the knock-out criteria to the advisory committee for discussion and (if required) adjustment.		<input type="checkbox"/>
Individually study available criteria for the assessment of market segment attractiveness.		<input type="checkbox"/>
Discuss the criteria with the other segmentation team members and agree on a subset of no more than six criteria.		<input type="checkbox"/>
Individually distribute 100 points across the segment attractiveness criteria you have agreed upon with the segmentation team. Distribute them in a way that reflects the relative importance of each attractiveness criterion.		<input type="checkbox"/>
Discuss weightings with other segmentation team members and agree on a weighting.		<input type="checkbox"/>
Present the selected segment attractiveness criteria and the proposed weights assigned to each of them to the advisory committee for discussion and (if required) adjustment.		<input type="checkbox"/>

References

- Croft M (1994) Market segmentation. Routledge, London
- Day GS (1984) Strategic market planning. West Publishing Company, Minnesota
- Dibb S, Simkin L (2008) Market segmentation success: making it happen! Routledge, New York
- Jain SC (2012) Marketing: planning and strategy. Cengage Learning Australia, South Melbourne
- Karlsson L (2015) The impact of checklists on organizational target segment selection. Ph.D. thesis, School of Management, Operations and Marketing, University of Wollongong, Wollongong
- Kotler P (1994) Marketing management, 8th edn. Prentice-Hall, Englewood Cliffs
- Kotler P, Keller KL (2012) Marketing management. Pearson Education, Paris
- 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

- McDonald M, Dunbar I (2004) Market segmentation: how to do it, how to profit from it. Elsevier Butterworth-Heinemann, Oxford
- McDonald M, Dunbar I (2012) Market segmentation: how to do it and how to profit from it, 4th edn. Wiley, Chichester
- Myers JH (1996) Segmentation and positioning for strategic marketing decisions. American Marketing Association, Chicago
- Perreault Jr WD, McCarthy JE (2002) Basic marketing: a global management approach. McGraw-Hill Irwin, New York
- Pride WM, Ferrell OC, Lukas BA, Schembri S, Niininen O (2012) Marketing principles. Cengage Learning Australia, South Melbourne
- Sharp B (2013) Marketing: theory, evidence, practice. Oxford University Press, Oxford
- Solomon MR, Hughes A, Chitty B, Fripp G, Marshall GW, EW S (2011) Marketing 2: real people, real choices. Pearson Education Australia, Frenchs Forest
- Sternthal B, Tybout AM (2001) Segmentation and targeting. In: Iacobucci D (ed) Kellogg on marketing. Wiley-Blackwell, Hoboken
- Tynan AC, Drayton J (1987) Market segmentation. *J Mark Manag* 2(3):301–335
- Wedel M, Kamakura W (2000) Market segmentation – conceptual and methodological foundations. Kluwer Academic, Boston
- West D, Ford J, Ibrahim E (2010) Strategic marketing: creating competitive advantage, 2nd edn. Oxford University Press, Oxford
- Winer R, Dhar R (2011) Marketing management. Pearson, London

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

Step 3: Collecting Data



5.1 Segmentation Variables

Empirical data forms the basis of both commonsense and data-driven market segmentation. Empirical data is used to identify or create market segments and – later in the process – describe these segments in detail.

Throughout this book we use the term *segmentation variable* to refer to the variable in the empirical data used in commonsense segmentation to split the sample into market segments. In commonsense segmentation, the segmentation variable is typically one single characteristic of the consumers in the sample. This case is illustrated in Table 5.1. Each row in this table represents one consumer, each variable represents one characteristic of that consumer. An entry of 1 in the data set indicates that the consumer has that characteristic. An entry of 0 indicates that the consumer does not have that characteristic. The commonsense segmentation illustrated in Table 5.1 uses gender as the segmentation variable. Market segments are created by simply splitting the sample using this segmentation variable into a segment of women and a segment of men.

All the other personal characteristics available in the data – in this case: age, the number of vacations taken, and information about five benefits people seek or do not seek when they go on vacation – serve as so-called *descriptor variables*. They are used to describe the segments in detail. Describing segments is critical to being able to develop an effective marketing mix targeting the segment. Typical descriptor variables include socio-demographics, but also information about media behaviour, allowing marketers to reach their target segment with communication messages.

The difference between commonsense and data-driven market segmentation is that data-driven market segmentation is based not on one, but on multiple segmentation variables. These segmentation variables serve as the starting point for identifying naturally existing, or artificially creating market segments useful to the organisation. An illustration is provided in Table 5.2 using the same data as in Table 5.1.

Table 5.1 Gender as a possible segmentation variable in commonsense market segmentation

Sociodemographics		Travel behaviour		Benefits sought			
gender	age	Nº of vacations	relaxation	action	culture	explore	meet people
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	87	2	1	0	1	0	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
segmentation variable		descriptor variables					

Table 5.2 Segmentation variables in data-driven market segmentation

Sociodemographics		Travel behaviour		Benefits sought			
gender	age	Nº of vacations	relaxation	action	culture	explore	meet people
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Male	87	2	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
descriptor variables				segmentation variables			

In the data-driven case we may, for example, want to extract market segments of tourists who do not necessarily have gender in common, but rather share a common set of benefits they seek when going on vacation. Sorting the data from Table 5.1 using this set of segmentation variables reveals one segment (shown in the first three rows) characterised by seeking relaxation, culture and meeting people, but not interested in action and exploring. In this case, the benefits sought represent the segmentation variables. The socio-demographic variables, gender, age, and the number of vacations undertaken per annum serve as descriptor variables.

These two simple examples illustrate how critical the quality of empirical data is for developing a valid segmentation solution. When commonsense segments are extracted – even if the nature of the segments is known in advance – data quality is critical to both (1) assigning each person in the sample to the correct market segment, and (2) being able to correctly describe the segments. The correct description, in turn, makes it possible to develop a customised product, determine the most appropriate pricing strategy, select the best distribution channel, and the most effective communication channel for advertising and promotion.

The same holds for data-driven market segmentation where data quality determines the quality of the extracted data-driven market segments, and the quality of the descriptions of the resulting segments. Good market segmentation analysis requires good empirical data.

Empirical data for segmentation studies can come from a range of sources: from survey studies; from observations such as scanner data where purchases are recorded and, frequently, are linked to an individual customer's long-term purchase history via loyalty programs; or from experimental studies. Optimally, data used in segmentation studies should reflect consumer behaviour. Survey data – although it arguably represents the most common source of data for market segmentation studies – can be unreliable in reflecting behaviour, especially when the behaviour of interest is socially desirable, such as donating money to a charity or behaving in an environmentally friendly way (Karlsson and Dolnicar 2016). Surveys should therefore not be seen as the default source of data for market segmentation studies. Rather, a range of possible sources should be explored. The source that delivers data most closely reflecting actual consumer behaviour is preferable.

5.2 Segmentation Criteria

Long before segments are extracted, and long before data for segment extraction is collected, the organisation must make an important decision: it must choose which segmentation criterion to use (Tynan and Drayton 1987). The term *segmentation criterion* is used here in a broader sense than the term segmentation variable. The term segmentation variable refers to one measured value, for example, one item in a survey, or one observed expenditure category. The term segmentation criterion relates to the nature of the information used for market segmentation. It can also relate to one specific construct, such as benefits sought.

The decision which segmentation criterion to use cannot easily be outsourced to either a consultant or a data analyst because it requires prior knowledge about the market. The most common segmentation criteria are geographic, socio-demographic, psychographic and behavioural.

Bock and Uncles (2002) argue that the following differences between consumers are the most relevant in terms of market segmentation: profitability, bargaining power, preferences for benefits or products, barriers to choice and consumer interaction effects. With so many different segmentation criteria available, which is the best to use? As Hoek et al. (1996) note, few guidelines as to the most appropriate base to use in a given marketing context exist (p. 26). Generally, the recommendation is to use the simplest possible approach. Cahill (2006) states this very clearly in his book on lifestyle segmentation (p. 159): Do the least you can. If demographic segmentation will work for your product or service, then use demographic segmentation. If geographic segmentation will work because your product will only appeal to people in a certain region, then use it. Just because psychographic segmentation is sexier and more sophisticated than demographic or geographic segmentation does not make it better. Better is what works for your product or service at the least possible cost.

5.2.1 Geographic Segmentation

Geographic information is seen as the original segmentation criterion used for the purpose of market segmentation (Lewis et al. 1995; Tynan and Drayton 1987). Typically – when geographic segmentation is used – the consumer's location of residence serves as the only criterion to form market segments. While simple, the geographic segmentation approach is often the most appropriate. For example: if the national tourism organisation of Austria wants to attract tourists from neighbouring countries, it needs to use a number of different languages: Italian, German, Slovenian, Hungarian, Czech. Language differences across countries represent a very pragmatic reason for treating tourists from different neighbouring countries as different segments. Interesting examples are also provided by global companies such as Amazon selling its Kindle online: one common web page is used for the description of the base product, then customers are asked to indicate their country of residence and country specific additional information is provided. IKEA offers a similar product range worldwide, yet slight differences in offers, pricing as well as the option to purchase online exist in dependence of the customer's geographic location.

The key advantage of geographic segmentation is that each consumer can easily be assigned to a geographic unit. As a consequence, it is easy to target communication messages, and select communication channels (such as local newspapers, local radio and TV stations) to reach the selected geographic segments.

The key disadvantage is that living in the same country or area does not necessarily mean that people share other characteristics relevant to marketers, such

as benefits they seek when purchasing a product. While, for example, people residing in luxury suburbs may all be a good target market for luxury cars, location is rarely the reason for differences in product preference. Even in the case of luxury suburbs, it is more likely that socio-demographic criteria are the reason for both similar choice of suburb to live in and similar car preferences. The typical case is best illustrated using tourism: people from the same country of origin are likely to have a wide range of different ideal holidays, depending on whether they are single or travel as a family, whether they are into sports or culture.

Despite the potential shortcomings of using geographic information as the segmentation variable, the location aspect has experienced a revival in international market segmentation studies aiming to extract market segments across geographic boundaries. Such an approach is challenging because the segmentation variable(s) must be meaningful across all the included geographic regions, and because of the known biases that can occur if surveys are completed by respondents from different cultural backgrounds (Steenkamp and Ter Hofstede 2002). An example of such an international market segmentation study is provided by Haverila (2013) who extracted market segments of mobile phone users among young customers across national borders.

5.2.2 *Socio-Demographic Segmentation*

Typical socio-demographic segmentation criteria include age, gender, income and education. Socio-demographic segments can be very useful in some industries. For example: luxury goods (associated with high income), cosmetics (associated with gender; even in times where men are targeted, the female and male segments are treated distinctly differently), baby products (associated with gender), retirement villages (associated with age), tourism resort products (associated with having small children or not).

As is the case with geographic segmentation, socio-demographic segmentation criteria have the advantage that segment membership can easily be determined for every consumer. In some instances, the socio-demographic criterion may also offer an explanation for specific product preferences (having children, for example, is the actual reason that families choose a family vacation village where previously, as a couple, their vacation choice may have been entirely different). But in many instances, the socio-demographic criterion is not the *cause* for product preferences, thus not providing sufficient market insight for optimal segmentation decisions. Haley (1985) estimates that demographics explain about 5% of the variance in consumer behaviour. Yankelovich and Meer (2006) argue that socio-demographics do not represent a strong basis for market segmentation, suggesting that values, tastes and preferences are more useful because they are more influential in terms of consumers' buying decisions.

5.2.3 *Psychographic Segmentation*

When people are grouped according to psychological criteria, such as their beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product, the term psychographic segmentation is used. Haley (1985) explains that the word psychographics was intended as an umbrella term to cover all measures of the mind (p. 7). Benefit segmentation, which Haley (1968) is credited for, is arguably the most popular kind of psychographic segmentation. Lifestyle segmentation is another popular psychographic segmentation approach (Cahill 2006); it is based on people's activities, opinions and interests.

Psychographic criteria are, by nature, more complex than geographic or socio-demographic criteria because it is difficult to find a single characteristic of a person that will provide insight into the psychographic dimension of interest. As a consequence, most psychographic segmentation studies use a number of segmentation variables, for example: a number of different travel motives, a number of perceived risks when going on vacation.

The psychographic approach has the advantage that it is generally more reflective of the underlying reasons for differences in consumer behaviour. For example, tourists whose primary motivation to go on vacation is to learn about other cultures, have a high likelihood of undertaking a cultural holiday at a destination that has ample cultural treasures for them to explore. Not surprisingly, therefore, travel motives have been frequently used as the basis for data-driven market segmentation in tourism (Bieger and Laesser 2002; Laesser et al. 2006; Boksberger and Laesser 2009). The disadvantage of the psychographic approach is the increased complexity of determining segment memberships for consumers. Also, the power of the psychographic approach depends heavily on the reliability and validity of the empirical measures used to capture the psychographic dimensions of interest.

5.2.4 *Behavioural Segmentation*

Another approach to segment extraction is to search directly for similarities in behaviour or reported behaviour. A wide range of possible behaviours can be used for this purpose, including prior experience with the product, frequency of purchase, amount spent on purchasing the product on each occasion (or across multiple purchase occasions), and information search behaviour. In a comparison of different segmentation criteria used as segmentation variables, behaviours reported by tourists emerged as superior to geographic variables (Moscardo et al. 2001).

The key advantage of behavioural approaches is that – if based on actual behaviour rather than stated behaviour or stated intended behaviour – the very behaviour of interest is used as the basis of segment extraction. As such, behavioural segmentation groups people by the similarity which matters most. Examples of such segmentation analyses are provided by Tsai and Chiu (2004) who use actual

expenses of consumers as segmentation variables, and Heilman and Bowman (2002) who use actual purchase data across product categories. Brand choice behaviour over time has also been used as segmentation variable by several authors (Poulsen 1990; Bockenholt and Langeheine 1996; Ramaswamy 1997, see also Section 7.3.3). Using behavioural data also avoids the need for the development of valid measures for psychological constructs.

But behavioural data is not always readily available, especially if the aim is to include in the segmentation analysis potential customers who have not previously purchased the product, rather than limiting oneself to the study of existing customers of the organisation.

5.3 Data from Survey Studies

Most market segmentation analyses are based on survey data. Survey data is cheap and easy to collect, making it a feasible approach for any organisation. But survey data – as opposed to data obtained from observing actual behaviour – can be contaminated by a wide range of biases. Such biases can, in turn, negatively affect the quality of solutions derived from market segmentation analysis. A few key aspects that need to be considered when using survey data are discussed below.

5.3.1 Choice of Variables

Carefully selecting the variables that are included as segmentation variable in commonsense segmentation, or as segmentation variables in data-driven segmentation, is critical to the quality of the market segmentation solution.

In data-driven segmentation, all variables relevant to the construct captured by the segmentation criterion need to be included. At the same time, unnecessary variables must be avoided. Including unnecessary variables can make questionnaires long and tedious for respondents, which, in turn, causes respondent fatigue. Fatigued respondents tend to provide responses of lower quality (Johnson et al. 1990; Dolnicar and Rossiter 2008). Including unnecessary variables also increases the dimensionality of the segmentation problem without adding relevant information, making the task of extracting market segments unnecessarily difficult for any data analytic technique. The issue of the appropriate ratio of the number of variables and the available sample is discussed later in this chapter. Unnecessary variables included as segmentation variables divert the attention of the segment extraction algorithm away from information critical to the extraction of optimal market segments. Such variables are referred to as *noisy variables* or *masking variables* and have been repeatedly shown to prevent algorithms from identifying the correct segmentation solution (Brusco 2004; Carmone et al. 1999; DeSarbo et al. 1984; DeSarbo and Mahajan 1984; Milligan 1980).

Noisy variables do not contribute any information necessary for the identification of the correct market segments. Instead, their presence makes it more difficult for the algorithm to extract the correct solution. Noisy variables can result from not carefully developing survey questions, or from not carefully selecting segmentation variables from among the available survey items. The problem of noisy variables negatively affecting the segmentation solution can be avoided at the data collection and the variable selection stage of market segmentation analysis.

The recommendation is to ask all necessary and unique questions, while resisting the temptation to include unnecessary or redundant questions. Redundant questions are common in survey research when scale development follows traditional psychometric principles (Nunally 1978), as introduced to marketing most prominently by Churchill (1979). More recently, Rossiter (2002, 2011) has questioned this practice, especially in the context of measuring concrete objects and attributes that are interpreted consistently as meaning the same by respondents. Redundant items are particularly problematic in the context of market segmentation analysis because they interfere substantially with most segment extraction algorithms' ability to identify correct market segmentation solutions (Dolnicar et al. 2016).

Developing a good questionnaire typically requires conducting exploratory or qualitative research. Exploratory research offers insights about people's beliefs that survey research cannot offer. These insights can then be categorised and included in a questionnaire as a list of answer options. Such a two-stage process involving both qualitative, exploratory and quantitative survey research ensures that no critically important variables are omitted.

5.3.2 *Response Options*

Answer options provided to respondents in surveys determine the scale of the data available for subsequent analyses. Because many data analytic techniques are based on distance measures, not all survey response options are equally suitable for segmentation analysis.

Options allowing respondents to answer in only one of two ways, generate *binary* or *dichotomous data*. Such responses can be represented in a data set by 0s and 1s. The distance between 0 and 1 is clearly defined and, as such, poses no difficulties for subsequent segmentation analysis. Options allowing respondents to select an answer from a range of unordered categories correspond to *nominal variables*. If asked about their occupation, respondents can select only one option from a list of unordered options. Nominal variables can be transformed into binary data by introducing a binary variable for each of the answer options.

Options allowing respondents to indicate a number, such as age or nights stayed at a hotel, generate *metric data*. Metric data allow any statistical procedure to be performed (including the measurement of distance), and are therefore well suited for segmentation analysis. The most commonly used response option in survey research, however, is a limited number of ordered answer options larger than two.

Respondents are asked, for example, to express – using five or seven response options – their agreement with a series of statements. This answer format generates *ordinal data*, meaning that the options are ordered. But the distance between adjacent answer options is not clearly defined. As a consequence, it is not possible to apply standard distance measures to such data, unless strong assumptions are made. Step 5 provides a detailed discussion of suitable distance measures for each scale level.

Preferably, therefore, either metric or binary response options should be provided to respondents if those options are meaningful with respect to the question asked. Using binary or metric response options prevents subsequent complications relating to the distance measure in the process of data-driven segmentation analysis. Although ordinal scales dominate both market research and academic survey research, using binary or metric response options instead is usually not a compromise. If, for example, there is a strong reason to believe that very fine nuances of responses need to be captured, and if capturing those fine nuances does not come at the cost of also capturing response styles, this can be achieved using visual analogue scales. The visual analogue scale allows respondents to indicate a position along a continuous line between two end-points, and leads to data that can be assumed to be metric. The visual analogue scale has experienced a revival with the popularity of online survey research, where it is frequently used and referred to as a slider scale. In many contexts, binary response options have been shown to outperform ordinal answer options (Dolnicar 2003; Dolnicar et al. 2011, 2012), especially when formulated in a level free way (see the discussion of the doubly level free answer format with individually inferred thresholds, or DLF IIIST, in Rossiter et al. 2010; Rossiter 2011; Dolnicar and Grün 2013).

5.3.3 *Response Styles*

Survey data is prone to capturing biases. A response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (i.e., what the items were designed to measure) (Paulhus 1991, p. 17). If a bias is displayed by a respondent consistently over time, and independently of the survey questions asked, it represents a response style.

A wide range of response styles manifest in survey answers, including respondents' tendencies to use extreme answer options (STRONGLY AGREE, STRONGLY DISAGREE), to use the midpoint (NEITHER AGREE NOR DISAGREE), and to agree with all statements. Response styles affect segmentation results because commonly used segment extraction algorithms cannot differentiate between a data entry reflecting the respondent's belief from a data entry reflecting both a respondent's belief and a response style. For example, some respondents displaying an acquiescence bias (a tendency to agree with all questions) could result in one market segment having much higher than average agreement with all answers. Such a segment could be misinterpreted. Imagine a market segmentation based on responses to a series of

questions asking tourists to indicate whether or not they spent money on certain aspects of their vacation, including DINING OUT, VISITING THEME PARKS, USING PUBLIC TRANSPORT, etc. A market segment saying YES to all those items would, no doubt, appear to be highly attractive for a tourist destination holding the promise of the existence of a high-spending tourist segment. It could equally well just reflect a response style. It is critical, therefore, to minimise the risk of capturing response styles when data is collected for the purpose of market segmentation. In cases where attractive market segments emerge with response patterns potentially caused by a response style, additional analyses are required to exclude this possibility. Alternatively, respondents affected by such a response style must be removed before choosing to target such a market segment.

5.3.4 Sample Size

Many statistical analyses are accompanied by sample size recommendations. Not so market segmentation analysis. Figure 5.1 illustrates the problem any segmentation algorithm faces if the sample is insufficient. The market segmentation problem in this figure is extremely simple because only two segmentation variables are used. Yet, when the sample size is insufficient (left plot), it is impossible to determine which the correct number of market segments is. If the sample size is sufficient, however (right plot) it is very easy to determine the number and nature of segments in the data set.

Only a small number of studies have investigated this problem. Viennese psychologist Formann (1984) recommends that the sample size should be at least 2^p (better five times 2^p), where p is the number of segmentation variables. This rule of thumb relates to the specific purpose of goodness-of-fit testing in the context of latent class analysis when using binary variables. It can therefore not be assumed

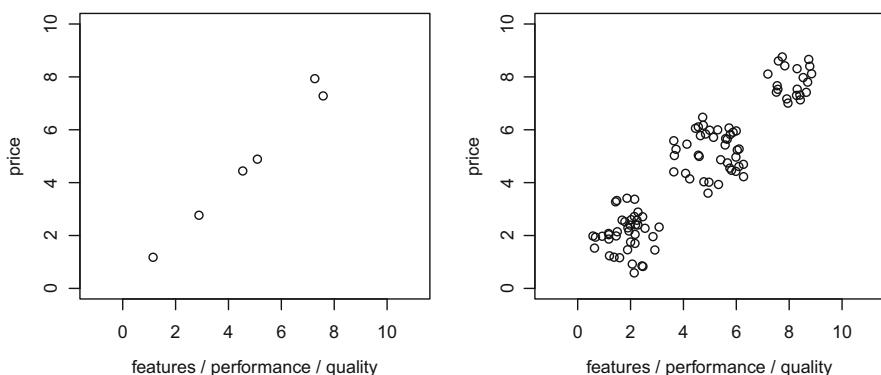


Fig. 5.1 Illustrating the importance of sufficient sample size in market segmentation analysis

to be generalisable to other algorithms, inference methods, and scales. Qiu and Joe (2015) developed a sample size recommendation for constructing artificial data sets for studying the performance of clustering algorithms. According to Qiu and Joe (2015), the sample size should – in the simple case of equal cluster sizes – be at least ten times the number of segmentation variables times the number of segments in the data ($10 \cdot p \cdot k$ where p represents the number of segmentation variables and k represents the number of segments). If segments are unequally sized, the smallest segment should contain a sample of at least $10 \cdot p$.

Dolnicar et al. (2014) conducted extensive simulation studies with artificial data modelled after typical data sets used in applied tourism segmentation studies. Knowing the true structure of the data sets, they tested sample size requirement for algorithms to correctly identify the true segments. Figure 5.2 shows the effect of sample size on the correctness of segment recovery for this particular study. The adjusted Rand index serves as the measure of correctness of segment recovery. The adjusted Rand index assesses the congruence between two segmentation solutions. Higher values indicate better alignment. Its maximum possible value is 1. The expected value is 0 if the two segmentation solutions are derived independently in a random way. To assess segment recovery, the adjusted Rand index is calculated for the true segment solution and the extracted one.

In Fig. 5.2, the x -axis plots the sample size (ranging from 10 to 100 times the number of segmentation variables). The y -axis plots the effect of an increase in sample size on the adjusted Rand index. The higher the effect, the better the algorithm identified the correct market segmentation solution.

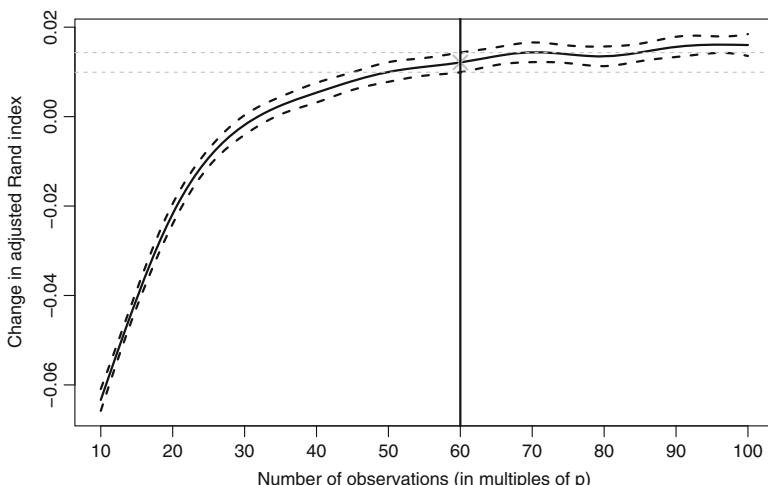


Fig. 5.2 Effect of sample size on the correctness of segment recovery in artificial data. (Modified from Dolnicar et al. 2014)

Not surprisingly, increasing the sample size improves the correctness of the extracted segments. Interestingly, however, the biggest improvement is achieved by increasing very small samples. As the sample size increases, the marginal benefit of further increasing the sample size decreases. Based on the results shown in Fig. 5.2, a sample size of at least $60 \cdot p$ is recommended. For a more difficult artificial data scenario Dolnicar et al. (2014) recommend using a sample size of at least $70 \cdot p$; no substantial improvements in identifying the correct segments were identified beyond this point.

Dolnicar et al. (2016) extended this line of research to account for key features of typical survey data sets, making it more difficult for segmentation algorithms to identify correct segmentation solutions. Specifically, they investigated the effect on sample size requirements resulting from market characteristics not under the control of the data analyst and, data characteristics – at least to some degree – under the control of the data analyst.

Market characteristics studied included: the number of market segments present in the data, whether those market segments are equal or unequal in size, and the extent to which market segments overlap. De Craen et al. (2006) show that the presence of unequally sized segments makes it more difficult for an algorithm to extract the correct market segments. Steinley (2003) shows the same for the case of overlapping segments.

In addition, some of the characteristics of survey data discussed above have been shown to affect segment recovery, specifically: sampling error, response biases and response styles, low data quality, different response options, the inclusion of irrelevant items, and correlation between blocks of items. Figure 5.3 shows the

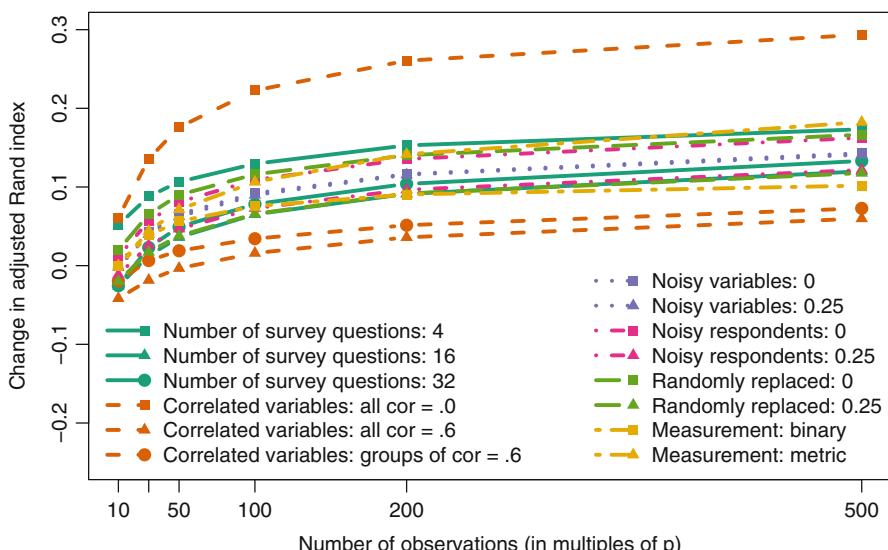


Fig. 5.3 Sample size requirements in dependence of market and data characteristics. (Modified from Dolnicar et al. 2016)

results from this large-scale simulation study using artificial data. Again, the axes plot the sample size, and the effect of increasing sample size on the adjusted Rand index, respectively.

As can be seen in Fig. 5.3, larger sample sizes always improve an algorithm's ability to identify the correct market segmentation solution. The extent to which this is the case, however, varies substantially across market and data characteristics. Also, some of the challenging market and data characteristics can be compensated by increasing sample size; others cannot. For example, using uncorrelated segmentation variables leads to very good segment recovery. But, correlation cannot be well compensated for by increasing sample size, as can be seen in Fig. 5.3: the top-most and the two bottom-most curves in Fig. 5.3 show three different levels of correlation between segmentation variables. If the variables are not correlated at all, the algorithm has no difficulty extracting the correct segments. If, however, the variables are highly correlated, the task becomes so difficult for the algorithm, that even increasing the sample size dramatically does not help. A small number of noisy variables, on the other hand, has a lower effect.

Overall, this study demonstrates the importance of having a sample size sufficiently large to enable an algorithm to extract the correct segments (if segments naturally exist in the data). The recommendation by Dolnicar et al. (2016) is to ensure the data contains at least 100 respondents for each segmentation variable. Results from this study also highlight the importance of collecting high-quality unbiased data as the basis for market segmentation analysis.

It can be concluded from the body of work studying the effects of survey data quality on the quality of market segmentation results based on such data that, optimally, data used in market segmentation analyses should

- contain all necessary items;
- contain no unnecessary items;
- contain no correlated items;
- contain high-quality responses;
- be binary or metric;
- be free of response styles;
- include responses from a suitable sample given the aim of the segmentation study; and
- include a sufficient sample size given the number of segmentation variables (100 times the number of segmentation variables).

5.4 Data from Internal Sources

Increasingly organisations have access to substantial amounts of internal data that can be harvested for the purpose of market segmentation analysis. Typical examples are scanner data available to grocery stores, booking data available through airline loyalty programs, and online purchase data. The strength of such data lies in the

fact that they represent *actual* behaviour of consumers, rather than statements of consumers about their behaviour or intentions, known to be affected by imperfect memory (Niemi 1993), as well as a range of response biases, such as social desirability bias (Fisher 1993; Paulhus 2002; Karlsson and Dolnicar 2016) or other response styles (Paulhus 1991; Dolnicar and Grün 2007a,b, 2009).

Another advantage is that such data are usually automatically generated and – if organisations are capable of storing data in a format that makes them easy to access – no extra effort is required to collect data.

The danger of using internal data is that it may be systematically biased by over-representing existing customers. What is missing is information about other consumers the organisation may want to win as customers in future, which may differ systematically from current customers in their consumption patterns.

5.5 Data from Experimental Studies

Another possible source of data that can form the basis of market segmentation analysis is experimental data. Experimental data can result from field or laboratory experiments. For example, they can be the result of tests how people respond to certain advertisements. The response to the advertisement could then be used as a segmentation criterion. Experimental data can also result from choice experiments or conjoint analyses. The aim of such studies is to present consumers with carefully developed stimuli consisting of specific levels of specific product attributes. Consumers then indicate which of the products – characterised by different combinations of attribute levels – they prefer. Conjoint studies and choice experiments result in information about the extent to which each attribute and attribute level affects choice. This information can also be used as a segmentation criterion.

5.6 Step 3 Checklist

Task	Who is responsible?	Completed?
Convene a market segmentation team meeting.		<input type="checkbox"/>
Discuss which consumer characteristics could serve as promising segmentation variables. These variables will be used to extract groups of consumers from the data.		<input type="checkbox"/>
Discuss which other consumer characteristics are required to develop a good understanding of market segments. These variables will later be used to describe the segments in detail.		<input type="checkbox"/>
Determine how you can collect data to most validly capture both the segmentation variables and the descriptor variables.		<input type="checkbox"/>
Design data collection carefully to keep data contamination through biases and other sources of systematic error to a minimum.		<input type="checkbox"/>
Collect data.		<input type="checkbox"/>

References

- Bieger T, Laesser C (2002) Market segmentation by motivation: the case of Switzerland. *J Travel Res* 41(1):68–76
- Bock T, Uncles M (2002) A taxonomy of differences between consumers for market segmentation. *Int J Res Mark* 19(3):215–224
- Bockenholt U, Langeheine UR (1996) Latent change in recurrent choice data. *Psychometrika* 61(2):285–301
- Boksberger PE, Laesser C (2009) Segmentation of the senior travel market by the means of travel motivations. *J Vacat Mark* 15(4):311–322
- Brusco M (2004) Clustering binary data in the presence of masking variables. *Psychol Methods* 9(4):510–523
- Cahill DJ (2006) Lifestyle market segmentation. Haworth Press, New York
- Carmone FJ, Kara A, Maxwell S (1999) HINoV: a new model to improve market segment definition by identifying noisy variables. *J Mark Res* 36(4):501–509
- Churchill JGA (1979) A paradigm for developing better measures of marketing constructs. *J Mark Res* 16:64–73
- De Craen S, Commandeur JJF, Frank LE, Heiser WJ (2006) Effects of group size and lack of sphericity on the recovery of clusters in k-means cluster analysis. *Multivar Behav Res* 41(2):127–145
- DeSarbo W, Mahajan V (1984) Constrained classification: the use of a priori information in cluster analysis. *Psychometrika* 49(2):187–215

- DeSarbo WS, Carroll JD, Clark LA, Green PE (1984) Synthesized clustering: a method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika* 49(1):57–78
- Dolnicar S (2003) Simplifying three-way questionnaires – do the advantages of binary answer categories compensate for the loss of information? In: ANZMAC CD proceedings, Australia and New Zealand marketing academy (ANZMAC)
- Dolnicar S, Grün B (2007a) Assessing analytical robustness in cross-cultural comparisons. *Int J Tour Cult Hosp Res* 1(2):140–160
- Dolnicar S, Grün B (2007b) Cross-cultural differences in survey response patterns. *Int Mark Rev* 24(2):127–143
- Dolnicar S, Grün B (2009) Response style contamination of student evaluation data. *J Mark Educ* 31(2):160–172
- Dolnicar S, Grün B (2013) Validly measuring destination images in survey studies. *J Travel Res* 52(1):3–13
- Dolnicar S, Rossiter JR (2008) The low stability of brand-attribute associations is partly due to measurement factors. *Int J Res Mark* 25(2):104–108
- Dolnicar S, Grün B, Leisch F (2011) Quick, simple and reliable: forced binary survey questions. *Int J Mark Res* 53(2):231–252
- Dolnicar S, Rossiter JR, Grün B (2012) “Pick-any” measures contaminate brand image studies. *Int J Mark Res* 54(6):821–834
- Dolnicar S, Grün B, Leisch F, Schmidt K (2014) Required sample sizes for data-driven market segmentation analyses in tourism. *J Travel Res* 53(3):296–306
- Dolnicar S, Grün B, Leisch F (2016) Increasing sample size compensates for data problems in segmentation studies. *J Bus Res* 69:992–999
- Fisher RJ (1993) Social desirability bias and the validity of indirect questioning. *J Consum Res* 20:303–315
- Formann A (1984) Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung. Beltz, Weinheim
- Haley R (1968) Benefit segmentation: a decision-oriented tool. *J Mark* 32:30–35
- Haley RI (1985) Developing effective communications strategy – a benefit segmentation approach. Wiley, New York
- Haverila MJ (2013) Market segmentation in the cell phone market among adolescents and young adults. *Asia Pac J Mark Logist* 25(3):346–368
- Heilman CM, Bowman D (2002) Segmenting consumers using multiple-category purchase data. *Int J Res Mark* 19(3):225–252
- Hoek J, Gendall P, Esslemont D (1996) Market segmentation: a search for the holy grail? *J Mark Pract Appl Mark Sci* 2(1):25–34
- Johnson M, Lehmann D, Horne D (1990) The effects of fatigue on judgments of interproduct similarity. *Int J Res Mark* 7(1):35–43
- Karlsson L, Dolnicar S (2016) Does eco certification sell tourism services? Evidence from a quasi-experimental observation study in Iceland. *J Sustain Tour* 24(5):604–714
- Laesser C, Crouch GI, Beritelli P (2006) Market segmentation by reasons and influences to visit a destination: the case of international visitors to Australia. *Tour Anal* 11(4):241–249
- Lewis RC, Chambers RE, Chacko HE (1995) Marketing leadership in hospitality: foundations and practices. van Nostrand Reinhold, New York
- Milligan GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* 45(3):325–342
- Moscardo G, Pearce P, Morrison A (2001) Evaluating different bases for market segmentation. *J Travel Tour Mark* 10(1):29–49
- Niemi I (1993) Systematic error in behavioural measurement: comparing results from interview and time budget studies. *Soc Indic Res* 30(2–3):229–244
- Nunally J (1978) Psychometric theory, 3rd edn. McGraw Hill, New York

- Paulhus D (1991) Measurement and control of response bias. In: Robinson J, Shaver P, Wrightsman L (eds) Measures of personality and social psychological attitudes. Academic Press, San Diego, pp 17–59
- Paulhus DL (2002) Socially desirable responding: the evolution of a construct. In: Braun HI, Jackson DN, Wiley DE (eds) The role of constructs in psychological and educational measurement, chap 4. Routledge, Lawrence Erlbaum Associates Mahwah, New Jersey, London pp 51–76
- Poulsen CS (1990) Mixed Markov and latent Markov modelling applied to brand choice behaviour. *Int J Res Mark* 7:5–19
- Qiu W, Joe H (2015) clusterGeneration: random cluster generation (with specified degree of separation). <https://CRAN.R-project.org/package=clusterGeneration>, R package version 1.3.4
- Ramaswamy V (1997) Evolutionary preference segmentation with panel survey data: an application to new products. *Int J Res Mark* 14(1):57–89
- Rossiter JR (2002) The C-OAR-SE procedure for scale development in marketing. *Int J Res Mark* 19(4):305–335
- Rossiter JR (2011) Measurement for the social sciences. The C-OAR-SE method and why it must replace psychometrics. Springer, New York
- Rossiter JR, Dolnicar S, Grün B (2010) The LFFB comparative judgment measure of brand attribute beliefs. Working Paper, Faculty of Commerce, University of Wollongong
- Steenkamp JBEM, Ter Hofstede F (2002) International market segmentation: issues and perspectives. *Int J Res Mark* 19(3):185–213
- Steinley D (2003) Local optima in k-means clustering: what you don't know may hurt you. *Psychol Methods* 8(3):294–304
- Tsai CY, Chiu CC (2004) A purchase-based market segmentation methodology. *Exp Syst Appl* 27(2):265–276
- Tynan AC, Drayton J (1987) Market segmentation. *J Mark Manag* 2(3):301–335
- Yankelovich D, Meer D (2006) Rediscovering market segmentation. *Harv Bus Rev* 84(2):122–131

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Step 4: Exploring Data



6.1 A First Glimpse at the Data

After data collection, exploratory data analysis cleans and – if necessary – pre-processes the data. This exploration stage also offers guidance on the most suitable algorithm for extracting meaningful market segments.

At a more technical level, data exploration helps to (1) identify the measurement levels of the variables; (2) investigate the univariate distributions of each of the variables; and (3) assess dependency structures between variables. In addition, data may need to be pre-processed and prepared so it can be used as input for different segmentation algorithms. Results from the data exploration stage provide insights into the suitability of different segmentation methods for extracting market segments.

To illustrate data exploration using real data, we use a travel motives data set. This data set contains 20 travel motives reported by 1000 Australian residents in relation to their last vacation. One example of such a travel motive is: I AM INTERESTED IN THE LIFE STYLE OF LOCAL PEOPLE. Detailed information about the data is provided in Appendix C.4. A comma-separated values (CSV) file of the data is contained in the R package **MSA** and can be copied to the current working directory using the command

```
R> vaccsv <- system.file("csv/vacation.csv",
+   package = "MSA")
R> file.copy(vaccsv, ".")
```

Alternatively, the CSV file can be downloaded from the web page of the book (<http://www.MarketSegmentationAnalysis.org>). The CSV file can be explored with a spreadsheet program before commencing analyses in R.

To read the data set into R, we use the following command:

```
R> vac <- read.csv("vacation.csv", check.names = FALSE)
```

`check.names = FALSE` prevents `read.csv()` to convert blanks in column names to dots (which is the default). After reading the data set into R, we store it in a data frame named `vac`.

We can inspect the the `vac` object, and learn about column names, and the size of the data set using the commands:

```
R> colnames(vac)
```

```
[1] "Gender"
[2] "Age"
[3] "Education"
[4] "Income"
[5] "Income2"
[6] "Occupation"
[7] "State"
[8] "Relationship.Status"
[9] "Obligation"
[10] "Obligation2"
[11] "NEP"
[12] "Vacation.Behaviour"
[13] "rest and relax"
[14] "luxury / be spoilt"
[15] "do sports"
[16] "excitement, a challenge"
[17] "not exceed planned budget"
[18] "realise creativity"
[19] "fun and entertainment"
[20] "good company"
[21] "health and beauty"
[22] "free-and-easy-going"
[23] "entertainment facilities"
[24] "not care about prices"
[25] "life style of the local people"
[26] "intense experience of nature"
[27] "cosiness/familiar atmosphere"
[28] "maintain unspoilt surroundings"
[29] "everything organised"
[30] "unspoilt nature/natural landscape"
[31] "cultural offers"
[32] "change of surroundings"
```

```
R> dim(vac)
```

```
[1] 1000   32
```

`summary(vac)` generates a full summary of the data set. Below we select only four columns to show `Gender` (column 1 of the data set), `Age` (column 2), `Income` (column 4), and `Income2` (column 5).

```
R> summary(vac[, c(1, 2, 4, 5)])
      Gender          Age           Income
Female:488   Min.   : 18.00   $30,001 to $60,000 :265
Male  :512    1st Qu.: 32.00   $60,001 to $90,000 :233
              Median : 42.00   Less than $30,000  :150
              Mean   : 44.17   $90,001 to $120,000 :146
              3rd Qu.: 57.00   $120,001 to $150,000: 72
              Max.   :105.00   (Other)            : 68
                           NA's                : 66

Income2
<30k   :150
>120k  :140
30-60k :265
60-90k :233
90-120k:146
NA's   : 66
```

As can be seen from this summary, the Australian travel motives data set contains answers from 488 women and 512 men. The age of the respondents is a metric variable summarised by the minimum value (Min.), the first quartile (1st Qu.), the median, the mean, the third quartile (3rd Qu.), and the maximum (Max.). The youngest respondent is 18, and the oldest 105 years old. Half of the respondents are between 32 and 57 years old. The summary also indicates that the Australian travel motives data set contains two income variables: Income2 consists of fewer categories than Income. Income2 represents a transformation of Income where high income categories (which occur less frequently) have been merged. The summary of the variables Income and Income2 indicates that these variables contain missing data. This means that not all respondents provided information about their income in the survey. Missing values are coded as NAs in R. NA stands for “not available”. The summary shows that 66 respondents did not provide income information.

6.2 Data Cleaning

The first step before commencing data analysis is to clean the data. This includes checking if all values have been recorded correctly, and if consistent labels for the levels of categorical variables have been used. For many metric variables, the range of plausible values is known in advance. For example, age (in years) can be expected to lie between 0 and 110. It is easy to check whether any implausible values are contained in the data, which might point to errors during data collection or data entry.

Similarly, levels of categorical variables can be checked to ensure they contain only permissible values. For example, gender typically has two values in surveys: female and male. Unless the questionnaire did offer a third option, only those two should appear in the data. Any other values are not permissible, and need to be corrected as part of the data cleaning procedure.

Returning to the Australian travel motives data set, the summary for the variables `Gender` and `Age` indicates that no data cleaning is required for these variables. The summary of the variable `Income2` reveals that the categories are not sorted in order. This is a consequence of how data is read into R. R functions like `read.csv()` or `read.table()` convert columns containing information other than numbers into factors. Factors are the default format for storing categorical variables in R. The possible categories of these variables are called levels. By default, levels of factors are sorted alphabetically. This explains the counter-intuitive ordering of the income variable in the Australian travel motives data set. The categories can be re-ordered. One way to achieve this is to copy the column to a helper variable `inc2`, store its levels in `lev`, find the correct re-ordering of the levels, and then convert the variable into an ordinal variable (an ordered factor in R):

```
R> inc2 <- vac$Income2
R> levels(inc2)
[1] "<30k"      ">120k"     "30-60k"    "60-90k"    "90-120k"

R> lev <- levels(inc2)
R> lev
[1] "<30k"      ">120k"     "30-60k"    "60-90k"    "90-120k"

R> lev[c(1, 3, 4, 5, 2)]
[1] "<30k"      "30-60k"    "60-90k"    "90-120k"   ">120k"

R> inc2 <- factor(inc2, levels = lev[c(1, 3, 4, 5, 2)],
+   ordered = TRUE)
```

Before overwriting the – oddly ordered – column of the original data set, we double-check that the transformation was implemented correctly. An easy way to do this is to cross-tabulate the original column with the new, re-ordered version:

```
R> table(orig = vac$Income2, new = inc2)
```

	new				
orig	<30k	30-60k	60-90k	90-120k	>120k
<30k	150	0	0	0	0
>120k	0	0	0	0	140
30-60k	0	265	0	0	0
60-90k	0	0	233	0	0
90-120k	0	0	0	146	0

As can be seen, all row values in this cross-tabulation have exactly one corresponding column value, and the names coincide. It can be concluded that no errors were introduced during re-ordering, and the original column of the data set can safely be overwritten:

```
R> vac$Income2 <- inc2
```

We can re-order variable `Income` in the same way. We keep all R code relating to data transformations to ensure that every step of data cleaning, exploration, and analysis can be reproduced in future. Reproducibility is important from a documentation point of view, and enables other data analysts to replicate the analysis. In addition, it enables the use of the exact same procedure when new data is added on a continuous basis or in regular intervals, as is the case when we monitor segmentation solutions on an ongoing basis (see Step 10). Cleaning data using code (as opposed to clicking in a spreadsheet), requires time and discipline, but makes all steps fully documented and reproducible. After cleaning the data set, we save the corresponding data frame using function `save()`. We can easily re-load this data frame in future R work sessions using function `load()`.

6.3 Descriptive Analysis

Being familiar with the data avoids misinterpretation of results from complex analyses. Descriptive numeric and graphic representations provide insights into the data. Statistical software packages offer a wide variety of tools for descriptive analysis. In R, we obtain a numeric summary of the data with command `summary()`. This command returns the range, the quartiles, and the mean for numeric variables. For categorical variables, the command returns frequency counts. The command also returns the number of missing values for each variable.

Helpful graphical methods for numeric data are histograms, boxplots and scatter plots. Bar plots of frequency counts are useful for the visualisation of categorical variables. Mosaic plots illustrate the association of multiple categorical variables. We explain mosaic plots in Step 7 where we use them to compare market segments.

Histograms visualise the distribution of numeric variables. They show how often observations within a certain value range occur. Histograms reveal if the distribution of a variable is unimodal and symmetric or skewed. To obtain a histogram, we first need to create categories of values. We call this binning. The bins must cover the entire range of observations, and must be adjacent to one another. Usually, they are of equal length. Once we have created the bins, we plot how many of the observations fall into each bin using one bar for each bin. We plot the bin range on the x -axis, and the frequency of observations in each bin on the y -axis.

A number of R packages can construct histograms. We use package `lattice` (Sarkar 2008) because it enables us to create histograms by segments in Step 7. We can construct a histogram for variable `AGE` using:

```
R> library("lattice")
R> histogram(~ Age, data = vac)
```

The left plot in Fig. 6.1 shows the resulting histogram.

By default, this command automatically creates bins. We can gain a deeper understanding of the data by inspecting histograms for different bin widths by specifying the number of bins using the argument `breaks`:

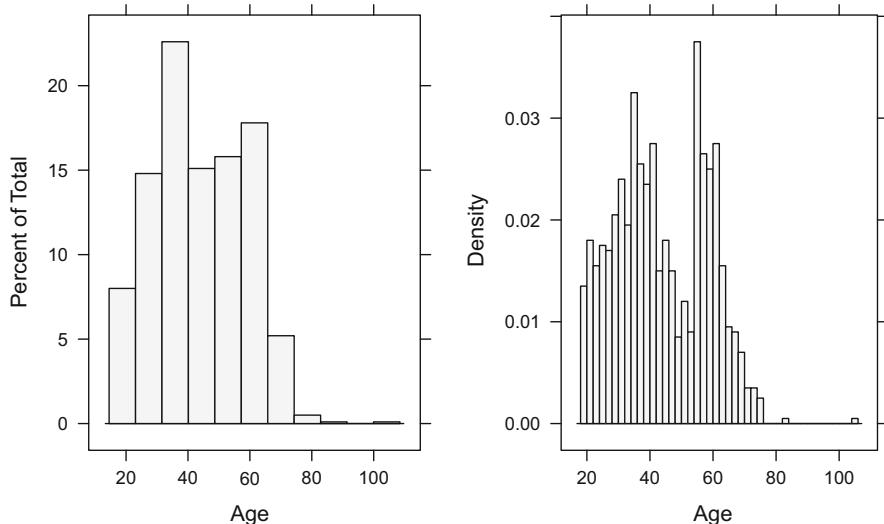


Fig. 6.1 Histograms of tourists' age in the Australian travel motives data set

```
R> histogram(~ Age, data = vac, breaks = 50,
+   type = "density")
```

This command leads to finer bins, as shown in the right plot of Fig. 6.1. The finer bins are more informative, revealing that the distribution is bi-modal with many respondents aged around 35–40 and around 60 years.

Argument `type = "density"` rescales the y-axis to display density estimates. The sum of the areas of all bars in this plot adds up to 1. Plotting density estimates allows us to superimpose probability density functions of parametric distributions. This scaling is in general viewed as the default representation for a histogram.

We can avoid selecting bin widths by using the *box-and-whisker* plot or boxplot (Tukey 1977). The boxplot is the most common graphical visualisation of unimodal distributions in statistics. It is widely used in the natural sciences, but does not enjoy the same popularity in business, and the social sciences more generally. The simplest version of a boxplot compresses a data set into minimum, first quartile, median, third quartile and maximum. These five numbers are referred to as the *five number summary*. R uses the five number summary, and the mean by default to create a numeric summary of a metric variable:

```
R> summary(vac$Age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	32.00	42.00	44.17	57.00	105.00

As can be seen from the output generated by this command, the youngest survey participant in the Australian travel motives study is 18 years old. One quarter of respondents are younger than 32; half of the respondents are younger than 42; and

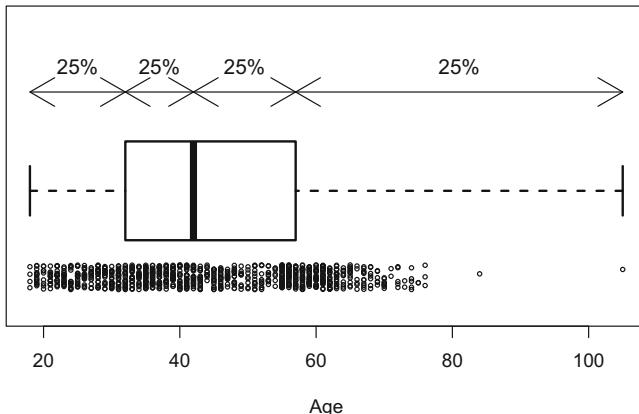


Fig. 6.2 Construction principles for box-and-whisker plots (tourists' age distribution)

three quarters of respondents are younger than 57. The oldest survey respondent is either an astonishing 105 years old, or has made a mistake when completing the survey. The minimum, first quartile, median, third quartile, and maximum are used to generate the boxplot. An illustration of how this is done is provided in Fig. 6.2.

The box-and-whisker plot itself is shown in the middle row of Fig. 6.2. The bottom row plots actual respondent values. Each respondent is represented by a small circle. The circles are jittered randomly in y -axis direction to avoid overplotting in regions of high density. The top row shows the quartiles. The inner box of the box-and-whisker plot extends from the first quartile at 32 to the third quartile at 57. The median is at 42 and depicted by a thick line in the middle of the box. The inner box contains half of the respondents. The whiskers mark the smallest and largest values observed among the respondents, respectively.

Such a simple box-and-whisker plot provides insight into several distributional properties of the sample assuming unimodality. For the Australian travel motives data set, the boxplot shows that the data is right skewed with respect to age because the median is not in the middle of the box but located more to the left. A symmetric distribution would have the median located in the middle of the inner box.

As can also be seen from Fig. 6.2, the 105-year old respondent is solely responsible for the whisker reaching all the way to a value of 105. This, obviously is not an optimal representation of the data, given most other respondents are 70 or younger. The 105-year old respondent is clearly an outlier. The version of the box-and-whisker plot used in Fig. 6.2 is heavily outlier-dependent. To get rid of this dependency on outliers, most statistical packages do not draw whiskers all the way to the minimum and maximum values contained in the data. Rather, they impose a restriction on the length of the whiskers. In R, whiskers are, by default, no longer than 1.5 times the size of the box. This length corresponds approximately to a 99% confidence interval for the normal distribution. Values outside of this range appear as circles. Depicting outliers as circles ensures that information about outliers in the data does not get lost in the box-and-whisker plot.

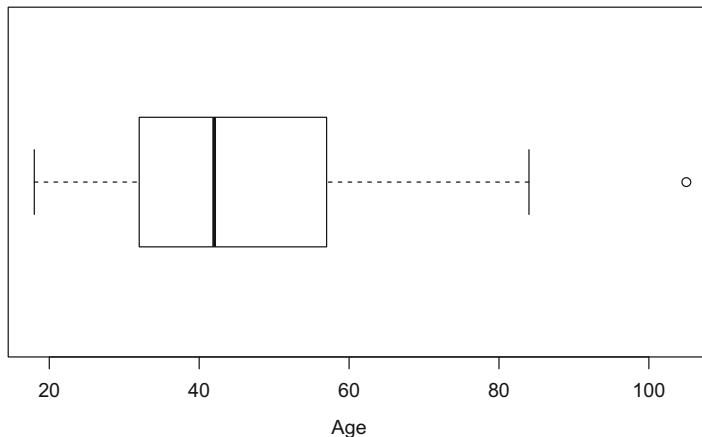


Fig. 6.3 Box-and-whisker plot of tourists' age in the Australian travel motives data set

The standard box-and-whisker plot for variable AGE in R results from:

```
R> boxplot(vac$Age, horizontal = TRUE, xlab = "Age")
```

`horizontal = TRUE` indicates that the box is horizontally aligned, otherwise it would be rotated by 90° . The result is shown in Fig. 6.3.

A comprehensive discussion of graphical methods for numeric data can be found in Putler and Krider (2012) and Chapman and Feit (2015).

To further illustrate the value of graphical methods, we visualise the percentages of agreement with the travel motives contained in the last 20 columns of the Australian travel motives data set. The numeric summaries introduced earlier offer some insights into the data, but they fail to provide an overview of the structure of the data that is intuitively easy and quick to understand. Using R, a graphical representation of this data can be generated with only two commands. Columns 13 to 32 of the data set contain the travel motives, and "yes" means that the motive does apply. Searching for string "yes" returns TRUE or FALSE (for "no"), function `colMeans()` computes the mean number of TRUEs (that is, "yes") for each column as a fraction between 0 and 1. Multiplying by 100 gives a percentage value between 0 and 100. The mean percentages are sorted, and a dot chart with a customised *x*-axis (argument `xlab` for the label and `xlim` for the range) is created:

```
R> yes <- 100 * colMeans(vac[, 13:32] == "yes")
R> dotchart(sort(yes), xlab = "Percent 'yes'", 
+   xlim = c(0, 100))
```

The resulting chart in Fig. 6.4 shows – for the travel motives contained in the data set – the percentage of respondents indicating that each of the travel motives was important to them on the last vacation.

One look at this dot chart illustrates the wide range of agreement levels with the travel motives. The vast majority of tourists want to rest and relax, but realising one's

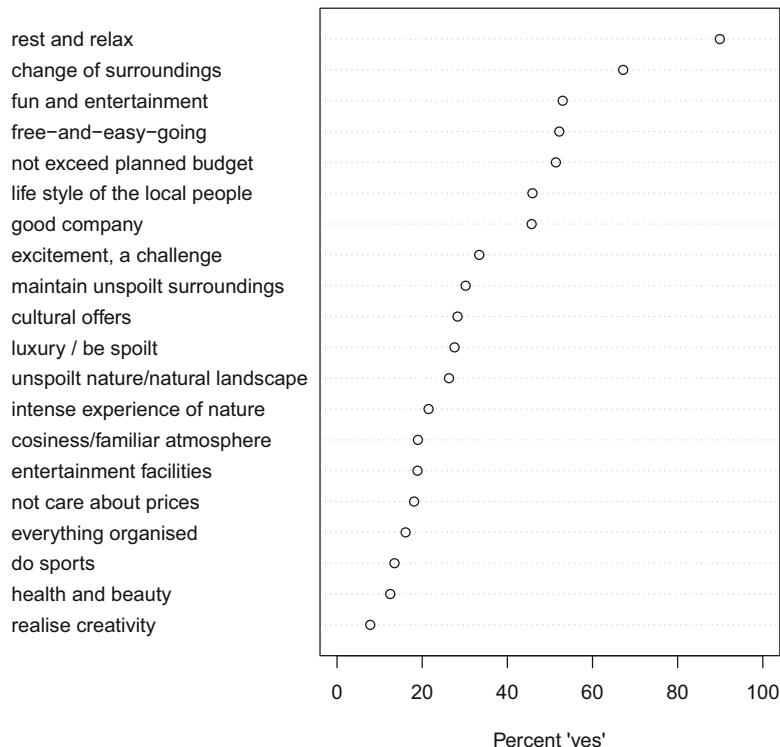


Fig. 6.4 Dot chart of percentages of YES answers in the Australian travel motives data set

creativity is important to only a very small proportion of respondents. The graphical inspection of the data also confirms the suitability of the Australian travel motives variables as segmentation variables because of the heterogeneity in the importance attributed to different motives. In other words: not all respondents say either YES or NO to most of those travel motives; differences exist between people. Such differences between people stand at the centre of market segmentation analysis.

6.4 Pre-Processing

6.4.1 Categorical Variables

Two pre-processing procedures are often used for categorical variables. One is merging levels of categorical variables before further analysis, the other one is converting categorical variables to numeric ones, if it makes sense to do so.

Merging levels of categorical variables is useful if the original categories are too differentiated (too many). Thinking back to the income variables, for example, the original income variable as used in the survey has the following categories:

```
R> sort(table(vac$Income))

$210,001 to $240,000    more than $240,001
                           10                      11
$180,001 to $210,000   $150,001 to $180,000
                           15                      32
$120,001 to $150,000   $90,001 to $120,000
                           72                      146
Less than $30,000       $60,001 to $90,000
                           150                     233
$30,001 to $60,000
                           265
```

The categories are sorted by the number of respondents. Only 68 people had an income higher than \$150,000. The three top income categories contain only between 10 and 15 people each, which corresponds to only 1% to 1.5% of the observations in the data set with 1000 respondents. Merging all these categories with the next income category (72 people with an income between \$120,001 and \$150,000), results in the new variable `Income2`, which has much more balanced frequencies:

```
R> table(vac$Income2)

<30k  30-60k  60-90k  90-120k  >120k
150     265     233     146      140
```

Many methods of data analysis make assumptions about the measurement level or scale of variables. The distance-based clustering methods presented in Step 5 assume that data are numeric, and measured on comparable scales. Sometimes it is possible to transform categorical variables into numeric variables.

Ordinal data can be converted to numeric data if it can be assumed that distances between adjacent scale points on the ordinal scale are approximately equal. This is a reasonable assumption for income, where the underlying metric construct is classified into categories covering ranges of equal length.

Another ordinal scale or multi-category scale frequently used in consumer surveys is the popular agreement scale which is often – but not always correctly – referred to as Likert scale (Likert 1932). Typically items measured on such a multi-category scale are bipolar and offer respondents five or seven answer options. The verbal labelling is usually worded as follows: STRONGLY DISAGREE, DISAGREE, NEITHER AGREE NOR DISAGREE, AGREE, STRONGLY AGREE. The assumption is frequently made that the distances between these answer options are the same. If this can be convincingly argued, such data can be treated as numerical. Note, however, that there is ample evidence that this may not be the case due to response styles at both the individual and cross-cultural level (Paulhus 1991; Marin et al. 1992; Hui and Triandis 1989; Baumgartner and Steenkamp 2001; Dolnicar and Grün 2007). It is therefore important to consider the consequences of the chosen survey response

options before collecting data in Step 3. Unless there is a strong argument for using multi-category scales (with uncertain distances between scale points), it may be preferable to use binary answer options.

Binary answer options are less prone to capturing response styles, and do not require data pre-processing. Pre-processing inevitably alters the data in some way. Binary variables can always be converted to numeric variables, and most statistical procedures work correctly after conversion if there are only two categories. Converting dichotomous ordinal or nominal variables to binary 0/1 variables is not a problem. For example, to use the travel motives as segmentation variables, they can be converted to a numeric matrix with 0 and 1 for NO and YES:

```
R> vacmot <- (vac[, 13:32] == "yes") + 0
```

Adding 0 to the logical matrix resulting from comparing the entries in the data frame to string "yes" converts the logical matrix to a numeric matrix with 0 for FALSE and 1 for TRUE. We will use matrix `vacmot` several times in the book. R package `flexclust` (Leisch 2006) contains it as a sample data set. We can load the data into R using `data("vacmot", package = "flexclust")`. This does not only load the data matrix containing the travel motives `vacmot`, but also the data frame `vacmotdesc` containing socio-demographic descriptor variables.

6.4.2 Numeric Variables

The range of values of a segmentation variable affects its relative influence in distance-based methods of segment extraction. If, for example, one of the segmentation variables is binary (with values 0 or 1 indicating whether or not a tourist likes to dine out during their vacation), and a second variable indicates the expenditure in dollars per person per day (and ranges from zero to \$1000), a difference in spend per person per day of one dollar is weighted equally as the difference between liking to dine out or not. To balance the influence of segmentation variables on segmentation results, variables can be standardised. Standardising variables means transforming them in a way that puts them on a common scale.

The default standardisation method in statistics subtracts the empirical mean \bar{x} and divides by the empirical standard deviation s :

$$z_i = \frac{x_i - \bar{x}}{s},$$

with

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

for the n observations of a variable $\mathbf{x} = \{x_1, \dots, x_n\}$. This implies that the empirical mean and the empirical standard deviation of \mathbf{z} are 0 and 1, respectively. Standardisation can be done in R using function `scale()`.

```
R> vacmot.scaled <- scale(vacmot)
```

Alternative standardisation methods may be required if the data contains observations located very far away from most of the data (outliers). In such situations, robust estimates for location and spread – such as the median and the inter quartile range – are preferable.

6.5 Principal Components Analysis

Principal components analysis (PCA) transforms a multivariate data set containing metric variables to a new data set with variables – referred to as principal components – which are uncorrelated and ordered by importance. The first variable (principle component) contains most of the variability, the second principle component contains the second most variability, and so on. After transformation, observations (consumers) still have the same relative positions to one another, and the dimensionality of the new data set is the same because principal components analysis generates as many new variables as there were old ones. Principal components analysis basically keeps the data space unchanged, but looks at it from a different angle.

Principal components analysis works off the covariance or correlation matrix of several numeric variables. If all variables are measured on the same scale, and have similar data ranges, it is not important which one to use. If the data ranges are different, the correlation matrix should be used (which is equivalent to standardising the data).

In most cases, the transformation obtained from principal components analysis is used to project high-dimensional data into lower dimensions for plotting purposes. In this case, only a subset of principal components are used, typically the first few because they capture the most variation. The first two principal components can easily be inspected in a scatter plot. More than two principal components can be visualised in a scatter plot matrix.

The following command generates a principal components analysis for the Australian travel motives data set:

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

In `prcomp`, the data is centered, but not standardised by default. Given that all variables are binary, not standardising is reasonable. We can inspect the resulting object `vacmot.pca` by printing it:

```
R> vacmot.pca
```

The print output shows the standard deviations of the principal components:

Standard deviations (1, ..., p=20) :

```
[1] 0.81 0.57 0.53 0.51 0.47 0.45 0.43 0.42 0.41 0.38
[11] 0.36 0.36 0.35 0.33 0.33 0.32 0.31 0.30 0.28 0.24
```

These standard deviations reflect the importance of each principal component. The print output also shows the rotation matrix, specifying how to rotate the original data matrix to obtain the principal components:

Rotation (n x k) = (20 x 20) :

	PC1	PC2	PC3
rest and relax	-0.063	0.0120	0.1345
luxury / be spoilt	-0.109	0.3932	-0.1167
do sports	-0.095	0.1456	-0.0456
excitement, a challenge	-0.277	0.2227	-0.2103
not exceed planned budget	-0.286	-0.1561	0.5831
realise creativity	-0.110	-0.0122	-0.0153
fun and entertainment	-0.279	0.5205	0.0865
good company	-0.284	-0.0097	0.1291
health and beauty	-0.140	0.0509	0.0039
free-and-easy-going	-0.317	0.0575	0.2445
entertainment facilities	-0.118	0.3207	0.0050
not care about prices	-0.049	0.2397	-0.2988
life style of the local people	-0.353	-0.2672	-0.3982
intense experience of nature	-0.241	-0.2133	-0.0763
cosiness/familiar atmosphere	-0.132	-0.0133	0.2017
Maintain unspoilt surroundings	-0.307	-0.3361	0.0052
everything organised	-0.092	0.1649	0.0780
unspoilt nature/natural landscape	-0.269	-0.1831	-0.0556
cultural offers	-0.260	-0.1160	-0.4282
change of surroundings	-0.259	0.0919	0.1043

Only the part of the rotation matrix corresponding to the first three principal components is shown here. The column PC1 indicates how the first principal component is composed of the original variables. This shows that the first principal component separates the two answer tendencies “almost no motives apply” and “all motives apply”, and therefore is not of much managerial value. For the second principal component, the variables loading highest are FUN and ENTERTAINMENT, LUXURY / BE SPOILT and to MAINTAIN AN UNSPOILT SURROUNDING. For the third principal component not exceeding the planned budget, cultural offers, and the life style of the local people are important variables.

We can obtain further information on the fitted object with the summary function. For objects returned by function prcomp, the function summary gives:

```
R> print(summary(vacmot.pca), digits = 2)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	0.81	0.57	0.529	0.509	0.47	0.455
Proportion of Variance	0.18	0.09	0.077	0.071	0.06	0.057

```

Cumulative Proportion 0.18 0.27 0.348 0.419 0.48 0.536
                      PC7   PC8   PC9  PC10  PC11  PC12
Standard deviation    0.431 0.420 0.405 0.375 0.364 0.360
Proportion of Variance 0.051 0.048 0.045 0.039 0.036 0.035
Cumulative Proportion 0.587 0.635 0.681 0.719 0.756 0.791
                      PC13  PC14  PC15  PC16  PC17  PC18
Standard deviation    0.348 0.33 0.33 0.320 0.306 0.297
Proportion of Variance 0.033 0.03 0.03 0.028 0.026 0.024
Cumulative Proportion 0.824 0.85 0.88 0.912 0.938 0.962
                      PC19  PC20
Standard deviation    0.281 0.243
Proportion of Variance 0.022 0.016
Cumulative Proportion 0.984 1.000

```

We interpret the output as follows: for each principal component (PC), the matrix lists standard deviation, proportion of explained variance of the original variables, and cumulative proportion of explained variance. The latter two are the most important pieces of information. Principal component 1 explains about one fifth (18%) of the variance of the original data; principal component 2 about one tenth (9%). Together, they explain 27% of the variation in the original data. Principal components 3 to 15 explain only between 8% and 3% of the original variation.

The fact that the first few principal components do not explain much of the variance indicates that all the original items (survey questions) are needed as segmentation variables. They are not redundant. They all contribute valuable information. From a projection perspective, this is bad news because it is not easy to project the data into lower dimensions. If a small number of principal components explains a substantial proportion of the variance, illustrating data using those components only gives a good visual representation of how close observations are to one another.

Returning to the Australian travel motives data set: we now want to plot the data in two-dimensional space. Usually we would do that by taking the first and second principal component. Inspecting the rotation matrix reveals that the first principal component does not differentiate well between motives because all motives load on it negatively. Principal components 2 and 3 display a more differentiated loading pattern of motives. We therefore use principal components 2 and 3 to create a perceptual map (Fig. 6.5):

```

R> library("flexclust")
R> plot(predict(vacmot.pca) [, 2:3], pch = 16,
+       col = "grey80")
R> projAxes(vacmot.pca, which = 2:3)

```

`predict(vacmot.pca) [, 2:3]` contains the rotated data and selects principal components 2 and 3. Points are drawn as filled circles (`pch = 16`) in light grey (`col`). Function `projAxes` plots how the principal components are composed of the original variables, and visualises the rotation matrix. As can be seen, NOT EXCEEDING THE PLANNED BUDGET (represented by the arrow pointing in the top slightly left direction) is a travel motive that is quite unique, whereas, for example,

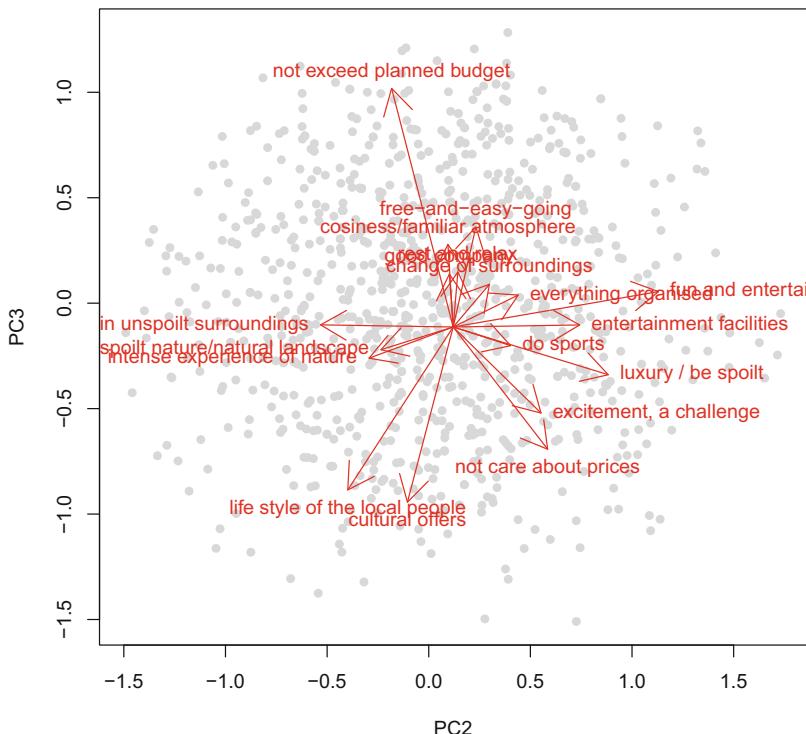


Fig. 6.5 Principal components 2 and 3 for the Australian travel motives data set

interest in the LIFESTYLE OF LOCAL PEOPLE, and interest in CULTURAL OFFERS available at destinations often occur simultaneously (as indicated by the two arrows both pointing to the left bottom of Fig. 6.5). A group of nature-oriented travel motives (arrows pointing to the left side of the chart) stands in direct contrast to the travel motives of LUXURY, EXCITEMENT, and NOT CARING ABOUT PRICES (arrows pointing to the right side of the chart).

Sometimes principal components analysis is used for the purpose of reducing the number of segmentation variables before extracting market segments from consumer data. This idea is appealing because more variables mean that the dimensionality of the problem the segment extraction technique needs to manage increases, thus making extraction more difficult and increasing sample size requirements (Dolnicar et al. 2014, 2016). Reducing dimensionality by selecting only a limited number of principal components has also been recommended in the early segmentation literature (Beane and Ennis 1987; Tynan and Drayton 1987), but has been since shown to be highly problematic (Sheppard 1996; Dolnicar and Grün 2008).

This will be discussed in detail in Sect. 7.4.3, but the key problem is that this procedure *replaces* original variables with a subset of factors or principal components. If all principal components would be used, the same data would be

used; it would merely be looked at from a different angle. But because typically only a small subset of resulting components is used, a different space effectively serves as the basis for extracting market segments. While using a subset of principal components as segmentation variables is therefore not recommended, it is safe to use principal components analysis to explore data, and identify highly correlated variables. Highly correlated variables will display high loadings on the same principal components, indicating redundancy in the information captured by them. Insights gained from such an exploratory analysis can be used to remove some of the original – redundant – variables from the segmentation base. This approach also achieves a reduction in dimensionality, but still works with the original variables collected.

6.6 Step 4 Checklist

Task	Who is responsible?	Completed?
Explore the data to determine if there are any inconsistencies and if there are any systematic contaminations.		<input type="checkbox"/>
If necessary, clean the data.		<input type="checkbox"/>
If necessary, pre-process the data.		<input type="checkbox"/>
Check if the number of segmentation variables is too high given the available sample size. You should have information from a minimum of 100 consumers for each segmentation variable.		<input type="checkbox"/>
If you have too many segmentation variables, use one of the available approaches to select a subset.		<input type="checkbox"/>
Check if the segmentation variables are correlated. If they are, choose a subset of uncorrelated segmentation variables.		<input type="checkbox"/>
Pass on the cleaned and pre-processed data to Step 5 where segments will be extracted from it.		<input type="checkbox"/>

References

- Baumgartner H, Steenkamp JBEM (2001) Response styles in marketing research: a cross-national investigation. *J Mark Res* 38(2):143–156
- Beane TP, Ennis DM (1987) Market segmentation: a review. *Eur J Mark* 21(5):20–42
- Chapman CN, McDonnell Feit E (2015) R for marketing research and analytics. *UseR!*. Springer International Publishing, Cham

- Dolnicar S, Grün B (2007) Assessing analytical robustness in cross-cultural comparisons. *Int J Tour Cult Hosp Res* 1(2):140–160
- Dolnicar S, Grün B (2008) Challenging “factor-cluster segmentation”. *J Travel Res* 47(1):63–71
- Dolnicar S, Grün B, Leisch F, Schmidt K (2014) Required sample sizes for data-driven market segmentation analyses in tourism. *J Travel Res* 53(3):296–306
- Dolnicar S, Grün B, Leisch F (2016) Increasing sample size compensates for data problems in segmentation studies. *J Bus Res* 69:992–999
- Hui C, Triandis H (1989) Effects of culture and response format on extreme response style. *J Cross Cult Psychol* 20(3):2960–309
- Leisch F (2006) A toolbox for k-centroids cluster analysis. *Comput Stat Data Anal* 51(2):526–544
- Likert R (1932) A technique for the measurement of attitudes. *Arch Psychol* 140:1–55
- Marin G, Gamba R, Marin B (1992) Extreme response style and acquiescence among hispanics – the role of acculturation and education. *J Cross Cult Psychol* 23(4):498–509
- Paulhus D (1991) Measurement and control of response bias. In: Robinson J, Shaver P, Wrightsman L (eds) *Measures of personality and social psychological attitudes*. Academic, San Diego, pp 17–59
- Putler DS, Krider RE (2012) Customer and business analytics: applied data mining for business decision making using R. Chapman&Hall/CRC, Boca Raton
- Sarkar D (2008) *Lattice: multivariate data visualization with R*. Springer, New York
- Sheppard AG (1996) The sequence of factor analysis and cluster analysis: differences in segmentation and dimensionality through the use of raw and factor scores. *Tour Anal* 1:49–57
- Tukey J (1977) *Exploratory data analysis*. Addison-Wesley, Reading
- Tynan AC, Drayton J (1987) Market segmentation. *J Mark Manag* 2(3):301–335

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Step 5: Extracting Segments



7.1 Grouping Consumers

Data-driven market segmentation analysis is exploratory by nature. Consumer data sets are typically not well structured. Consumers come in all shapes and forms; a two-dimensional plot of consumers' product preferences typically does not contain clear groups of consumers. Rather, consumer preferences are spread across the entire plot. The combination of exploratory methods and unstructured consumer data means that results from any method used to extract market segments from such data will strongly depend on the assumptions made on the structure of the segments implied by the method. The result of a market segmentation analysis, therefore, is determined as much by the underlying data as it is by the extraction algorithm chosen. Segmentation methods shape the segmentation solution.

Many segmentation methods used to extract market segments are taken from the field of cluster analysis. In that case, market segments correspond to clusters. As pointed out by Hennig and Liao (2013), selecting a suitable clustering method requires matching the data analytic features of the resulting clustering with the context-dependent requirements that are desired by the researcher (p. 315). It is, therefore, important to explore market segmentation solutions derived from a range of different clustering methods. It is also important to understand how different algorithms impose structure on the extracted segments.

One of the most illustrative examples of how algorithms impose structure is shown in Fig. 7.1. In this figure, the same data set – containing two spiralling segments – is segmented using two different algorithms, and two different numbers of segments. The top row in Fig. 7.1 shows the market segments obtained when running k -means cluster analysis (for details see Sect. 7.2.3) with 2 (left) and 8 segments (right), respectively. As can be seen, k -means cluster analysis fails to identify the naturally existing spiral-shaped segments in the data. This is because k -means cluster analysis aims at finding compact clusters covering a similar range in all dimensions.

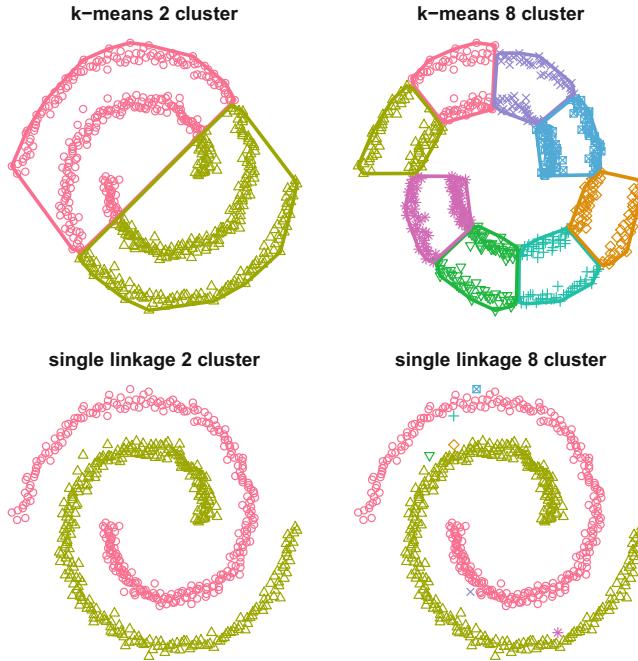


Fig. 7.1 k -means and single linkage hierarchical clustering of two spirals

The bottom row in Fig. 7.1 shows the market segments obtained from single linkage hierarchical clustering (for details see Sect. 7.2.2). This algorithm correctly identifies the existing two spiralling segments, even if the incorrect number of segments is specified up front. This is because the single linkage method constructs snake-shaped clusters. When asked to return too many (8) segments, outliers are defined as micro-segments, but the two main spirals are still correctly identified. k -means cluster analysis fails to identify the spirals because it is designed to construct round, equally sized clusters. As a consequence, the k -means algorithm ignores the spiral structure and, instead, places consumers in the same market segments if they are located close to one another (in Euclidean space), irrespective of the spiral they belong to.

This illustration gives the impression that single linkage clustering is much more powerful, and should be preferred over other approaches of extracting market segments from data. This is not the case. This particular data set was constructed specifically to play to the strengths of the single linkage algorithm allowing single linkage to identify the grouping corresponding to the spirals, and highlighting how critical the interaction between data and algorithm is. There is no single best algorithm for all data sets. If consumer data is well-structured, and well-separated, distinct market segments exist, tendencies of different algorithms matter less. If, however, data is not well-structured, the tendency of the algorithm influences the solution substantially. In such situations, the algorithm will impose a structure that suits the objective function of the algorithm.

Table 7.1 Data set and segment characteristics informing extraction algorithm selection

Data set characteristics:	<ul style="list-style-type: none"> – Size (number of consumers, number of segmentation variables) – Scale level of segmentation variables (nominal, ordinal, metric, mixed) – Special structure, additional information
Segment characteristics:	<ul style="list-style-type: none"> – Similarities of consumers in the same segment – Differences between consumers from different segments – Number and size of segments

The aim of this chapter is to provide an overview of the most popular extraction methods used in market segmentation, and point out their specific tendencies of imposing structure on the extracted segments. None of these methods outperform other methods in all situations. Rather, each method has advantages and disadvantages.

So-called *distance-based methods* are described first. Distance-based methods use a particular notion of similarity or distance between observations (consumers), and try to find groups of similar observations (market segments). So-called *model-based methods* are described second. These methods formulate a concise stochastic model for the market segments. In addition to those main two groups of extraction methods, a number of methods exist which try to achieve multiple aims in one step. For example, some methods perform variable selection during the extraction of market segments. A few such specialised algorithms are also discussed in this chapter.

Because no single best algorithm exists, investigating and comparing alternative segmentation solutions is critical to arriving at a good final solution. Data characteristics and expected or desired segment characteristics allow a pre-selection of suitable algorithms to be included in the comparison. Table 7.1 contains the information needed to guide algorithm selection.

The size of the available data set indicates if the number of consumers is sufficient for the available number of segmentation variables, the expected number of segments, and the segment sizes. The minimum segment size required from a target segment has been defined as one of the knock-out criteria in Step 2. It informs the expectation about how many segments of which size will be extracted. If the target segment is expected to be a niche segment, larger sample sizes are required. Larger samples allow a more fine-grained extraction of segments. If the number of segmentation variables is large, but not all segmentation variables are expected to be key characteristics of segments, extraction algorithms which simultaneously select variables are helpful (see Sect. 7.4).

The scale level of the segmentation variables determines the most suitable variant of an extraction algorithms. For distance-based methods, the choice of the distance measure depends on the scale level of the data. The scale level also determines the set of suitable segment-specific models in the model-based approach. Other special structures of the data can restrict the set of suitable algorithms. If the data set contains repeated measurements of consumers over time, for example, an algorithm that takes this longitudinal nature of the data into account is needed. Such data generally requires a model-based approach.

We also need to specify the characteristics consumers should have in common to be placed in the same segment, and how they should differ from consumers in other segments. These features have, conceptually, been specified in Step 2, and need to be recalled here. The structure of segments extracted by the algorithm needs to align with these expected characteristics.

We distinguish directly observable characteristics from those that are only indirectly accessible. Benefits sought are an example of a directly observable characteristic. They are contained directly in the data, placing no restrictions on the segment extraction algorithm to be chosen. An example of an indirect characteristic is consumer price sensitivity. If the data contains purchase histories and price information, and market segments are based on similar price sensitivity levels, regression models are needed. This, in turn calls for the use of a model-based segment extraction algorithm.

In the case of binary segmentation variables, another aspect needs to be considered. We may want consumers in the same segments to have both the presence and absence of segmentation variables in common. In this case, we need to treat the binary segmentation variables symmetrically (with 0s and 1s treated equally). Alternatively, we may only care about segmentation variables consumers have in common. In this case, we treat them asymmetrically (with only common 1s being of interest). An example of where it makes sense to treat them asymmetrically is if we use vacation activities as the segmentation variables. It is very interesting if two tourists both engage in horse-riding during their vacation. It is not so interesting if two tourists do not engage in horse-riding. Biclustering (see Sect. 7.4.1) uses binary information asymmetrically. Distance-based methods can use distance measures that account for this asymmetry, and extract segments characterised by common 1s.

7.2 Distance-Based Methods

Consider the problem of finding groups of tourists with similar activity patterns when on vacation. A fictitious data set is shown in Table 7.2. It contains seven people indicating the percentage of time they spend enjoying BEACH, ACTION, and CULTURE when on vacation. Anna and Bill only want to relax on the beach, Frank likes beach and action, Julia and Maria like beach and culture, Michael wants action and a little bit of culture, and Tom does everything.

Market segmentation aims at grouping consumers into groups with similar needs or behaviour, in this example: groups of tourists with similar patterns of vacation activities. Anna and Bill have exactly the same profile, and should be in the same segment. Michael is the only one not interested in going to the beach, which differentiates him from the other tourists. In order to find groups of similar tourists one needs a notion of similarity or dissimilarity, mathematically speaking: a distance measure.

Table 7.2 Artificial data set on tourist activities:
percentage of time spent on three activities

	beach	action	culture
Anna	100	0	0
Bill	100	0	0
Frank	60	40	0
Julia	70	0	30
Maria	80	0	20
Michael	0	90	10
Tom	50	20	30

7.2.1 Distance Measures

Table 7.2 is a typical data matrix. Each row represents an observation (in this case a tourist), and every column represents a variable (in this case a vacation activity). Mathematically, this can be represented as an $n \times p$ matrix where n stands for the number of observations (rows) and p for the number of variables (columns):

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

The vector corresponding to the i -th row of matrix \mathbf{X} is denoted as $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ in the following, such that $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$ is the set of all observations. In the example above, Anna's vacation activity profile is vector $\mathbf{x}_1 = (100, 0, 0)'$ and Tom's vacation activity profile is vector $\mathbf{x}_7 = (50, 20, 30)'$.

Numerous approaches to measuring the distance between two vectors exist; several are used routinely in cluster analysis and market segmentation. A distance is a function $d(\cdot, \cdot)$ with two arguments: the two vectors \mathbf{x} and \mathbf{y} between which the distance is being calculated. The result is the distance between them (a nonnegative value). A good way of thinking about distance is in the context of geography. If the distance between two cities is of interest, the location of the cities are the two vectors, and the length of the air route in kilometres is the distance. But even in the context of geographical distance, other measures of natural distance between two cities are equally valid, for example, the distance a car has to drive on roads to get from one city to the other.

A distance measure has to comply with a few criteria. One criterion is symmetry, that is:

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}).$$

A second criterion is that the distance of a vector to itself and only to itself is 0:

$$d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}.$$

In addition, most distance measures fulfil the so-called triangle inequality:

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}).$$

The triangle inequality says that if one goes from \mathbf{x} to \mathbf{z} with an intermediate stop in \mathbf{y} , the combined distance is at least as long as going from \mathbf{x} to \mathbf{z} directly.

Let $\mathbf{x} = (x_1, \dots, x_p)'$ and $\mathbf{y} = (y_1, \dots, y_p)'$ be two p -dimensional vectors. The most common distance measures used in market segmentation analysis are:

Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}$$

Manhattan or absolute distance:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p |x_j - y_j|$$

Asymmetric binary distance: applies only to binary vectors, that is, all x_j and y_j are either 0 or 1.

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \mathbf{x} = \mathbf{y} = \mathbf{0} \\ (\#\{j|x_j = 1 \text{ and } y_j = 1\}) / (\#\{j|x_j = 1 \text{ or } y_j = 1\}) & \end{cases}$$

In words: the number of dimensions where both \mathbf{x} and \mathbf{y} are equal to 1 divided by the number of dimensions where at least one of them is 1.

Euclidean distance is the most common distance measure used in market segmentation analysis. Euclidean distance corresponds to the direct “straight-line” distance between two points in two-dimensional space, as shown in Fig. 7.2 on the left. Manhattan distance derives its name from the fact that it gives the distance between two points assuming that streets on a grid (like in Manhattan) need to be used to get from one point to another. Manhattan distance is illustrated in Fig. 7.2 on the right. Both Euclidean and Manhattan distance use all dimensions of the vectors \mathbf{x} and \mathbf{y} .

Fig. 7.2 A comparison of Euclidean and Manhattan distance



The asymmetric binary distance does not use all dimensions of the vectors. It only uses dimensions where at least one of the two vectors has a value of 1. It is asymmetric because it treats 0s and 1s differently. Similarity between two observations is only concluded if they share 1s, but not if they share 0s. The dissimilarity between two observations is increased if one has a 1 and the other not. This has implications for market segmentation analysis. Imagine, for example, that the tourist vacation activity profiles not only include common vacation activities, but also unusual activities, such as HORSEBACK RIDING and BUNGEE JUMPING. The fact that two tourists have in common that they do not ride horses or that they do not bungee jump is not very helpful in terms of extracting market segments because the overall proportion of horse riders and bungee jumpers in the tourist population is low. If, however, two tourists do horse ride or bungee jump, this represents key information about similarities between them.

The asymmetric binary distance corresponds to the proportion of common 1s over all dimensions where at least one vector contains a 1. In the tourist example: the number of common vacation activities divided by the number of vacation activities at least one of the two tourists engages in. A symmetric binary distance measure (which treats 0s and 1s equally) emerges from using the Manhattan distance between the two vectors. The distance is then equal to the number of vacation activities where values are different.

The standard R function to calculate distances is called `dist()`. It takes as arguments a data matrix `x` and – optionally – the distance method. If no distance method is explicitly specified, Euclidean distance is the default. The R function returns all pairwise distances between the rows of `x`.

Using the vacation activity data in Table 7.2, we first need to load the data:

```
R> data("annabill", package = "MSA")
```

Then, we can calculate the Euclidean distance between all tourists with the following command:

```
R> D1 <- dist(annabill)
R> round(D1, 2)
```

	Anna	Bill	Frank	Julia	Maria	Michael
Bill	0.00					
Frank	56.57	56.57				
Julia	42.43	42.43	50.99			
Maria	28.28	28.28	48.99	14.14		
Michael	134.91	134.91	78.74	115.76	120.83	
Tom	61.64	61.64	37.42	28.28	37.42	88.32

The distance between Anna and Bill is zero because they have identical vacation activity profiles. The distance between Michael and all other people in the data set is substantial because Michael does not go to the beach where most other tourists spend a lot of time.

Manhattan distance – which is also referred to as absolute distance – is very similar to Euclidean distance for this data set:

```
R> D2 <- dist(annabill, method = "manhattan")
R> D2
```

	Anna	Bill	Frank	Julia	Maria	Michael
Bill	0					
Frank	80	80				
Julia	60	60	80			
Maria	40	40	80	20		
Michael	200	200	120	180	180	
Tom	100	100	60	40	60	140

No rounding is necessary because the Manhattan distance is automatically integer if all values in the data matrix are integer.

The printout contains only six rows and columns in both cases. To save computer memory, `dist()` does not return the full symmetric matrix of all pairwise distances. It only returns the lower triangle of the matrix. If the full matrix is required, it can be obtained by coercing the return object of `dist()` to the full 7×7 matrix:

```
R> as.matrix(D2)
```

	Anna	Bill	Frank	Julia	Maria	Michael	Tom
Anna	0	0	80	60	40	200	100
Bill	0	0	80	60	40	200	100
Frank	80	80	0	80	80	120	60
Julia	60	60	80	0	20	180	40
Maria	40	40	80	20	0	180	60
Michael	200	200	120	180	180	0	140
Tom	100	100	60	40	60	140	0

Both Euclidean and Manhattan distance treat all dimensions of the data equally; they take a sum over all dimensions of squared or absolute differences. If the different dimensions of the data are not on the same scale (for example, dimension 1 indicates whether or not a tourist plays golf, and dimension 2 indicates how many dollars the tourist spends per day on dining out on average), the dimension with the larger numbers will dominate the distance calculation between two observations. In such situations data needs to be standardised before calculating distances (see Sect. 6.4.2).

Function `dist` can only be used if the segmentation variables are either all metric or all binary. In R package `cluster` (Maechler et al. 2017), function `daisy` calculates the dissimilarity matrix between observations contained in a data frame. In this data frame the variables can be numeric, ordinal, nominal and binary. Following Gower (1971), all variables are rescaled to a range of [0, 1] which allows for a suitable weighting between variables. If variables are metric, the results are the same as for `dist`:

```
R> library("cluster")
R> round(daisy(annabill), digits = 2)
```

```
Dissimilarities :
      Anna   Bill  Frank  Julia  Maria Michael
Bill      0.00
Frank    56.57  56.57
Julia    42.43  42.43  50.99
Maria    28.28  28.28  48.99  14.14
Michael 134.91 134.91  78.74 115.76 120.83
Tom      61.64  61.64  37.42  28.28  37.42   88.32

Metric : euclidean
Number of objects : 7
```

7.2.2 Hierarchical Methods

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). If the aim is to have one large market segment ($k = 1$), the only possible solution is one big market segment containing all consumers in data X . At the other extreme, if the aim is to have as many market segments as there are consumers in the data set ($k = n$), the number of market segments has to be n , with each segment containing exactly one consumer. Each consumer represents their own cluster. Market segmentation analysis occurs between those two extremes.

Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment.

Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment.

Both approaches result in a sequence of nested partitions. A partition is a grouping of observations such that each observation is exactly contained in one group. The sequence of partitions ranges from partitions containing only one group (segment) to n groups (segments). They are nested because the partition with $k + 1$ groups (segments) is obtained from the partition with k groups by splitting one of the groups.

Numerous algorithms have been proposed for both strategies. The unifying framework for agglomerative clustering – which was developed in the seminal paper by Lance and Williams (1967) – contains most methods still in use today. In each step, standard implementations of hierarchical clustering perform the optimal step. This leads to a deterministic algorithm. This means that every time the hierarchical clustering algorithm is applied to the same data set, the exactly same sequence of nested partitions is obtained. There is no random component.

Underlying both divisive and agglomerative clustering is a measure of distance between groups of observations (segments). This measure is determined by specifying (1) a distance measure $d(\mathbf{x}, \mathbf{y})$ between observations (consumers) \mathbf{x} and \mathbf{y} , and (2) a *linkage method*. The linkage method generalises how, given a distance between pairs of observations, distances between groups of observations are obtained. Assuming two sets \mathcal{X} and \mathcal{Y} of observations (consumers), the following linkage methods are available in the standard R function `hclust()` for measuring the distance $l(\mathcal{X}, \mathcal{Y})$ between these two sets of observations:

Single linkage: distance between the two closest observations of the two sets.

$$l(\mathcal{X}, \mathcal{Y}) = \min_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

Complete linkage: distance between the two observations of the two sets that are farthest away from each other.

$$l(\mathcal{X}, \mathcal{Y}) = \max_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

Average linkage: mean distance between observations of the two sets.

$$l(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} d(\mathbf{x}, \mathbf{y}),$$

where $|\mathcal{X}|$ denotes the number of elements in \mathcal{X} .

These linkage methods are illustrated in Fig. 7.3, and all of them can be combined with any distance measure. There is no correct combination of distance and linkage method. Clustering in general, and hierarchical clustering in specific, are exploratory techniques. Different combinations can reveal different features of the data.

Single linkage uses a “next neighbour” approach to join sets, meaning that the two closest consumers are united. As a consequence, single linkage hierarchical clustering is capable of revealing non-convex, non-linear structures like the spirals in Fig. 7.1. In situations where clusters are not well-separated – and this means in

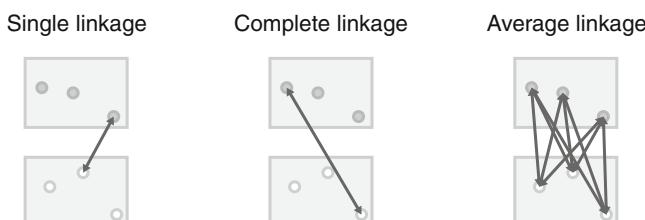


Fig. 7.3 A comparison of different linkage methods between two sets of points

most consumer data situations – the next neighbour approach can lead to undesirable chain effects where two groups of consumers form a segment only because two consumers belonging to each of those segments are close to one another. Average and complete linkage extract more compact clusters.

A very popular alternative hierarchical clustering method is named after Ward (1963), and based on squared Euclidean distances. Ward clustering joins the two sets of observations (consumers) with the minimal weighted squared Euclidean distance between cluster centers. Cluster centers are the midpoints of each cluster. They result from taking the average over the observations in the cluster. We can interpret them as segment representatives.

When using Ward clustering we need to check that the correct distance is used as input (Murtagh and Legendre 2014). The two options are Euclidean distance or squared Euclidean distance. Function `hclust()` in R can deal with both kinds of input. The input, along with the suitable linkage method, needs to be specified in the R command as either Euclidean distance with `method = "ward.D2"`, or as squared Euclidean distance with `method = "ward.D"`.

The result of hierarchical clustering is typically presented as a dendrogram. A dendrogram is a tree diagram. The root of the tree represents the one-cluster solution where one market segment contains all consumers. The leaves of the tree are the single observations (consumers), and branches in-between correspond to the hierarchy of market segments formed at each step of the procedure. The height of the branches corresponds to the distance between the clusters. Higher branches point to more distinct market segments. Dendograms are often recommended as a guide to select the number of market segments. Based on the authors' experience with market segmentation analysis using consumer data, however, dendograms rarely provide guidance of this nature because the data sets underlying the analysis are not well structured enough.

As an illustration of the dendrogram, consider the seven tourists in Table 7.2 and the Manhattan distances between them. Agglomerative hierarchical clustering with single linkage will first identify the two people with the smallest distance (Anna and Bill with a distance of 0). Next, Julia and Maria are joined into a market segment because they have the second smallest distance between them (20). The single linkage distance between these two groups is 40, because that is the distance from Maria to Anna and Bill. Tom has a distance of 40 to Julia, hence Anna, Bill, Julia, Maria and Tom are joined to a group of five in the third step. This process continues until all tourists are united in one big group. The resulting dendrogram is shown in Fig. 7.4 on the left.

The result of complete linkage clustering is provided in the right dendrogram in Fig. 7.4. For this small data set, the result is very similar. The only major difference is that Frank and Tom are first grouped together in a segment of two, before they are merged into a segment with all other tourists (except for Michael) in the data set. In both cases, Michael is merged last because his activity profile is very different. The result from average linkage clustering is not shown because the corresponding dendrogram is almost identical to that of complete linkage clustering.

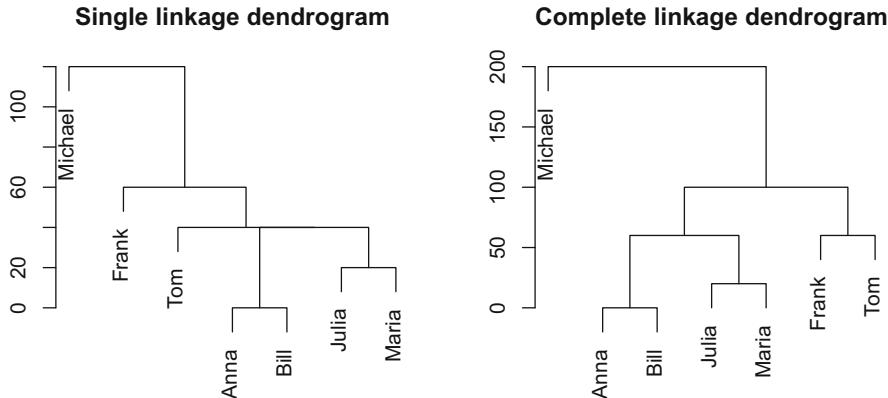


Fig. 7.4 Single and complete linkage clustering of the tourist data shown in Table 7.2

The order of the leaves of the tree (the observations or consumers) is not unique. At every split into two branches, the left and right branch could be exchanged, resulting in 2^n possible dendrograms for exactly the same clustering where n is the number of consumers in the data set. As a consequence, dendrograms resulting from different software packages may look different although they represent exactly the same market segmentation solution. Another possible source of variation between software packages is how ties are broken, meaning, which two groups are joined first when several have exactly the same distance.

Example: Tourist Risk Taking

A data set on “tourist disasters” contains survey data collected by an online research panel company in October 2015 commissioned by UQ Business School (Hajibaba et al. 2017). The target population were adult Australian residents who had undertaken at least one personal holiday in the past 12 months. The following commands load the data matrix:

```
R> library("MSA")
R> data("risk", package = "MSA")
R> dim(risk)
```

```
[1] 563   6
```

This data set contains 563 respondents who state how often they take risks from the following six categories:

1. recreational risks: e.g., rock-climbing, scuba diving
2. health risks: e.g., smoking, poor diet, high alcohol consumption
3. career risks: e.g., quitting a job without another to go to
4. financial risks: e.g., gambling, risky investments

5. safety risks: e.g., speeding
 6. social risks: e.g., standing for election, publicly challenging a rule or decision
- Respondents are presented with an ordinal scale consisting of five answer options (1=NEVER, 5=VERY OFTEN). In the subsequent analysis, we assume equidistance between categories. Respondents, on average, display risk aversion with mean values for all columns close to 2 (=RARELY):

```
R> colMeans(risk)
```

	Recreational	Health	Career	Financial
Safety	2.190053	2.396092	2.007105	2.026643
Social	2.266430	2.017762		

The following command extracts market segments from this data set using Manhattan distance and complete linkage:

```
R> risk.dist <- dist(risk, method = "manhattan")
R> risk.hcl <- hclust(risk.dist, method = "complete")
R> risk.hcl
```

Call:

```
hclust(d = risk.dist, method = "complete")
```

```
Cluster method : complete
Distance       : manhattan
Number of objects: 563
```

`plot(risk.hcl)` generates the dendrogram shown in Fig. 7.5. The dendrogram visualises the sequence of nested partitions by indicating each merger or split. The straight line at the top of the dendrogram indicates the merger of the last two groups into a single group. The y-axis indicates the distance between these two groups. At the bottom each single observation is one line.

The dendrogram in Fig. 7.5 indicates that the largest additional distance between two clusters merged occurred when the last two clusters were combined to the single cluster containing all observations. Cutting the dendrogram at a specific height selects a specific partition. The boxes numbered 1–6 in Fig. 7.5 illustrate how this dendrogram or tree can be cut into six market segments. The reason that the boxes are not numbered from left to right is that the market segment labelled number 1 contains the first observation (the first consumer) in the data set. Which consumers have been assigned to which market segment can be computed using function `cutree()`, which takes an object as returned by `hclust` and either the height `h` at which to cut or the number `k` of segments to cut the tree into.

```
R> c2 <- cutree(risk.hcl, h = 20)
R> table(c2)
```

```
c2
 1   2
511 52
```

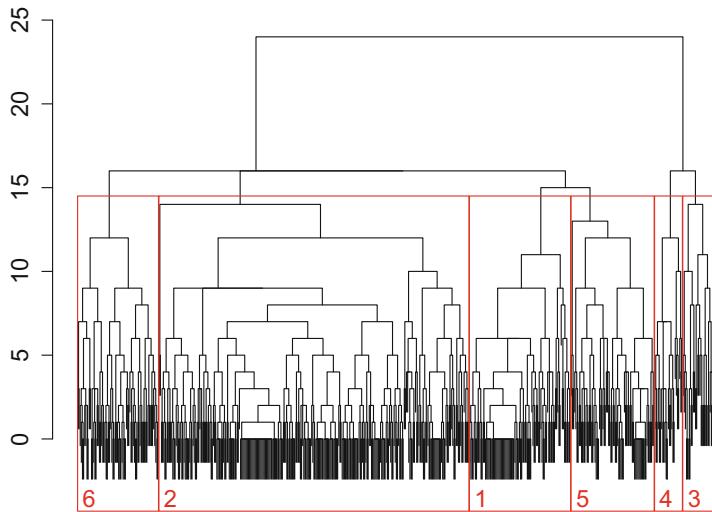


Fig. 7.5 Complete linkage hierarchical cluster analysis of the tourist risk taking data set

```
R> c6 <- cutree(risk.hcl, k = 6)
R> table(c6)
```

```
c6
 1   2   3   4   5   6
90 275 27  25  74  72
```

A simple way to assess the characteristics of the clusters is to look at the column-wise means by cluster.

```
R> c6.means <- aggregate(risk, list(Cluster = c6), mean)
R> round(c6.means, 1)
```

	Recreational	Health	Career	Financial	Safety	Social
1	2.0	2.2	1.9	2.0	2.2	2.8
2	1.9	1.8	1.5	1.6	2.0	1.4
3	3.9	4.4	2.9	3.2	3.3	4.1
4	4.1	3.3	4.1	2.8	3.4	3.2
5	2.3	2.6	3.2	2.6	2.6	2.2
6	2.0	3.8	1.8	2.4	2.3	2.0

But it is much easier to understand the cluster characteristics by visualising the column-wise means by clusters using a barchart (Fig. 7.6). `barchart(risk.hcl, risk, k=6)` from R package `flexclust` results in such a barchart. (A refined version of this plot – referred to as the segment profile plot – is described in detail in Sect. 8.3). The dark red dots correspond to the total mean values across all respondents; the bars indicate the mean values within each one of the segments. Segments are interpreted by inspecting the difference between the total population (red dots) and the segments (bars). For the tourist risk taking data set, the largest

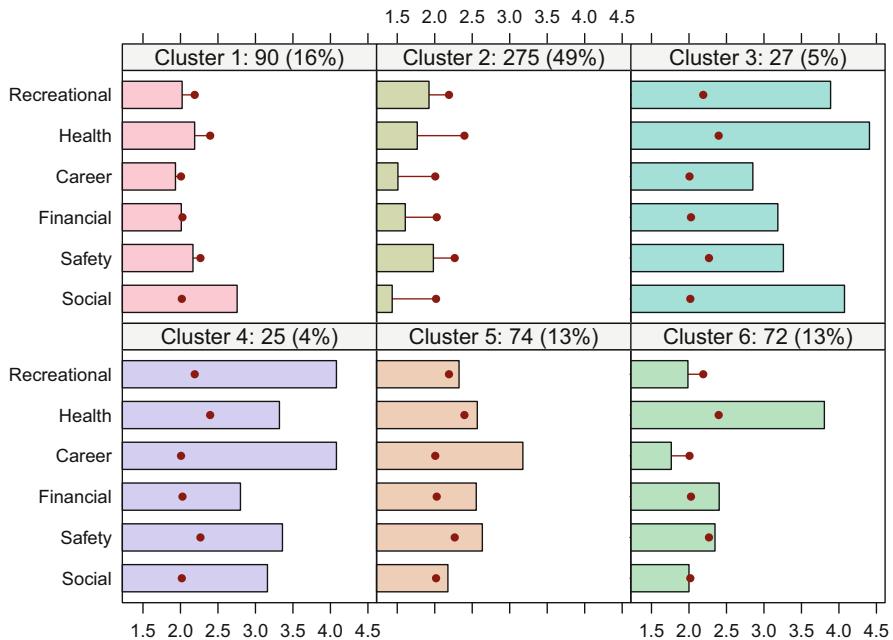


Fig. 7.6 Bar chart of cluster means from hierarchical clustering for the tourist risk taking data set

segment is cluster 2. People assigned to this segment avoid all types of risks as indicated by all bars being lower than all the red dots. Segments 3 and 4 display above average risk taking in all areas, while segments 1, 5 and 6 have average risk taking values for 5 of the 6 categories, but are characterised by their willingness to take above average risk in one category. Members of segment 1 are more willing to accept social risks than the overall population, members of segment 5 are more willing to accept career risks, and members of segment 6 are more willing to accept health risks.

7.2.3 Partitioning Methods

Hierarchical clustering methods are particularly well suited for the analysis of small data sets with up to a few hundred observations. For larger data sets, dendograms are hard to read, and the matrix of pairwise distances usually does not fit into computer memory. For data sets containing more than 1000 observations (consumers), clustering methods creating a single partition are more suitable than a nested sequence of partitions. This means that – instead of computing all distances between all pairs of observations in the data set at the beginning of a hierarchical partitioning

cluster analysis using a standard implementation – only distances between each consumer in the data set and the centre of the segments are computed. For a data set including information about 1000 consumers, for example, the agglomerative hierarchical clustering algorithm would have to calculate $(1000 \times 999)/2 = 499,500$ distances for the pairwise distance matrix between all consumers in the data set.

A partitioning clustering algorithm aiming to extract five market segments, in contrast, would only have to calculate between 5 and 5000 distances at each step of the iterative or stepwise process (the exact number depends on the algorithm used). In addition, if only a few segments are extracted, it is better to optimise specifically for that goal, rather than building the complete dendrogram and then heuristically cutting it into segments.

7.2.3.1 k -Means and k -Centroid Clustering

The most popular partitioning method is k -means clustering. Within this method, a number of algorithms are available. R function `kmeans()` implements the algorithms by Forgy (1965), Hartigan and Wong (1979), Lloyd (1982) and MacQueen (1967). These algorithms use the squared Euclidean distance. A generalisation to other distance measures, also referred to as k -centroid clustering, is provided in R package `flexclust`.

Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a set of observations (consumers) in a data set. Partitioning clustering methods divide these consumers into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible. The representative of a market segment is referred to in many partitioning clustering algorithms as the centroid. For the k -means algorithm based on the squared Euclidean distance, the centroid consists of the column-wise mean values across all members of the market segment. The data set contains observations (consumers) in rows, and variables (behavioural information or answers to survey questions) in columns. The column-wise mean, therefore, is the average response pattern across all segmentation variables for all members of the segment (Fig. 7.6).

The following generic algorithm represents a heuristic for solving the optimisation problem of dividing consumers into a given number of segments such that consumers are similar to their fellow segment members, but dissimilar to members of other segments. This algorithm is iterative; it improves the partition in each step, and is bound to converge, but not necessarily to the global optimum.

It involves five steps with the first four steps visualised in a simplified way in Fig. 7.7:

1. Specify the desired number of segments k .
2. Randomly select k observations (consumers) from data set \mathcal{X} (see Step 2 in Fig. 7.7) and use them as initial set of cluster centroids $C = \{\mathbf{c}_1, \dots, \mathbf{c}_k\}$. If five market segments are being extracted, then five consumers are randomly drawn from the data set, and declared the representatives of the five market segments. Of

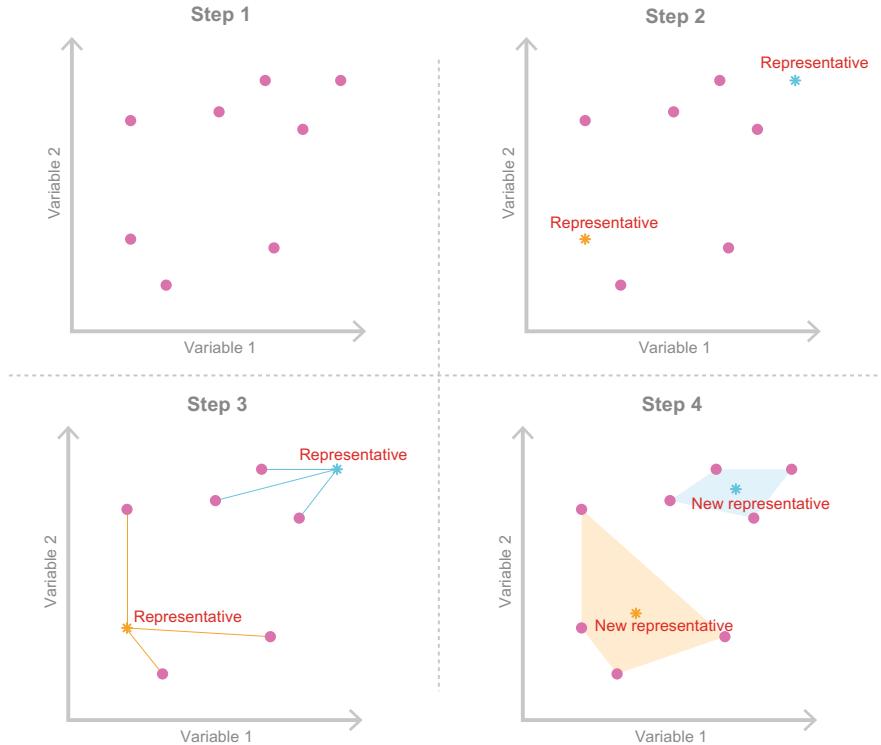


Fig. 7.7 Simplified visualisation of the k -means clustering algorithm

course, these randomly chosen consumers will – at this early stage of the process – not be representing the optimal segmentation solution. They are needed to get the step wise (iterative) partitioning algorithm started.

3. Assign each observation \mathbf{x}_i to the closest cluster centroid (segment representative, see Step 3 in Fig. 7.7) to form a partition of the data, that is, k market segments $\mathcal{S}_1, \dots, \mathcal{S}_k$ where

$$\mathcal{S}_j = \{\mathbf{x} \in \mathcal{X} | d(\mathbf{x}, \mathbf{c}_j) \leq d(\mathbf{x}, \mathbf{c}_h), 1 \leq h \leq k\}.$$

This means that each consumer in the data set is assigned to one of the initial segment representatives. This is achieved by calculating the distance between each consumer and each segment representative, and then assigning the consumer to the market segment with the most similar representative. If two segment representatives are equally close, one needs to be randomly selected. The result of this step is an initial – suboptimal – segmentation solution. All consumers in the data set are assigned to a segment. But the segments do not yet comply with the criterion that members of the same segment are as similar as possible, and members of different segments are as dissimilar as possible.

4. Recompute the cluster centroids (segment representatives) by holding cluster membership fixed, and minimising the distance from each consumer to the corresponding cluster centroid (representative see Step 4 in Fig. 7.7):

$$\mathbf{c}_j = \arg \min_{\mathbf{c}} \sum_{\mathbf{x} \in \mathcal{S}_j} d(\mathbf{x}, \mathbf{c}).$$

For squared Euclidean distance, the optimal centroids are the cluster-wise means, for Manhattan distance cluster-wise medians, resulting in the so-called k -means and k -medians procedures, respectively. In less mathematical terms: what happens here is that – acknowledging that the initial segmentation solution is not optimal – better segment representatives need to be identified. This is exactly what is achieved in this step: using the initial segmentation solution, one new representative is “elected” for each of the market segments. When squared Euclidean distance is used, this is done by calculating the average across all segment members, effectively finding the most typical, hypothetical segment members and declaring them to be the new representatives.

5. Repeat from step 3 until convergence or a pre-specified maximum number of iterations is reached. This means that the steps of assigning consumers to their closest representative, and electing new representatives is repeated until the point is reached where the segment representatives stay the same. This is when the stepwise process of the partitioning algorithm stops and the segmentation solution is declared to be the final one.

The algorithm will always converge: the stepwise process used in a partitioning clustering algorithm will always lead to a solution. Reaching the solution may take longer for large data sets, and large numbers of market segments, however. The starting point of the process is random. Random initial segment representatives are chosen at the beginning of the process. Different random initial representatives (centroids) will inevitably lead to different market segmentation solutions. Keeping this in mind is critical to conducting high quality market segmentation analysis because it serves as a reminder that running one single calculation with one single algorithm leads to nothing more than one out of many possible segmentation solutions. The key to a high quality segmentation analysis is systematic repetition, enabling the data analyst to weed out less useful solutions, and present to the users of the segmentation solution – managers of the organisation wanting to adopt target marketing – the best available market segment or set of market segments.

In addition, the algorithm requires the specification of the number of segments. This sounds much easier than it is. The challenge of determining the optimal number of market segments is as old as the endeavour of grouping people into segments itself (Thorndike 1953). A number of indices have been proposed to assist the data analyst (these are discussed in detail in Sect. 7.5.1). We prefer to assess the stability of different segmentation solutions before extracting market segments. The key idea is to systematically repeat the extraction process for different numbers of clusters (or market segments), and then select the number of segments that leads to

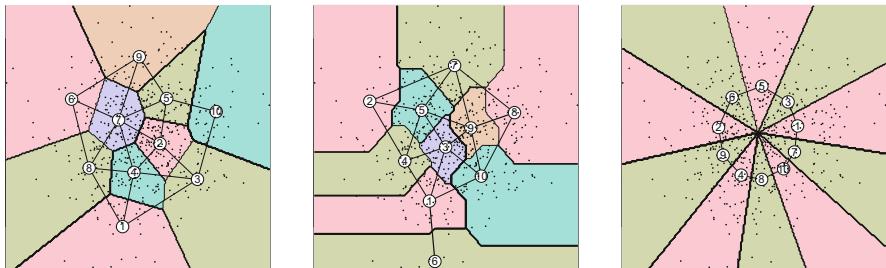


Fig. 7.8 Artificial Gaussian data clustered using squared Euclidean distance (left), Manhattan distance (middle) and angle distance (right)

either the most stable overall segmentation solution, or to the most stable individual segment. Stability analysis is discussed in detail in Sects. 7.5.3 and 7.5.4. In any case, partitioning clustering does require the data analyst to specify the number of market segments to be extracted in advance.

What is described above is a generic version of a partitioning clustering algorithm. Many variations of this generic algorithm are available; some are discussed in the subsequent subsections. The machine learning community has also proposed a number of clustering algorithms. Within this community, the term *unsupervised learning* is used to refer to clustering because groups of consumers are created without using an external (or dependent) variable. In contrast, *supervised learning* methods use a dependent variable. The equivalent statistical methods are regression (when the dependent variable is metric), and classification (when the dependent variable is nominal). Hastie et al. (2009) discuss the relationships between statistics and machine learning in detail. Machine learning algorithms essentially achieve the same thing as their statistical counterparts. The main difference is in the vocabulary used to describe the algorithms.

Irrespective of whether traditional statistical partitioning methods such as k -means are used, or whether any of the algorithms proposed by the machine learning community is applied, distance measures are the basic underlying calculation. Not surprisingly, therefore, the choice of the distance measure has a significant impact on the final segmentation solution. In fact, the choice of the distance measure typically has a bigger impact on the nature of the resulting market segmentation solution than the choice of algorithm (Leisch 2006). To illustrate this, artificial data from a bivariate normal distribution are clustered three times using a generalised version of the k -means algorithm. A different distance measure is used for each calculation: squared Euclidean distance, Manhattan distance, and the difference between angles when connecting observations to the origin.

Figure 7.8 shows the resulting three partitions. As can be seen, squared Euclidean and Manhattan distance result in similarly shaped clusters in the interior of the data. The direction of cluster borders in the outer region of the data set, however, are quite different. Squared Euclidean distance results in diagonal borders, while the borders for Manhattan distance are parallel to the axes. Angle distance slices the

data set into cake piece shaped segments. Figure 7.8 shows clearly the effect of the chosen distance measure on the segmentation solution. Note, however, that – while the three resulting segmentation solutions are different – neither of them is superior or inferior, especially given that no natural clusters are present in this data set.

Example: Artificial Mobile Phone Data

Consider a simple artificial data set for a hypothetical mobile phone market. It contains two pieces of information about mobile phone users: the number of features they want in a mobile phone, and the price they are willing to pay for it. We can artificially generate a random sample for such a scenario in R. To do this, we first load package `flexclust` which also contains a wide variety of partitioning clustering algorithms for many different distance measures:

```
R> library("flexclust")
R> set.seed(1234)
R> PF3 <- priceFeature(500, which = "3clust")
```

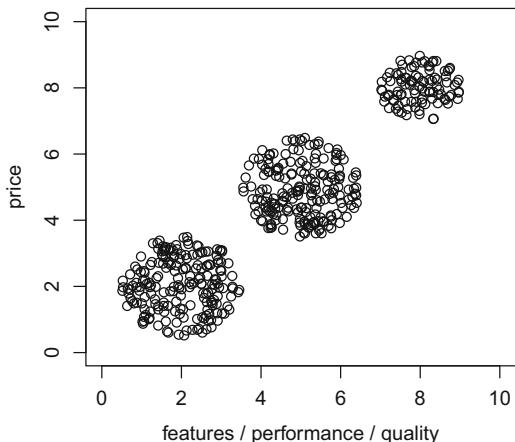
Next, we set the seed of the random number generator to 1234. We use seed 1234 throughout the book whenever randomness is involved to make all results reproducible. After setting the seed of the random number generator, it always produces exactly the same sequence of numbers. In the example above, function `priceFeature()` draws a random sample with uniform distribution on three circles. Data sets drawn with different seeds will all look very similar, but the exact location of points is different.

Figure 7.9 shows the data. The x -axis plots mobile phone features. The y -axis plots the price mobile phone users are willing to pay. The data contains three very distinct and well-separated market segments. Members of the bottom left market segment want a cheap mobile phone with a limited set of features. Members of the middle segment are willing to pay a little bit more, and expect a few additional features. Members of the small market segment located in the top right corner of Fig. 7.9 are willing to pay a lot of money for their mobile phone, but have very high expectations in terms of features.

Next, we extract market segments from this data. Figure 7.9 shows clearly that three market segments exist (when working with empirical data it is not known how many, if any, natural segments are contained in the data). To obtain a solution containing three market segments for the artificially generated mobile phone data set using k -means, we use function `cclust()` from package `flexclust`. Compared to the standard R function `kmeans()`, function `cclust()` returns richer objects, which are useful for the subsequent visualisation of results using tools from package `flexclust`. Function `cclust()` implements the k -means algorithm by determining the centroids using the average values across segment members, and by assigning each observation to the closest centroid using Euclidean distance.

```
R> PF3.km3 <- cclust(PF3, k = 3)
R> PF3.km3
```

Fig. 7.9 Artificial mobile phone data set



```
kcca object of family 'kmeans'
```

```
call:  
cclust(x = PF3, k = 3)
```

cluster sizes:

1	2	3
100	200	200

The cluster centres (centroids, representatives of each market segment), and the vector of cluster memberships (the assignment of each consumer to a specific market segment) can be extracted using

```
R> parameters(PF3.km3)
```

	features / performance / quality	price
[1,]	7.976827	8.027105
[2,]	5.021999	4.881439
[3,]	1.990105	2.062453

```
R> clusters(PF3.km3)[1:20]
```

```
[1] 1 2 3 3 2 3 2 3 1 1 3 1 3 2 2 3 2 1 2 1
```

The term [1:20] in the above R command asks for the segment memberships of only the first 20 consumers in the data set to be displayed (to save space). The numbering of the segments (clusters) is random; it depends on which consumers from the data set have been randomly chosen to be the initial segment representatives. Exactly the same solution could be obtained with a different numbering of segments; the market segment labelled cluster 1 in one calculation could be labelled cluster 3 in the next calculation, although the grouping of consumers is the same.

The information about segment membership can be used to plot market segments in colour, and to draw circles around them. These circles are referred to as convex

hulls. In two-dimensional space, the convex hull of a set of observations is a closed polygon connecting the outer points in a way that ensures that all points of the set are located within the polygon. An additional requirement is that the polygon has no “inward dents”. This means that any line connecting two data points of the set must not lie outside the convex hull. To generate a coloured scatter plot of the data with convex hulls for the segments – such as the one depicted in Fig. 7.10 – we can use function `clusterhull()` from package **MSA**:

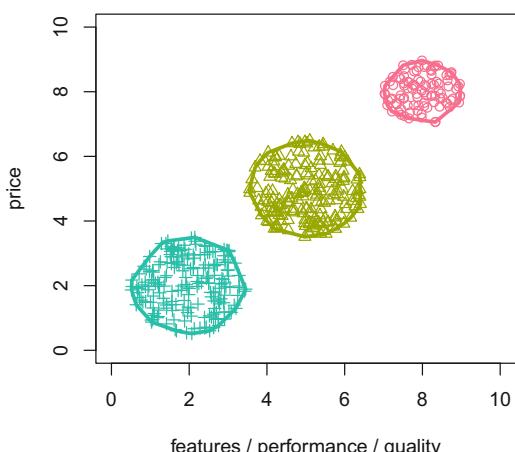
```
R> clusterhulls(PF3, clusters(PF3.km3))
```

Figure 7.10 visualises the segmentation solution resulting from a single run of the k -means algorithm with one specific set of initial segment representatives. The final segmentation solution returned by the k -means algorithm differs for different initial values. Because each calculation starts with randomly selected consumers serving as initial segment representatives, it is helpful to rerun the process of selecting random segment representatives a few times to eliminate a particularly bad initial set of segment representatives. The process of selecting random segment representatives is called random initialisation.

Specifying the number of clusters (number of segments) is difficult because, typically, consumer data does not contain distinct, well-separated naturally existing market segments. A popular approach is to repeat the clustering procedure for different numbers of market segments (for example: everything from two to eight market segments), and then compare – across those solutions – the sum of distances of all observations to their representative. The lower the distance, the better the segmentation solution because members of market segments are very similar to one another.

We now calculate 10 runs of the k -means algorithm for each number of segments using different random initial representatives (`nrep = 10`), and retain the best solution for each number of segments. The number of segments varies from 2 to 8 (`k = 2 : 8`):

Fig. 7.10 Three-segment k -means partition for the artificial mobile phone data set



```
R> PF3.km28 <- stepcclust(PF3, k = 2:8, nrep = 10)

2 : * * * * * * * * *
3 : * * * * * * * * *
4 : * * * * * * * * *
5 : * * * * * * * * *
6 : * * * * * * * * *
7 : * * * * * * * * *
8 : * * * * * * * * *

R> PF3.km28

stepFlexclust object of family 'kmeans'

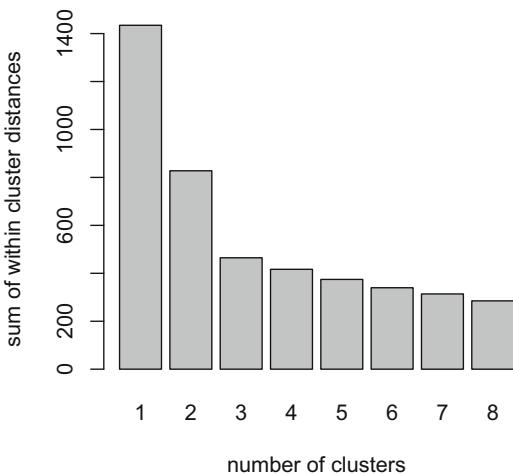
call:
stepcclust(PF3, k = 2:8, nrep = 10)

  iter converged   distsum
1     NA          NA 1434.6462
2      5          TRUE  827.6455
3      3          TRUE  464.7213
4      4          TRUE  416.6217
5     11          TRUE  374.4978
6     11          TRUE  339.6770
7     12          TRUE  313.8717
8     15          TRUE  284.9730
```

In this case, we extract market segmentation solutions containing between 2 and 8 segments (argument `k = 2:8`). For each one of those solutions, we retain the best out of ten random initialisations (`nrep = 10`), using the sum of Euclidean distances between the segment members and their segment representatives as criterion.

Function `stepcclust()` enables automated parallel processing on multiple cores of a computer (see `help("stepcclust")` for details). This is useful because the repeated calculations for different numbers of segments and different random initialisations are independent. In the example above $7 \times 10 = 70$ segment extractions are required. Without parallel computing, these 70 segment extractions run sequentially one after the other. Parallel computing means that a number of calculations can run simultaneously. Parallel computing is possible on most modern standard laptops, which can typically run at least four R processes in parallel, reducing the required runtime of the command by a factor of four (e.g., 15 s instead of 60 s). More powerful desktop machines or compute servers allow many more parallel R processes. For single runs of `stepcclust()` this makes little difference, but as soon as advanced bootstrapping procedures are used, the difference in runtime can be substantial. Calculations which would run for an hour, are processed in 15 min on a laptop, and in 1.5 min on a computer server running 40 parallel processes. The R commands used are exactly the same, but parallel processing needs to be enabled before using them. The help page for function `stepcclust()` offers examples on how to do that.

Fig. 7.11 Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the artificial mobile phone data set

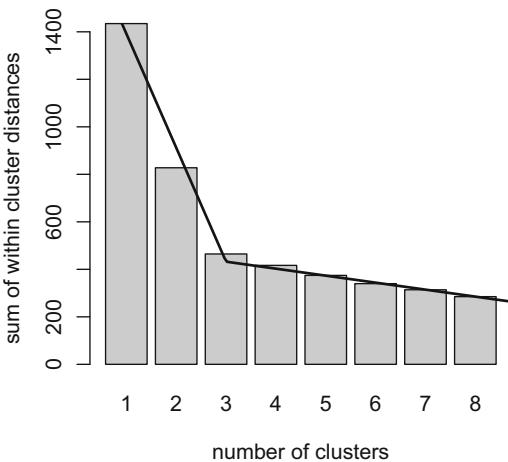


The sums of within-cluster distances for different numbers of clusters (number of market segments) are visualised using `plot(PF3.km28)`. Figure 7.11 shows the resulting *scree plot*. The scree plot displays – for each number of segments – the sum of within-cluster distances. For clustering results obtained using `stepcclust`, this is the sum of the Euclidean distances between each segment member and the representative of the segment. The smaller this number, the more homogeneous the segments; members assigned to the same market segment are similar to one another. Optimally, the scree plot shows distinct drops in the sum of within-cluster distances for the first numbers of segments, followed only by small decreases afterwards. The number of segments where the last distinct drop occurs is the optimal number of segments. After this point, homogeneous segments are split up artificially, resulting in no major decreases in the sum of within-cluster distances.

The point of the scree plot indicating the best number of segments is where an *elbow* occurs. The elbow is illustrated in Fig. 7.12. Figure 7.12 contains the scree plot as well as an illustration of the elbow. The elbow is visualised by the two intersecting lines with different slopes. The point where the two lines intersect indicates the optimal number of segments. In the example shown in Fig. 7.12, large distance drops are visible when the number of segments increases from one to two segments, and then again from two to three segments. A further increase in segments leads to small reductions in distance.

For this simple artificial data set – constructed to contain three distinct and exceptionally well-separated market segments – the scree plot in Fig. 7.11 correctly points to three market segments being a good choice. The scree plot only provides guidance if market segments are well-separated. If they are not, stability analysis – discussed in detail in Sects. 7.5.3 and 7.5.4 – can inform the number of segments decision.

Fig. 7.12 Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the artificial mobile phone data set including a visualisation of the elbow consisting of two intersecting lines with different slopes



Example: Tourist Risk Taking

To illustrate the difference between an artificially created data set (containing three textbook market segments), and a data set containing real consumer data, we use the tourist risk taking data set. We generate solutions for between 2 and 8 segments ($k = 2, \dots, 8$ clusters) using the following command:

```
R> set.seed(1234)
R> risk.km28 <- stepcclust(risk, k = 2:8, nrep = 10)
```

We use the default seed of 1234 for the random number generator, and initialise each k -means run with a different set of k random representatives. To make it possible for readers to get exactly the same results as shown in this book, the seed is actively set. Figure 7.13 contains the corresponding sum of distances. As can be seen immediately, the drops in distances are much less distinct for this consumer data set than they were for the artificial mobile phone data set. No obvious number of segments recommendation emerges from this plot. But if this plot were the only available decision tool, the two-segment solution would be chosen. We obtain the corresponding bar chart using

```
R> barchart(risk.km28[["2"]])
```

(Figure not shown). The solution containing two market segments splits the data into risk-averse people and risk-takers, reflecting the two main branches of the dendrogram in Fig. 7.5.

Figure 7.14 show the six-segment solution. It is similar to the partition resulting from the hierarchical clustering procedure, but not exactly the same. The six-segment solution resulting from the partitioning algorithm contains two segments of low risk takers (segments 1 and 4), two segments of high risk takers (segments 2 and 5), and two distinctly profiled segments, one of which contains people taking recreational and social risks (segment 3), and another one containing health risk

Fig. 7.13 Scree plot for k -means partitions with $k = 1, \dots, 8$ segments for the tourist risk taking data set

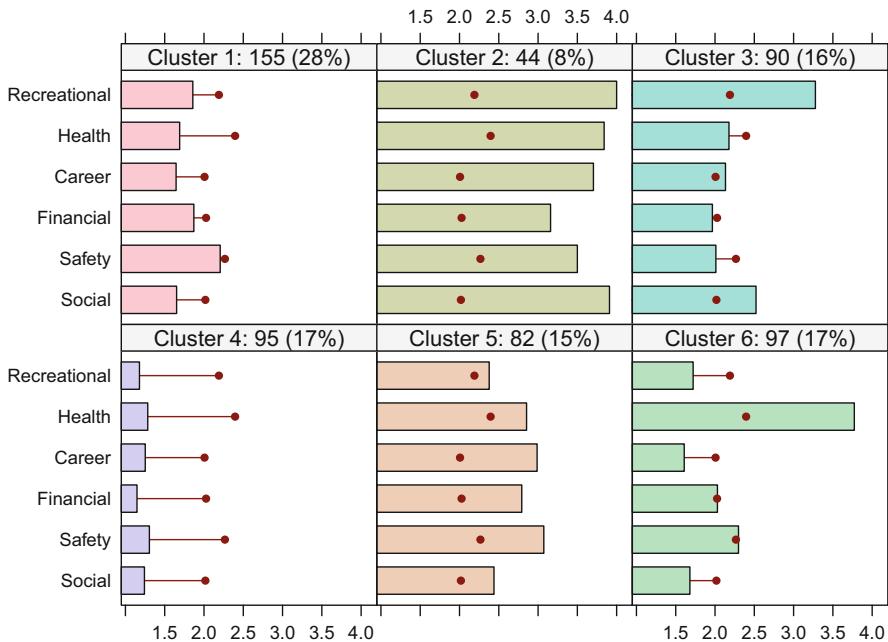
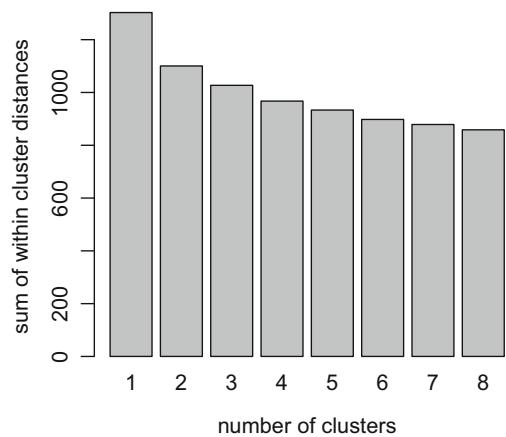


Fig. 7.14 Bar chart of cluster means from k -means clustering for the tourist risk taking data set

takers (segment 6). Both partitions obtained using either hierarchical or partitioning clustering methods are reasonable from a statistical point of view. Which partition is more suitable to underpin the market segmentation strategy of an organisation needs to be evaluated jointly by the data analyst and the user of the segmentation solution using the tools and methods presented in Sect. 7.5 and in Steps 6, 7 and 8.

7.2.3.2 “Improved” k -Means

Many attempts have been made to refine and improve the k -means clustering algorithm. The simplest improvement is to initialise k -means using “smart” starting values, rather than randomly drawing k consumers from the data set and using them as starting points. Using randomly drawn consumers is suboptimal because it may result in some of those randomly drawn consumers being located very close to one another, and thus not being representative of the data space. Using starting points that are not representative of the data space increases the likelihood of the k -means algorithm getting stuck in what is referred to as a *local optimum*. A local optimum is a good solution, but not the best possible solution. One way of avoiding the problem of the algorithm getting stuck in a local optimum is to initialise it using starting points evenly spread across the entire data space. Such starting points better represent the entire data set.

Steinley and Brusco (2007) compare 12 different strategies proposed to initialise the k -means algorithm. Based on an extensive simulation study using artificial data sets of known structure, Steinley and Brusco conclude that the best approach is to randomly draw many starting points, and select the best set. The best starting points are those that best represent the data. Good representatives are close to their segment members; the total distance of all segment members to their representatives is small (as illustrated on the left side of Fig. 7.15). Bad representatives are far away from their segment members; the total distance of all segment members to their representatives is high (as illustrated on the right side of Fig. 7.15).

7.2.3.3 Hard Competitive Learning

Hard competitive learning, also known as *learning vector quantisation* (e.g. Ripley 1996), differs from the standard k -means algorithm in how segments are extracted. Although hard competitive learning also minimises the sum of distances from

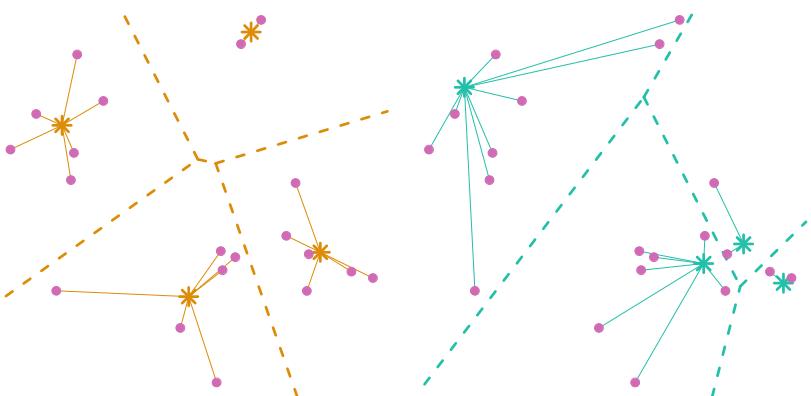


Fig. 7.15 Examples of good (left) and bad (right) starting points for k -means clustering

each consumer contained in the data set to their closest representative (centroid), the process by which this is achieved is slightly different. k -means uses *all* consumers in the data set at each iteration of the analysis to determine the new segment representatives (centroids). Hard competitive learning randomly picks one consumer and moves this consumer's closest segment representative a small step into the direction of the randomly chosen consumer.

As a consequence of this procedural difference, different segmentation solutions can emerge, even if the same starting points are used to initialise the algorithm. It is also possible that hard competitive learning finds the globally optimal market segmentation solution, while k -means gets stuck in a local optimum (or the other way around). Neither of the two methods is superior to the other; they are just different. An application of hard competitive learning in market segmentation analysis can be found in Boztug and Reutterer (2008), where the procedure is used for segment-specific market basket analysis. Hard competitive learning can be computed in R using function `cclust(x, k, method = "hardcl")` from package `flexclust`.

7.2.3.4 Neural Gas and Topology Representing Networks

A variation of hard competitive learning is the *neural gas* algorithm proposed by Martinetz et al. (1993). Here, not only the segment representative (centroid) is moved towards the randomly selected consumer. Instead, also the location of the second closest segment representative (centroid) is adjusted towards the randomly selected consumer. However, the location of the second closest representative is adjusted to a smaller degree than that of the primary representative. Neural gas has been used in applied market segmentation analysis (Dolnicar and Leisch 2010, 2014). Neural gas clustering can be performed in R using function `cclust(x, k, method = "neuralgas")` from package `flexclust`. An application with real data is presented in Sect. 7.5.4.1.

A further extension of neural gas clustering are *topology representing networks* (TRN, Martinetz and Schulten 1994). The underlying algorithm is the same as in neural gas. In addition, topology representing networks count how often each pair of segment representatives (centroids) is closest and second closest to a randomly drawn consumer. This information is used to build a virtual map in which “similar” representatives – those which had their values frequently adjusted at the same time – are placed next to one other. Almost the same information – which is central to the construction of the map in topology representing networks – can be obtained from any other clustering algorithms by counting how many consumers have certain representatives as closest and second closest in the final segmentation solution. Based on this information, the so-called *segment neighbourhood graph* (Leisch 2010) is generated. The segment neighbourhood graph is part of the default segment visualisation functions of package `flexclust`. Currently there appears to be no implementation of the original topology representing network (TRN) algorithm in R, but using neural gas in combination with neighbourhood graphs achieves similar

results. Function `cclust()` returns the neighbourhood graph by default (see Figs. 7.19, 7.41, 8.4 and 8.6 for examples). Neural gas and topology representing networks are not superior to the k -means algorithm or to hard competitive learning; they are different. As a consequence, they result in different market segmentation solutions. Given that data-driven market segmentation analysis is exploratory by very nature, it is of great value to have a larger toolbox of algorithms available for exploration.

7.2.3.5 Self-Organising Maps

Another variation of hard competitive learning are *self-organising maps* (Kohonen 1982, 2001), also referred to as *self-organising feature maps* or *Kohonen maps*. Self-organising maps position segment representatives (centroids) on a regular grid, usually a rectangular or hexagonal grid. Examples of grids are provided in Fig. 7.16.

The self-organising map algorithm is similar to hard competitive learning: a single random consumer is selected from the data set, and the closest representative for this random consumer moves a small step in their direction. In addition, representatives which are direct grid neighbours of the closest representative move in the direction of the selected random consumer. The process is repeated many times; each consumer in the data set is randomly chosen multiple times, and used to adjust the location of the centroids in the Kohonen map. What changes over the many repetitions, however, is the extent to which the representatives are allowed to change. The adjustments get smaller and smaller until a final solution is reached. The advantage of self-organising maps over other clustering algorithms is that the numbering of market segments is not random. Rather, the numbering aligns with the grid along which all segment representatives (centroids) are positioned. The price paid for this advantage is that the sum of distances between segment members and segment representatives can be larger than for other clustering algorithms. The reason is that the location of representatives cannot be chosen freely. Rather, the grid imposes restrictions on permissible locations. Comparisons of self-organising maps and topology representing networks with other clustering algorithms, such as the standard k -means algorithm, as well as for market segmentation applications are provided in Mazanec (1999) and Reutterer and Natter (2000).

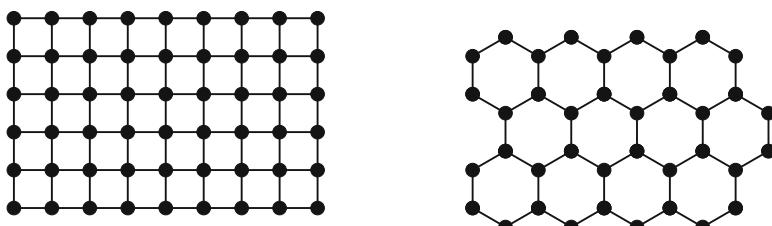
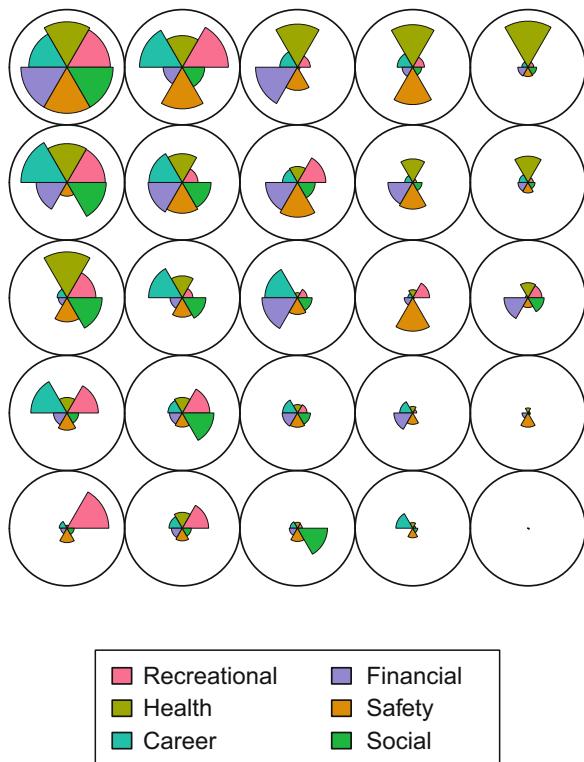


Fig. 7.16 Rectangular (*left*) and hexagonal (*right*) grid for self-organising maps

Fig. 7.17 5×5
self-organising map of the
tourist risk taking data set



Many implementations of self-organising maps are available in R packages. Here, we use function `som()` from package `kohonen` (Wehrens and Buydens 2007) because it offers good visualisations of the fitted maps. The following R commands load package `kohonen`, fit a 5×5 rectangular self-organising map to the tourist risk taking data, and plot it using the colour palette `flxPalette` from package `flexclust`:

```
R> library("kohonen")
R> set.seed(1234)
R> risk.som <- som(risk, somgrid(5, 5, "rect"))
R> plot(risk.som, palette.name = flxPalette, main = "")
```

The resulting map is shown in Fig. 7.17. As specified in the R code, the map has the shape of a five by five rectangular grid, and therefore extracts 25 market segments. Each circle on the grid represents one market segment. Neighbouring segments are more similar to one another than segments located far away from one another. The pie chart provided in Fig. 7.17 for each of the market segments contains basic information about the segmentation variables. Members of the segment in the top left corner take all six kinds of risks frequently. Members of the segment in the bottom right corner do not take any kind of risk ever. The market segments in-between display different risk taking tendencies. For example, members of the

market segment located at the very centre of the map take financial risks and career risks, but not recreational, health, safety and social risks.

7.2.3.6 Neural Networks

Auto-encoding neural networks for cluster analysis work mathematically differently than all cluster methods presented so far. The most popular method from this family of algorithms uses a so-called *single hidden layer perceptron*. A detailed description of the method and its usage in a marketing context is provided by Natter (1999). Hruschka and Natter (1999) compare neural networks and k -means.

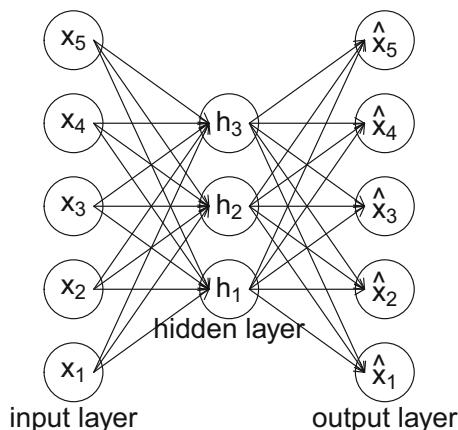
Figure 7.18 illustrates a single hidden layer perceptron. The network has three layers. The input layer takes the data as input. The output layer gives the response of the network. In the case of clustering this is the same as the input. In-between the input and output layer is the so-called hidden layer. It is named hidden because it has no connections to the outside of the network. The input layer has one so-called *node* for every segmentation variable. The example in Fig. 7.18 uses five segmentation variables. The values of the three nodes in the hidden layer h_1 , h_2 and h_3 are weighted linear combinations of the inputs

$$h_j = f_j \left(\sum_{i=1}^5 \alpha_{ij} x_i \right)$$

for a non-linear function f_j . Each weight α_{ij} in the formula is depicted by an arrow connecting nodes in input layer and hidden layer. The f_j are chosen such that $0 \leq h_j \leq 1$, and all h_j sum up to one ($h_1 + h_2 + h_3 = 1$).

In the simplest case, the outputs \hat{x}_i are weighted combinations of the hidden nodes

Fig. 7.18 Schematic representation of an auto-encoding neural network with one hidden layer



$$\hat{x}_i = \sum_{j=1}^3 \beta_{ji} h_i,$$

where coefficients β_{ji} correspond to the arrows between hidden nodes and output nodes. When training the network, the parameters α_{ij} and β_{ji} are chosen such that the squared Euclidean distance between inputs and outputs is as small as possible for the training data available (the consumers to be segmented). In neural network vocabulary, the term training is used for parameter estimation. This gives the network its name auto-encoder; it is trained to predict the inputs x_i as accurately as possible. The task would be trivial if the number of hidden nodes would be equal to the number available as inputs. If, however, fewer hidden nodes are used (which is usually the case), the network is forced to learn how to best represent the data using segment representatives.

Once the network is trained, parameters connecting the hidden layer to the output layer are interpreted in the same way as segment representatives (centroids) resulting from traditional cluster algorithms. The parameters connecting the input layer to the hidden layer can be interpreted in the following way: consider that for one particular consumer $h_1 = 1$, and hence $h_2 = h_3 = 0$. In this case $\hat{x}_i = \beta_{1i}$ for $i = 1, \dots, 5$. This is true for all consumers where h_1 is 1 or close to 1. The network predicts the same value for all consumers with $h_1 \approx 1$. All these consumers are members of market segment 1 with representative β_{1i} . All consumers with $h_2 \approx 1$, are members of segment 2, and so on.

Consumers who have no h_j value close to 1 can be seen as in-between segments. k -means clustering and hard competitive learning produce crisp segmentations, where each consumer belongs to exactly one segment. Neural network clustering is an example of a so-called fuzzy segmentation with membership values between 0 (not a member of this segment) and 1 (member of only this segment). Membership values between 0 and 1 indicate membership in multiple segments. Several implementations of auto-encoding neural networks are available in R. One example is function `autoencode()` in package `autoencoder` (Dubossarsky and Tyshetskiy 2015). Many other clustering algorithms generate fuzzy market segmentation solutions, see for example R package `fclust` (Ferraro and Giordani 2015).

7.2.4 Hybrid Approaches

Several approaches combine hierarchical and partitioning algorithms in an attempt to compensate the weaknesses of one method with the strengths of the other. The strengths of hierarchical cluster algorithms are that the number of market segments to be extracted does not have to be specified in advance, and that similarities of market segments can be visualised using a dendrogram. The biggest disadvantage of hierarchical clustering algorithms is that standard implementations require substantial memory capacity, thus restricting the possible sample size of

the data for applying these methods. Also, dendograms become very difficult to interpret when the sample size is large.

The strength of partitioning clustering algorithms is that they have minimal memory requirements during calculation, and are therefore suitable for segmenting large data sets. The disadvantage of partitioning clustering algorithms is that the number of market segments to be extracted needs to be specified in advance. Partitioning algorithms also do not enable the data analyst to track changes in segment membership across segmentation solutions with different number of segments because these segmentation solutions are not necessarily nested.

The basic idea behind hybrid segmentation approaches is to first run a partitioning algorithm because it can handle data sets of any size. But the partitioning algorithm used initially does not generate the number of segments sought. Rather, a much larger number of segments is extracted. Then, the original data is discarded and only the centres of the resulting segments (centroids, representatives of each market segment) and segment sizes are retained, and used as input for the hierarchical cluster analysis. At this point, the data set is small enough for hierarchical algorithms, and the dendrogram can inform the decision how many segments to extract.

7.2.4.1 Two-Step Clustering

IBM SPSS (IBM Corporation 2016) implemented a procedure referred to as two-step clustering (SPSS 2001). The two steps consist of run a partitioning procedure followed by a hierarchical procedure. The procedure has been used in a wide variety of application areas, including internet access types of mobile phone users (Okazaki 2006), segmenting potential nature-based tourists based on temporal factors (Tkaczynski et al. 2015), identifying and characterising potential electric vehicle adopters (Mohamed et al. 2016), and segmenting travel related risks (Ritchie et al. 2017).

The basic idea can be demonstrated using simple R commands. For this purpose we use the artificial mobile phone data set introduced in Sect. 7.2.3. First we cluster the original data using k -means with k much larger than the number of market segments sought, here $k = 30$:

```
R> set.seed(1234)
R> PF3.k30 <- stepcclust(PF3, k = 30, nrep = 10)
```

The exact number of clusters k in this first step is not crucial. Here, 30 clusters were extracted because the original data set only contains 500 observations. For large empirical data sets much larger numbers of clusters can be extracted (100, 500 or 1000). The choice of the original number of clusters to extract is not crucial because the primary aim of the first step is to reduce the size of the data set by retaining only one representative member of each of the extracted clusters. Such an application of cluster methods is often also referred to as *vector quantisation*. The following R command plots the result of running k -means to extract $k = 30$ clusters:

```
R> plot(PF3.k30, data = PF3)
```

This plot is shown in Fig. 7.19. The plot visualises the cluster solution using a *neighbourhood graph*. In a neighbourhood graph, the cluster means are the nodes, and are plotted using circles with the cluster number (label) in the middle. The edges between the nodes correspond to the similarity between clusters. In addition – if the data is provided – a scatter plot of the data with the observations coloured by cluster memberships and cluster hulls is plotted.

As can be seen, the 30 extracted clusters are located within the three segments contained in this artificially created data set. But because the number of clusters extracted is ten times larger (30) than the actual number of segments (3), each naturally existing market segment is split up into a number of even more homogeneous segments. The top right market segment – willing to pay a high price for a mobile phone with many features – has been split up in eight subsegments.

The representatives of each of these 30 market segments (centroids, cluster centres) as well as the segment sizes serve as the new data set for the second step of the procedure, the hierarchical cluster analysis. To achieve this, we need to extract the cluster centres and segment sizes from the k -means solution:

```
R> PF3.k30.cent <- parameters(PF3.k30)
R> sizes <- table(clusters(PF3.k30))
```

Based on this information, we can extract segments with hierarchical clustering using the following R command:

```
R> PF3.hc <- hclust(dist(PF3.k30.cent), members = sizes)
```

Figure 7.20 contains the resulting dendrogram produced by `plot(PF3.hc)`. The three long vertical lines in this dendrogram clearly point to the existence of three market segments in the data set. It cannot be determined from the hierarchical cluster analysis, however, which consumer belongs to which market segment. This cannot be determined because the original data was discarded. What needs to happen in the final step of two-step clustering, therefore, is to link the original data with the segmentation solution derived from the hierarchical analysis. This can be achieved using function `twoStep()` from package **MSA** which takes as argument the hierarchical clustering solution, the cluster memberships of the original data obtained with the partitioning clustering method, and the number k of segments to extract:

```
R> PF3.ts3 <- twoStep(PF3.hc, clusters(PF3.k30), k = 3)
R> table(PF3.ts3)
```

PF3.ts3			
1	2	3	
200	100	200	

As can be seen from this table (showing the number of members in each segment), the number of segment members extracted matches the number of segment members generated for this artificial data set. That the correct segments were indeed extracted is confirmed by inspecting the plot generated with the following R command: `plot(PF3, col = PF3.ts3)`. The resulting plot is not shown because it is in principal identical to that shown in Fig. 7.10.

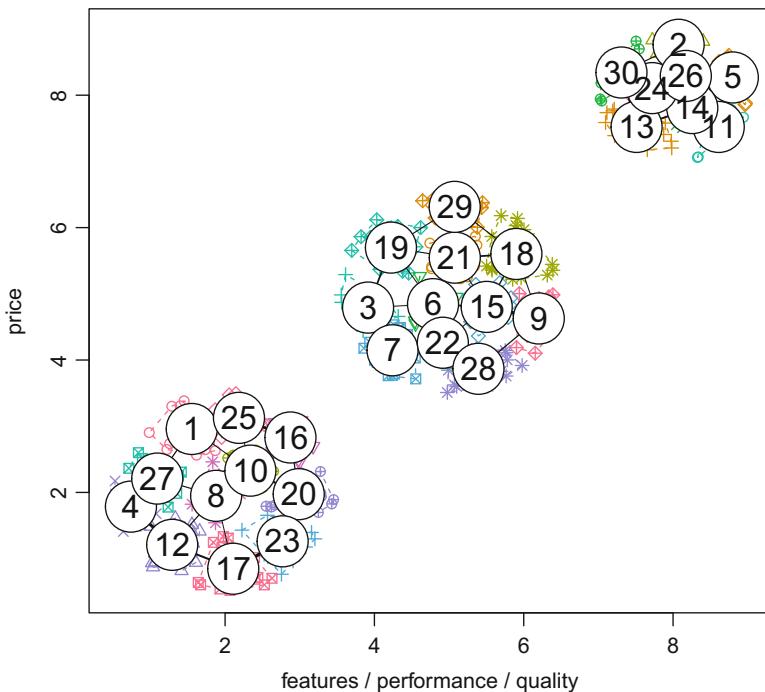


Fig. 7.19 k -means clustering of the artificial mobile phone data set into 30 clusters

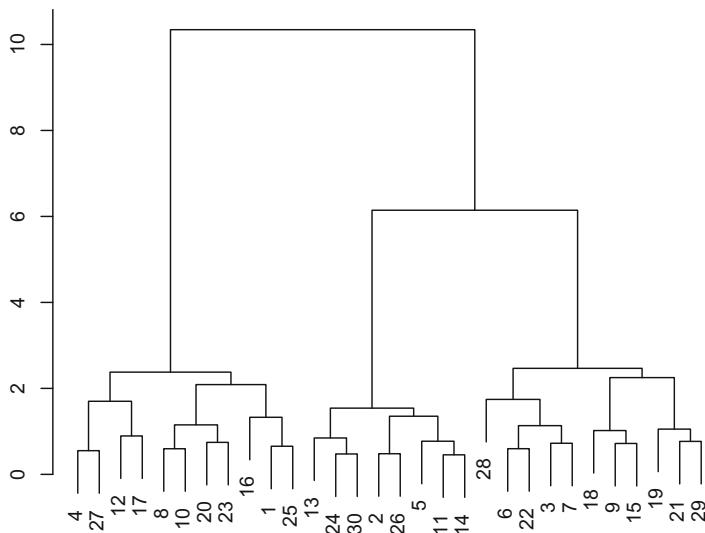


Fig. 7.20 Hierarchical clustering of the 30 k -means cluster centres of the artificial mobile phone data set

The R commands presented in this section may be slightly less convenient to use than the fully automated two-step procedure within SPSS. But they illustrate the key strength of R: the details of the algorithms used are known, and the data analyst can choose from the full range of hierarchical and partitioning clustering procedures available in R, rather than being limited to what has been implemented in a commercial statistical software package.

7.2.4.2 Bagged Clustering

Bagged clustering (Leisch 1998, 1999) also combines hierarchical clustering algorithms and partitioning clustering algorithms, but adds bootstrapping (Efron and Tibshirani 1993). Bootstrapping can be implemented by random drawing from the data set with replacement. That means that the process of extracting segments is repeated many times with randomly drawn (bootstrapped) samples of the data. Bootstrapping has the advantage of making the final segmentation solution less dependent on the exact people contained in consumer data.

In bagged clustering, we first cluster the bootstrapped data sets using a partitioning algorithm. The advantage of starting with a partitioning algorithm is that there are no restrictions on the sample size of the data. Next, we discard the original data set and all bootstrapped data sets. We only save the cluster centroids (segment representatives) resulting from the repeated partitioning cluster analyses. These cluster centroids serve as our data set for the second step: hierarchical clustering. The advantage of using hierarchical clustering in the second step is that the resulting dendrogram may provide clues about the best number of market segments to extract.

Bagged clustering is suitable in the following circumstances (Dolnicar and Leisch 2004; Leisch 1998):

- If we suspect the existence of niche markets.
- If we fear that standard algorithms might get stuck in bad local solutions.
- If we prefer hierarchical clustering, but the data set is too large.

Bagged clustering can identify niche segments because hierarchical clustering captures market niches as small distinct branches in the dendrogram. The increased chance of arriving at a good segmentation solution results from: (1) drawing many bootstrap samples from the original data set, (2) repeating the k -means analysis – or any other partitioning algorithm – many times to avoid a suboptimal initialisation (the random choice of initial segment representatives), (3) using only the centroids resulting from the k -means studies in the second (hierarchical) step of the analysis, and (4) using the deterministic hierarchical analysis in the final step.

Bagged clustering consists of five steps starting with a data set X of size n :

1. Create b bootstrap samples of size n by drawing with replacement consumers from the data set (using $b = 50$ or 100 bootstrap samples works well).
2. Repeat the preferred partitioning method for each bootstrap sample, generating $b \times k$ cluster centres (centroids, representatives of market segments) with k representing the number of clusters (segments). Leisch (1999) shows that the

exact number of clusters k selected is not important, as long as the number selected is higher than the number of segments expected to exist in the data. If k is larger than necessary, segments artificially split up in this step are merged during hierarchical clustering.

3. Use all cluster centres resulting from the repeated partitioning analyses to create a new, derived data set. Discard the original data. In the subsequent steps, replace the original data with the derived data set containing the cluster centres (centroids, representatives of market segments). It is for this reason that bagged clustering can deal with large data sets; it effectively discards the large data set once it has successfully extracted a number of cluster centres.
4. Calculate hierarchical clustering using the derived data set.
5. Determine the final segmentation solution by selecting a cut point for the dendrogram. Then, assign each original observation (consumer in the data set) to the market segment the representative of which is closest to that particular consumer.

Bagged clustering has been successfully applied to tourism data (Dolnicar and Leisch 2003; Prayag et al. 2015). For illustration purposes, we use the winter vacation activities data discussed in Dolnicar and Leisch (2003). The underlying marketing challenge for the Austrian winter tourist destination is to identify tourist market segments on the basis of their vacation activities. The available data set contains responses from 2961 tourists surveyed as part of the Austrian National Guest Survey (winter 1997/1998). Respondents indicated whether they have engaged in each of 27 winter vacation activities. As a consequence, 27 binary segmentation variables are available for market segmentation analysis. Activities include typical winter sports such as alpine skiing and ice skating, but also more generic tourist activities such as going to a spa or visiting museums. A detailed description of the data set is provided in Appendix C.2.

We first load the data set from package MSA, and inspect the labels of the 27 winter vacation activities used as segmentation variables:

```
R> data("winterActiv", package = "MSA")
R> colnames(winterActiv)
```

[1] "alpine skiing"	"cross-country skiing"
[3] "snowboarding"	"carving"
[5] "ski touring"	"ice-skating"
[7] "sleigh riding"	"tennis"
[9] "horseback riding"	"going to a spa "
[11] "using health facilities"	"hiking"
[13] "going for walks"	"organized excursions"
[15] "excursions"	"relaxing"
[17] "going out in the evening"	"going to discos/bars"
[19] "shopping"	"sight-seeing"
[21] "museums"	"theater/opera"
[23] "heurigen"	"concerts"
[25] "tyrolean evenings"	"local events"
[27] "pool sauna"	

We run bagged clustering using `bclust()` from package `flexclust`. We can specify the same number of `base.k = 10` market segments for the partitioning algorithm and `base.iter = 50` bootstrap samples as in Dolnicar and Leisch (2003) using the following R command:

```
R> set.seed(1234)
R> winter.bc <- bclust(winterActiv, base.k = 10,
+   base.iter = 50)

Committee Member:
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
40 41 42 43 44 45 46 47 48 49 50

Computing Hierarchical Clustering
```

`bclust` uses k -means as partitioning method, and the Euclidean distance together with average linkage in the hierarchical clustering part as the default.

Bagged clustering is an example of a so-called *ensemble clustering method* (Hornik 2005). These methods are called ensemble methods because they combine several segmentation solutions into one. Ensembles are also referred to as committees. Every repeated segment extraction using a different bootstrap sample contributes one committee member. The final step is equivalent to all committee members voting on the final market segmentation solution.

Figure 7.21 shows a dendrogram resulting from the second part of bagged clustering, the hierarchical cluster analysis of the $k \times b = 10 \times 50 = 500$ cluster centres (centroids, representatives of segments). This dendrogram appears

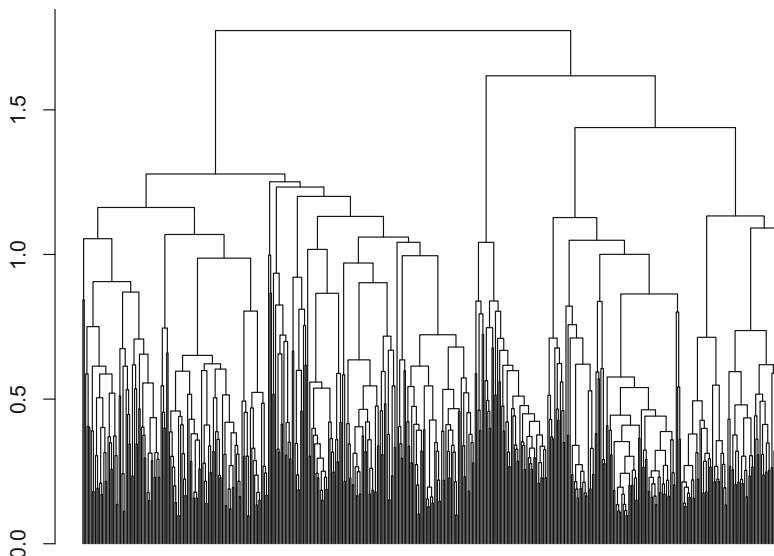


Fig. 7.21 Dendrogram for bagged cluster analysis of the winter vacation activities data set

to recommend four market segments. But assigning observations to these segments shows that the left branch of the dendrogram contains two thirds of all tourists. This large market segment is not very distinct.

Splitting this large segment up into two subsegments leads to a SNOWBOARD/PARTY SEGMENT and a SUPER ACTIVES segment. To gain insight into the characteristics of all resulting segments, we generate a bar chart (Fig. 7.22) using the following R command:

```
R> barchart(winter.bc, k = 5)
```

To inspect segmentation solutions containing fewer or more than five market segments, we can change the argument k to the desired number of clusters (number of segments).

Note that the bootstrapping procedure is based on artificial random numbers. Random number generators in R have changed over the last decade. As a consequence, the results presented here are not identical to those in Dolnicar and Leisch (2003), but qualitatively the same market segments emerge.

As can be seen from Fig. 7.22, the five segments extracted using bagged clustering vary substantially in size. The largest segment or cluster (segment 3) contains more than one third of all tourists in the sample. The smallest segment (segment 4) contains only 6%. This tiny segment is not particularly interesting from an organisational point of view, however: it is characterised by above average agreement with all vacation activities. As such, there is a risk that this segment may capture acquiescence response style (the tendency of survey respondents to agree with everything they are asked). Before selecting a segment of such nature as a target segment, it would have to be investigated (using other variables from the same survey) whether the profile is a reflection of overall high vacation activity or a response style.

The second smallest segment in this solution (segment 2) is still a niche segment, containing only 11% of respondents. Segment 2 displays some very interesting characteristics: members of this segment rarely go skiing. Instead, a large proportion of them goes to a spa or a health facility. They also go for walks, and hike more frequently than the average tourist visiting Austria in that particular winter season. Relaxation is also very high on the list of priorities for this market segment. Segment 2 (HEALTH TOURISTS) is a very interesting niche segment in the context of Austrian tourism. Austria has a large number of thermal baths built along thermal lines. Water from these hot thermal springs is believed to have health benefits. Thermal springs are popular, not only among people who are recovering from injuries, but also as a vacation or short break destination for (mainly older) tourists.

If the same data set had been analysed using a different algorithm, such as k -means, this niche segment of HEALTH TOURISTS would not have emerged.

An additional advantage of bagged clustering – compared to standard partitioning algorithms – is that the two-step process effectively has a built-in variable uncertainty analysis. This analysis provides element-wise uncertainty bands for the cluster centres. These bands are shown in Fig. 7.23, which contains a boxplot of the

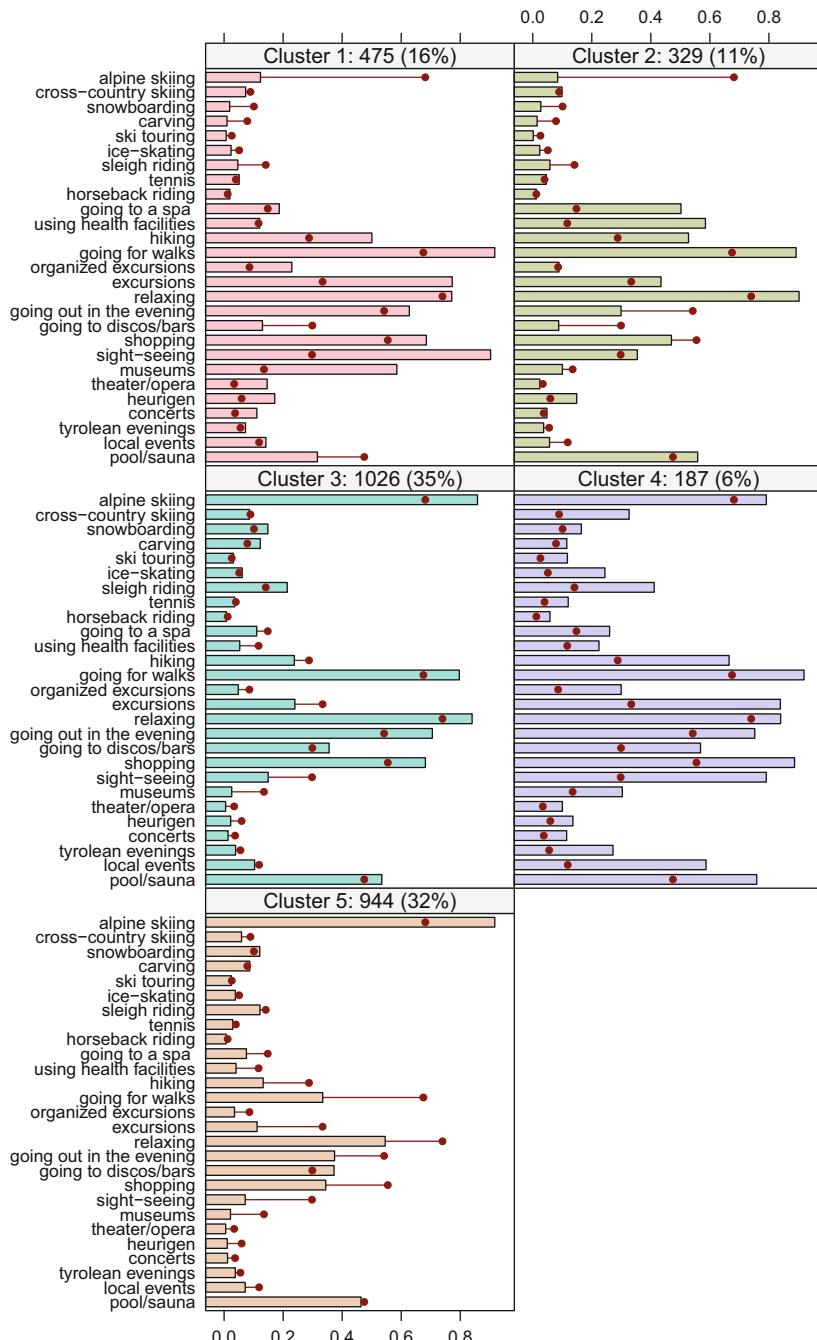


Fig. 7.22 Bar chart of cluster means from bagged cluster analysis of the winter vacation activities data set

133 cluster centres (centroids, representatives of market segments) forming segment 5 as generated by

```
R> bwplot(winter.bc, k = 5, clusters = 5)
```

Here, only the plot for segment 5 is provided. The same R code can generate boxplots for all other market segments resulting from bagged clustering.

A general explanation of boxplots and how they are interpreted is provided in Sect. 6.3 using Fig. 6.2. Looking at Fig. 7.23: if the 133 cluster centres are spread across the full width of the plot for a specific vacation activity, it indicates that the market segment is not very distinct with respect to this activity. If, however, all cluster centres are lumped together, this is a key characteristic of this particular market segment.

As can be seen in Fig. 7.23, cluster centres assigned to segment 5 display little variation with respect to a number of variables: going skiing (which most of them do), a range of cultural activities (which most of them do not engage in), and a few other activities, such as horseback riding and organised excursions.

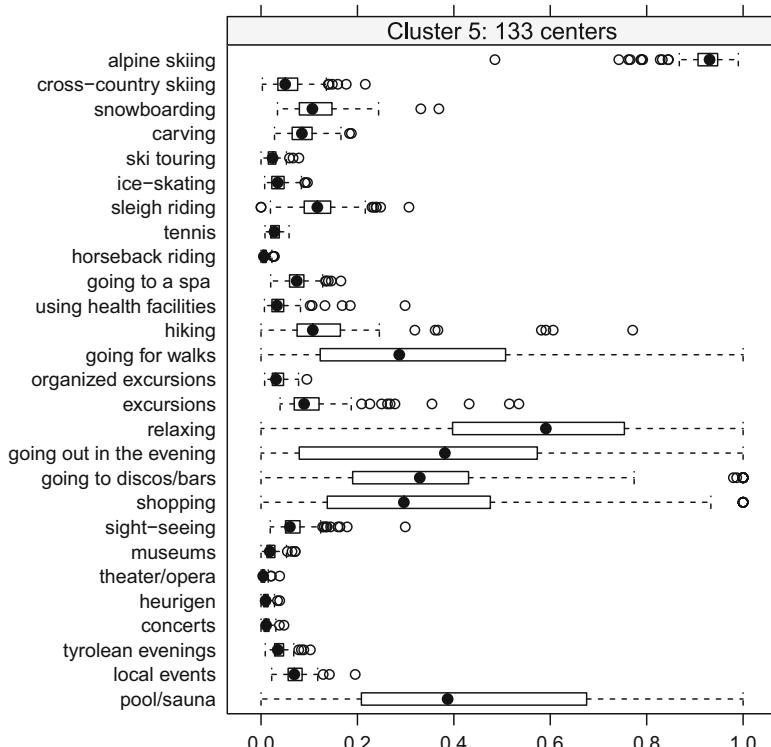


Fig. 7.23 Boxplot of cluster centres from bagged cluster analysis for segment 5 of the winter vacation activities data set

With respect to other vacation activities, however, there is a lot of variation among the cluster centres assigned to segment 5, including relaxation, going out in the evening, going to discos and bars, shopping, and going to the pool or sauna.

Note that the marginal probabilities in the total population for alpine skiing and relaxing are almost the same (both approximately 70%). The difference in variability is therefore not simply an artefact of how many people undertake these activities overall. Low variability in unpopular winter activities, on the other hand, is not unexpected: if almost nobody in the total tourist population goes horseback riding, it is not a key insight that cluster centres assigned to segment 5 do not go horseback riding either.

7.3 Model-Based Methods

Distance-based methods have a long history of being used in market segmentation analysis. More recently, model-based methods have been proposed as an alternative. According to Wedel and Kamakura (2000, p. XIX) – the pioneers of model-based methods in market segmentation analysis – mixture methodologies have attracted great interest from applied marketing researchers and consultants. Wedel and Kamakura (2000, p. XIX) predict that in terms of impact on academics and practitioners, next to conjoint analysis, mixture models will prove to be the most influential methodological development spawned by marketing problems to date.

Here, a slightly more pragmatic perspective is taken. Model-based methods are viewed as one additional segment extraction method available to data analysts. Given that extracting market segments is an exploratory exercise, it is helpful to use a range of extraction methods to determine the most suitable approach for the data at hand. Having model-based methods available is particularly useful because these methods extract market segments in a very different way, thus genuinely offering an alternative extraction technique.

As opposed to distance-based clustering methods, model-based segment extraction methods do not use similarities or distances to assess which consumers should be assigned to the same market segment. Instead, they are based on the assumption that the true market segmentation solution – which is unknown – has the following two general properties: (1) each market segment has a certain size, and (2) if a consumer belongs to market segment A , that consumer will have characteristics which are specific to members of market segment A . These two properties are assumed to hold, but the exact nature of these properties – the sizes of these segments, and the values of the segment-specific characteristics – is not known in advance. Model-based methods use the empirical data to find those values for segment sizes and segment-specific characteristics that best reflect the data.

Model-based methods can be seen as selecting a general structure, and then fine-tuning the structure based on the consumer data. The model-based methods used in this section are called *finite mixture models* because the number of market segments is finite, and the overall model is a mixture of segment-specific models. The two

properties of the finite mixture model can be written down in a more formal way. Property 1 (that each market segment has a certain size) implies that the segment membership z of a consumer is determined by the multinomial distribution with segment sizes π :

$$z \sim \text{Multinomial}(\pi).$$

Property 2 states that members of each market segment have segment-specific characteristics. These segment-specific characteristics are captured by the vector θ , containing one value for each segment-specific characteristic. Function $f()$, together with θ , captures how likely specific values y are to be observed in the empirical data, given that the consumer has segment membership z , and potentially given some additional pieces of information x for that consumer:

$$f(y|x, \theta_z).$$

These functions $f()$ together with their parameters θ are also referred to as segment-specific models and correspond to statistical distribution functions.

This leads to the following finite mixture model:

$$\sum_{h=1}^k \pi_h f(y|x, \theta_h), \quad \pi_h > 0, \quad \sum_{h=1}^k \pi_h = 1. \quad (7.1)$$

The values to be estimated – across all segments h ranging from 1 to k – consist of the segment sizes π (positive values summing to one), and the segment-specific characteristics θ . The values that need to be estimated are called parameters.

Different statistical frameworks are available for estimating the parameters of the finite mixture model. Maximum likelihood estimation (see for example Casella and Berger 2010) is commonly used. Maximum likelihood estimation aims at determining the parameter values for which the observed data is most likely to occur. The maximum likelihood estimate has a range of desirable statistical properties. The likelihood is given by interpreting the function in Eq. 7.1 as a function of the parameters instead of the data. However, even for the simplest mixture models, this likelihood function cannot be maximised in closed form. Iterative methods are required such as the EM algorithm (Dempster et al. 1977; McLachlan and Basford 1988; McLachlan and Peel 2000). This approach regards the segment memberships z as missing data, and exploits the fact that the likelihood of the complete data (where also the segment memberships are included as observed data) is easier to maximise. An alternative statistical inference approach is to use the Bayesian framework for estimation. If a Bayesian approach is pursued, mixture models are usually fitted using Markov chain Monte Carlo methods (see for example Frühwirth-Schnatter 2006).

Regardless of the way the finite mixture model is estimated, once values for the segment sizes, and the segment-specific characteristics are determined (for example

using the maximum likelihood or the posterior mode estimates), consumers in the empirical data set can be assigned to segments using the following approach. First, the probability of each consumer to be a member of each segment is determined. This is based on the information available for the consumer, which consists of y , the potentially available x , and the estimated parameter values of the finite mixture model:

$$\text{Prob}(z = h|x, y, \pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k) = \frac{\pi_h f(y|x, \theta_h)}{\sum_j^k \pi_j f(y|x, \theta_j)} \quad (7.2)$$

The consumers are then assigned to segments using these probabilities by selecting the segment with the highest probability.

As is the case with partitioning clustering methods, maximum likelihood estimation of the finite mixture model with the EM algorithm requires specifying the number of segments k to extract in advance. But the true number of segments is rarely known. A standard strategy to select a good number of market segments is to extract finite mixture models with a varying number of segments and compare them. Selecting the correct number of segments is as problematic in model-based methods as it is to select the correct number of clusters when using partitioning methods.

In the framework of maximum likelihood estimation, so-called *information criteria* are typically used to guide the data analyst in their choice of the number of market segments. Most common are the Akaike information criterion or AIC (Akaike 1987), the Bayesian information criterion or BIC (Schwarz 1978; Fraley and Raftery 1998), and the integrated completed likelihood or ICL (Biernacki et al. 2000). All these criteria use the likelihood as a measure of goodness-of-fit of the model to the data, and penalise for the number of parameters estimated. This penalisation is necessary because the maximum likelihood value increases as the model becomes more complex (more segments, more independent variables). Comparing models of different complexity using maximum likelihoods will therefore always lead to the recommendation of the larger model. The criteria differ in the exact value of the penalty. The specific formulae for AIC, BIC and ICL are given by:

$$\text{AIC} = 2df - 2\log(L) \quad (7.3)$$

$$\text{BIC} = \log(n)df - 2\log(L) \quad (7.4)$$

$$\text{ICL} = \log(n)df - 2\log(L) + 2ent \quad (7.5)$$

where df is the number of all parameters of the model, $\log(L)$ is the maximised log-likelihood, and n is the number of observations. ent is the mean entropy (Shannon 1948) of the probabilities given in Eq. 7.2. Mean entropy decreases if the assignment of observations to segments is clear. The entropy is lowest if a consumer has a 100% probability of being assigned to a certain segment. Mean entropy increases if segment assignments are not clear. The entropy is highest if a consumer has the same probability of being a member of each market segment.

All criteria decrease if fewer parameters are used or the likelihood increases. In contrast, more parameters or smaller likelihoods will increase them. The goal is to minimise them. Because $\log(n)$ is larger than 2 for n larger than 7, BIC penalises stronger than AIC for additional parameters, and prefers smaller models in case different model sizes are recommended. The ICL uses an additional penalty to the BIC, which takes the separatedness of segments into account. In addition to these three criteria, a number of other information criteria have been proposed; no one specific information criterion has been shown to consistently outperform the others in model-based clustering applications.

At first glance, finite mixture models may appear unnecessarily complicated. The advantage of using such models is that they can capture very complex segment characteristics, and can be extended in many different ways. One possible extension of the presented finite mixture model includes a model where the segment-specific models differ not only in the segment characteristics θ , but also in the general structure. There is an extensive literature available on finite mixture models including several research monographs (see for example McLachlan and Peel 2000; Frühwirth-Schnatter 2006). The finite mixture model literature uses the following terminology: market segments are referred to as *mixture components*, segment sizes as *prior probabilities* or component sizes, and the probability of each consumer to be a member of each segment given in Eq. 7.2 as *posterior probability*.

7.3.1 Finite Mixtures of Distributions

The simplest case of model-based clustering has no independent variables x , and simply fits a distribution to y . To compare this with distance-based methods, finite mixtures of distributions basically use the same segmentation variables: a number of pieces of information about consumers, such as the activities they engage in when on vacation. No additional information about these consumers, such as total travel expenditures, is simultaneously included in the model.

The finite mixture model reduces to

$$\sum_{h=1}^k \pi_h f(y|\theta_h), \quad \pi_h \geq 0, \quad \sum_{h=1}^k \pi_h = 1. \quad (7.6)$$

The formulae are the same as in Eq. 7.1, the only difference is that there is no x . The statistical distribution function $f()$ depends on the measurement level or scale of the segmentation variables y .

7.3.1.1 Normal Distributions

For metric data, the most popular finite mixture model is a mixture of several multivariate normal distributions. The multivariate normal distribution can easily model covariance between variables; and approximate multivariate normal distributions occur in both biology and business. For example, physical measurements on humans like height, arm length, leg length or foot length are almost perfectly modelled by a multivariate normal distribution. All these variables have an approximate univariate normal distribution individually, but are not independent of each other. Taller people have longer arms, longer legs and bigger feet. All measurements are positively correlated. An example from business is that prices in markets with many players can be modelled using (log-)normal distributions. In sum, a mixture of normal distributions can be used for market segmentation when the segmentation variables are metric, for example: money spent on different consumption categories, time spent engaging in different vacation activities, or body measurements for the segments of different clothes sizes.

Mathematically, $f()$ in Eq. 7.6 is the multivariate normal distribution which has two sets of parameters (mean and variance) like the univariate normal distribution. If p segmentation variables are used, these have p mean values, and each segment has a segment-specific mean vector μ_h of length p . In addition to the p variances of the p segmentation variables, the covariance structure can be modelled, resulting in a $p \times p$ covariance matrix Σ_h for each segment. The covariance matrix Σ_h contains the variances of the p segmentation variables in the diagonal and the covariances between pairs of segmentation variables in the other entries. The covariance matrix is symmetric, and contains $p(p + 1)/2$ unique values.

The segment-specific parameters θ_h are the combination of the mean vector μ_h and the covariance matrix Σ_h , and the number of parameters to estimate is $p + p(p + 1)/2$.

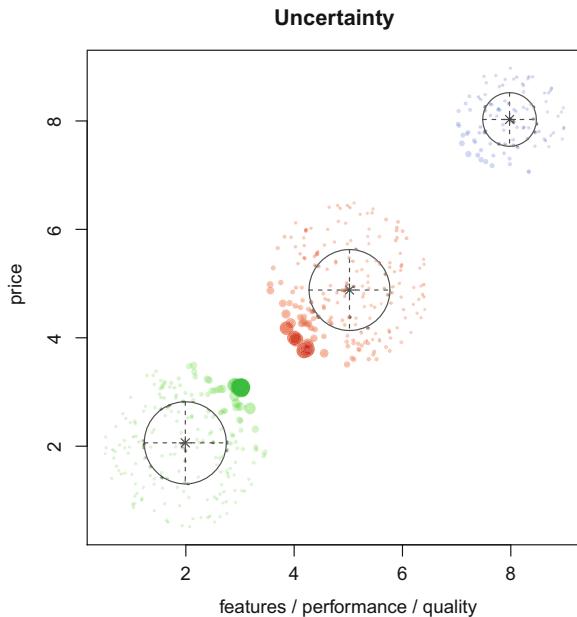
Mixtures of normal distributions can be illustrated using the simple artificial mobile phone data set presented in Sect. 7.2.3 and shown in Fig. 7.9:

```
R> library("flexclust")
R> set.seed(1234)
R> PF3 <- priceFeature(500, which = "3clust")
```

Fitting a mixture of normal distributions is best done in R with package `mclust` (Fraley et al. 2012; Fraley and Raftery 2002). Function `Mclust` fits models for different numbers of segments using the EM algorithm. Initialisation is deterministic using the partition inferred from a hierarchical clustering approach with a likelihood-based distance measure. Here, we extract two to eight market segments (argument `G` for number of segments):

```
R> library("mclust")
R> PF3.m28 <- Mclust(PF3, G = 2:8)
R> PF3.m28
```

Fig. 7.24 Uncertainty plot of the mixture of normal distributions for the artificial mobile phone data set



```
'Mclust' model object:  
best model: spherical, varying volume (VII) with  
3 components
```

Ignoring the statement about “spherical, varying volume (VII)” for the moment, we see that the BIC correctly recommends extracting three segments.

Figure 7.24 shows the market segments resulting from the mixture of normal distributions for the artificial mobile phone. We obtain this plot using the following R command:

```
R> plot(PF3.m28, what = "uncertainty")
```

The plot in Fig. 7.24 is referred to as an *uncertainty plot*. The uncertainty plot illustrates the ambiguity of segment assignment. A consumer who cannot be clearly assigned to one of the market segments is considered uncertain. The further away from 1 a consumer's maximum segment assignment probability is (as determined using Eq. 7.2), the less certain is the segment assignment. The uncertainty plot is a useful visualisation alerting the data analyst to solutions that do not induce clear partitions, and pointing to market segments being artificially created, rather than reflecting the existence of natural market segments in the data. The uncertainty plot consists of a scatter plot of observations (consumers). The colours of the observations indicate segment assignments. Larger solid coloured bubbles have higher assignment uncertainty. The means and covariance matrices of the segments are superimposed to provide insights into the fitted mixture of normal distributions.

The “spherical, varying volume (VII)” part of the Mclust output indicates which specific mixture model of normal distributions is selected according to the BIC. Model selection for mixtures of normal distributions does not only

require selecting the number of segments, but also choosing an appropriate shape of the covariance matrices of the segments.

For two-dimensional data (like in the mobile phone example), each market segment can be shaped like an ellipse. The ellipses can have different shapes, areas and orientations. The ellipse corresponding to one market segment could be very flat and point from bottom left to top right, while another one could be a perfect circle. For the mobile phone data set, the procedure correctly identifies that the ellipses are shaped as circles. But the areas covered by the three circles are not the same. The segment in the top right corner is less spread out and more compact.

A circle with more than two dimensions is a sphere. The area covered by a sphere is its volume. The “spherical, varying volume (VII)” part uses the terms for higher dimensional spaces because the dimensionality is larger than two in most applications. The output indicates that spherical covariance matrices are used for the segments but with different volume. This selected shape for the covariance matrices is shown in Fig. 7.24, where the axes of the ellipses are parallel to the coordinate axes, and have the shape of a circle.

Spherical covariance structures correspond to covariance matrices where only the main diagonal elements are non-zero, and they all have the same value. So – instead of $p(p + 1)/2$ parameters – only one parameter has to be estimated for each covariance matrix: the radius of the sphere (circle in the 2-dimensional example). If it were known in advance that only spherical clusters are present in the data, the task of fitting the mixture of normal distributions would be much simpler because fewer parameters have to be estimated.

The covariance matrices of the mixture of normal distributions used for the segments strongly affect the number of parameters that need to be estimated. Given that each Σ_h contains $p(p + 1)/2$ parameters for p segmentation variables, the number of parameters that has to be estimated grows quadratically with the number of segmentation variables p .

The simple mobile phone example contains only two segmentation variables ($p = 2$). The number of parameters for each market segment is 2 (length of μ_h) plus 3 (symmetric 2×2 matrix Σ_h), which sums up to $2 + 3 = 5$. If three market segments are extracted, a total of $3 \times 5 = 15$ parameters have to be estimated for the segments, plus two segment sizes (the three π_h have to sum up to one, such that $\pi_3 = 1 - \pi_1 - \pi_2$). In sum, a mixture of normal distributions with three segments for the artificial mobile phone data set has $15 + 2 = 17$ parameters.

If ten segmentation variables are used ($p = 10$), the number of parameters that need to be estimated increases to 10 mean values, covariance matrices with $10 \times 11/2 = 55$ parameters, and $10 + 55 = 65$ parameters per segment. For a three-segment model this means that $3 \times 65 + 2 = 197$ parameters have to be estimated. As a consequence, large sample sizes are required to ensure reliable estimates.

To reduce the number of parameters to estimate, package `mclust` imposes restrictions on the covariance matrices. One possible restriction is to use spherical instead of ellipsoidal covariances, such that only a single radius has to be estimated for each segment. An even more parsimonious model restricts all spheres for all segments to having the same radius (and hence the same volume).

Table 7.3 The 14 covariance models available in package `mclust`

EII	Spherical, equal volume
VII	Spherical, unequal volume
EEI	Diagonal, equal volume and shape
VEI	Diagonal, varying volume, equal shape
EVI	Diagonal, equal volume, varying shape
VVI	Diagonal, varying volume and shape
EEE	Ellipsoidal, equal volume, shape, and orientation
EVE	Ellipsoidal, equal volume and orientation
VEE	Ellipsoidal, equal shape and orientation
VVE	Ellipsoidal, equal orientation
EEV	Ellipsoidal, equal volume and equal shape
VEV	Ellipsoidal, equal shape
EVV	Ellipsoidal, equal volume
VVV	Ellipsoidal, varying volume, shape, and orientation

By default, `Mclust` tries a full model where all segments have different covariance matrices without any restrictions (called model VVV in Table 7.3 for varying volume, shape, and orientation). In addition, 13 restricted models are estimated: the smallest model assumes identical spheres for all segments (EII, spherical, equal volume). A list of all models is shown in Table 7.3, and illustrated in Fig. 7.25. Mathematical details are provided in the `mclust` documentation.

The BIC values obtained for each of the resulting models for different numbers of segments are shown in Fig. 7.26. We obtain this plot using:

```
R> plot(PF3.m28, what = "BIC")
```

R package `mclust` uses the negative BIC values (instead of the BIC values defined in Eq. 7.4), but refers to them as BIC values. It makes no difference to the results, except that we now want to maximise, not minimise the BIC.

Figure 7.26 plots the BIC value along the y-axis, and the number of segments (ranging from 2 to 8) along the x-axis. The BIC values obtained for each covariance model are joined using lines. The different colours and point characters used for each of the covariance models are indicated in the legend in the bottom right corner. As can be seen, BIC values are low for two segments, then dramatically increase for three segments, and show no further significant improvement for solutions with more than three segments. The BIC therefore recommends a spherical, varying volume (VII) model with three segments. This leads to selecting a model that allows to extract the three well-separated, distinct segments using a parsimonious mixture model. Unfortunately, if empirical consumer data is used as the basis for market segmentation analysis, it is not always possible to easily assess the quality of the recommendation made by information criteria such as the BIC.

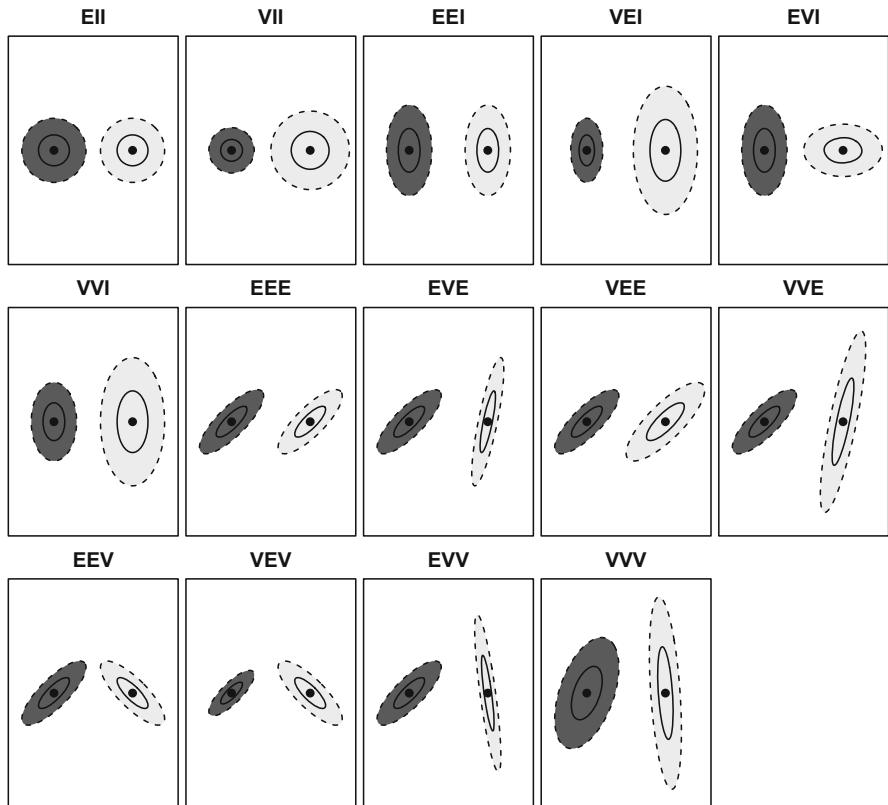
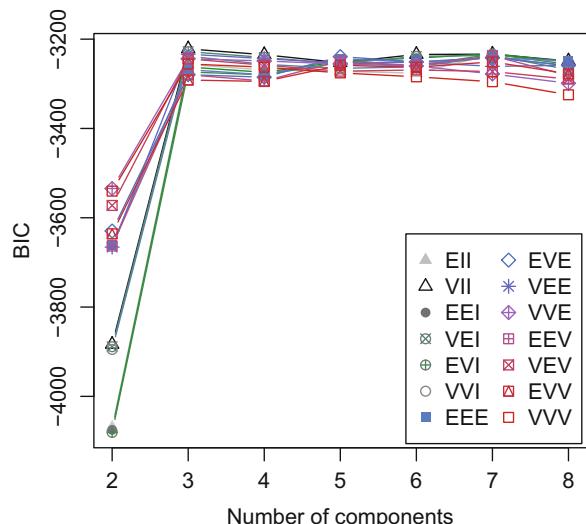


Fig. 7.25 Visualisation of the 14 covariance models available in package mclust

Fig. 7.26 BIC values of the mixtures of normal distributions for the artificial mobile phone data set



Example: Australian Vacation Motives

In addition to their vacation motives, survey respondents also answered a range of other questions. These answers are contained in the data frame `vacmotdesc`. The following three metric variables are available: moral obligation score, NEP score, and environmental behaviour score on vacation. We load the data set and extract the metric variables using:

```
R> data("vacmot", package = "flexclust")
R> vacmet <- vacmotdesc[, c("Obligation", "NEP",
+   "Vacation.Behaviour")]
R> vacmet <- na.omit(vacmet)
```

Because variable `VACATION.BEHAVIOUR` contains missing values, we remove respondents with missing values using `na.omit`. We then visualise the data:

```
R> pairs(vacmet, pch = 19, col = rgb(0, 0, 0, 0.2))
```

Solid points are drawn using `pch = 19`. To avoid losing information due to overplotting, the points are black with transparency using `rgb(0, 0, 0, 0.2)` with an α -shading value of 0.2. Figure 7.27 indicates that no clearly separated segments exist in the data.

Command `Mclust` fits all 14 different covariance matrix models by default, and returns the best model with respect to the BIC:

```
R> vacmet.m18 <- Mclust(vacmet, G = 1:8)
```

Alternatively, `Mclust` can fit only selected covariance matrix models. In the example below, we fit only covariance models where the covariance matrices have equal volume, shape and orientation over segments. We can look up those model names in Table 7.3:

```
R> vacmet.m18.equal <- Mclust(vacmet, G = 1:8,
+   modelNames = c("EEI", "EII", "EEE"))
```

The best models according to the BIC are:

```
R> vacmet.m18
'Mclust' model object:
best model:
  ellipsoidal, equal shape and orientation (VEE)
  with 2 components

R> vacmet.m18.equal
'Mclust' model object:
best model:
  ellipsoidal, equal volume, shape and orientation (EEE)
  with 3 components
```

Results indicate that – in the case where all 14 different covariance matrices are considered – a mixture model with two segments is selected. In the restricted case,

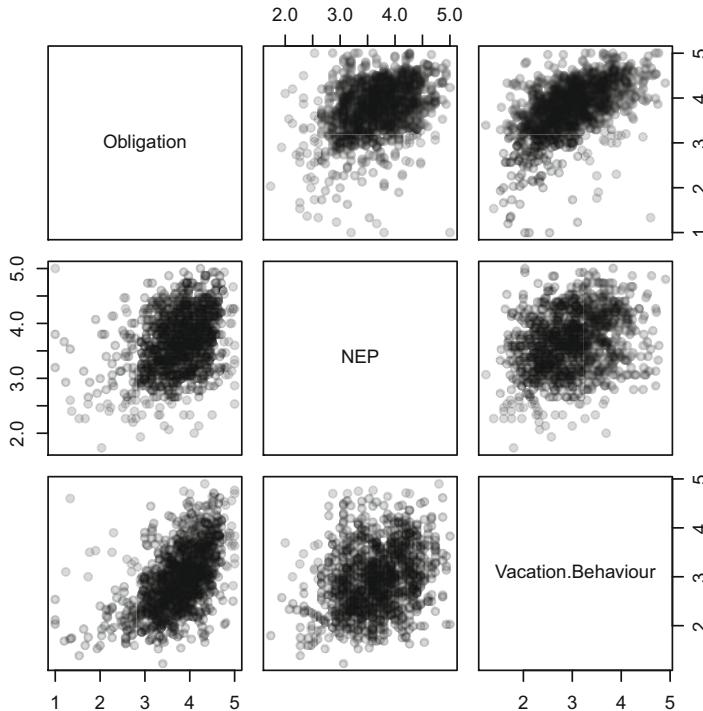


Fig. 7.27 Scatter plot of the metric variables in the Australian travel motives data set

a model with three segments emerges. Figures 7.28 and 7.29 visualise the fitted models using *classification plots*. The classification plot is similar to the uncertainty plot, except that all data points are of the same size regardless of their uncertainty of assignment.

```
R> plot(vacmet.m18, what = "classification")
R> plot(vacmet.m18.equal, what = "classification")
```

In both selected mixture models, the covariance matrices have identical orientation and shape. This implies that the correlation structure between the variables is the same across segments. However, in the case where all covariance models are considered, the covariance matrices differ in volume. Using mixtures of normal distributions means that the data points are not assigned to the segment where the mean is closest in Euclidean space (as is the case for k -means clustering). Rather, the distance induced by the covariance matrices (Mahalanobis distance) is used, and the segment sizes are taken into account. Assigning segment membership in this way implies that observations are not necessarily assigned to the segment representative closest to them in Euclidean space. However, restricting covariance matrices to be identical over segments at least ensures that the same distance measure is used

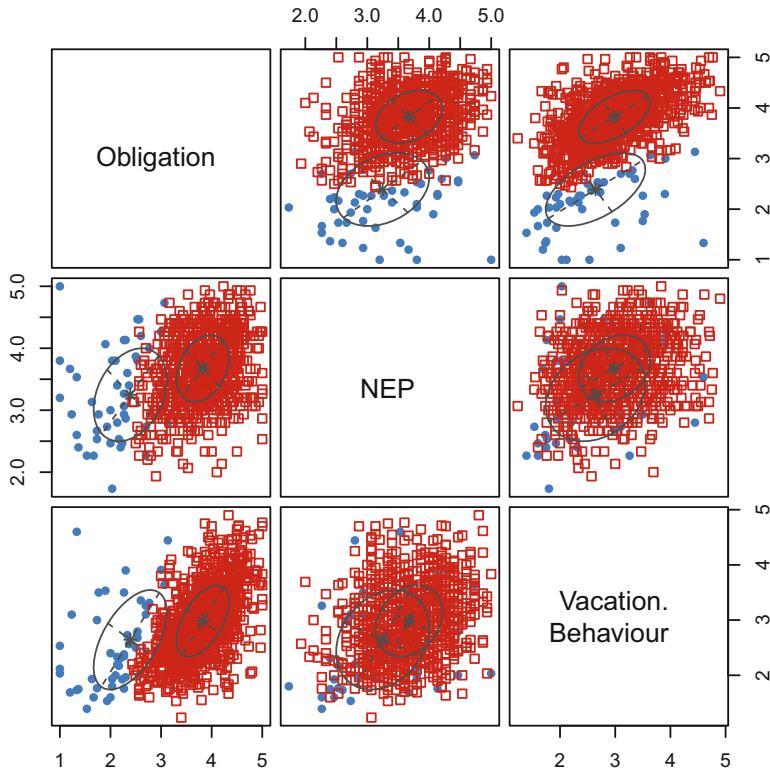


Fig. 7.28 Classification plot of the mixture of normal distributions for the Australian travel motives data set selected using the BIC among all covariance models

for all segment representatives for segment membership assignment except for the differences in segment sizes.

7.3.1.2 Binary Distributions

For binary data, finite mixtures of binary distributions, sometimes also referred to as latent class models or latent class analysis (Bhatnagar and Ghose 2004; Kemperman and Timmermanns 2006; Campbell et al. 2014) are popular. In this case, the p segmentation variables in the vector y are not metric, but binary (meaning that all p elements of y are either 0 or 1). The elements of y , the segmentation variables, could be vacation activities where a value of 1 indicates that a tourist undertakes this activity, and a value of 0 indicates that they do not.

The mixture model assumes that respondents in different segments have different probabilities of undertaking certain activities. For example, some respondents may be interested in alpine skiing and not interested in sight-seeing. This leads to these

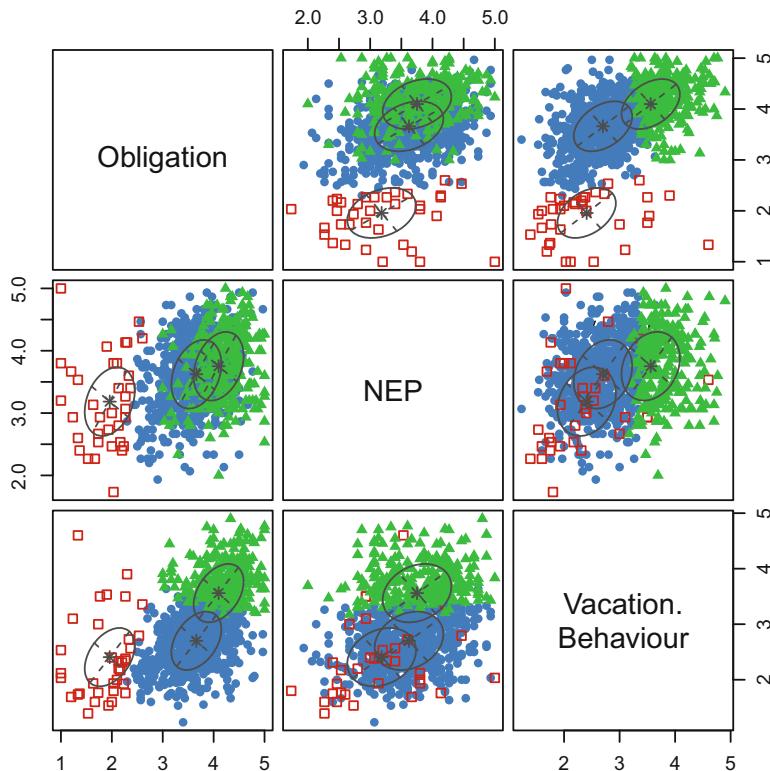


Fig. 7.29 Classification plot of the mixture of normal distributions for the Australian travel motives data set selected using the BIC among the models with identical covariance matrices across segments

two variables being negatively correlated in the overall data set. However, this correlation is due to groups of respondents interested in one of the two activities only.

To illustrate mixtures of binary distributions, we use the data set containing winter activities of Austrian tourists (introduced in the context of bagged clustering in Sect. 7.2.4). We first investigate the observed frequency patterns for the variables ALPINE SKIING and SIGHT-SEEING:

```
R> data("winterActiv", package = "MSA")
R> winterActiv2 <- winterActiv[, c("alpine skiing",
+ "sight-seeing")]
R> table(as.data.frame(winterActiv2))

sight-seeing
alpine skiing   0     1
              0 416  527
              1 1663 355
```

Of the 2961 respondents, only 355 (12%) stated they engaged in both activities. If the two activities were not associated, we would expect this percentage to be much higher:

```
R> p <- colMeans(winterActiv2)
R> p
alpine skiing sight-seeing
0.6815265      0.2978723

R> round(prod(p) * 100)
[1] 20
```

The expected percentage is 20%. This indicates an association between the two variables across the complete data set. The expected counts for the patterns (given the overall mean activity levels for the two activities) are:

```
R> n <- nrow(winterActiv2)
R> expected <- function(p) {
+   res <- outer(c(1 - p[1], p[1]), c(1 - p[2], p[2]))
+   dimnames(res) <- setNames(rep(list(c("0", "1")), 2),
+     names(p)))
+   res
+ }
R> round(n * expected(p))

           sight-seeing
alpine skiing      0    1
                  0 662 281
                  1 1417 601
```

The model of independent binary distributions does not represent the data well (as indicated by the discrepancy between the observed and expected frequencies). We thus fit a mixture of binary distributions to the data. The expected frequencies of a suitable mixture model should correspond to the observed frequencies.

The R package **flexmix** (Leisch 2004; Grün and Leisch 2008) implements a general framework for mixture modelling for a wide variety of segment models, including mixtures of regression models (see Sect. 7.3.2). We use function **flexmix** to fit the mixture model with one single run of the EM algorithm. We need to specify the dependent (`winterActiv2`) and the independent variables (1) using the formula interface. The formula is of the form $y \sim x$ where y are the dependent variables, and x are the independent variables. Because mixtures of distributions do not contain any independent variables x (see Eq. 7.6), the formula used for mixtures of distributions is $y \sim 1$. Here, we extract two market segments ($k = 2$), and we use independent binary distributions as segment-specific model (**FLXMCmvbinary**):

```
R> library("flexmix")
R> winterActiv2.m2 <- flexmix(winterActiv2 ~ 1, k = 2,
+   model = FLXMCmvbinary())
```

Function `flexmix()` initialises the EM algorithm by randomly assigning probabilities for each consumer to be a member of each of the market segments. The EM algorithm can get stuck in local optima of the likelihood. We can avoid that by using several random starts with different initialisations, and retain the solution with the highest likelihood using the function `stepFlexmix`. We specify the number of random restarts using `nrep = 10` for ten random restarts. The random restart procedure is undertaken for the full range of market segments specified, in this case 1 to 4 (`k = 1:4`). The argument `verbose = FALSE` prevents progress information on the calculations to be printed.

```
R> winterActiv2.m14 <- stepFlexmix(winterActiv2 ~ 1,
+   k = 1:4, model = FLXMCmvbinary(), nrep = 10,
+   verbose = FALSE)
R> winterActiv2.m14

Call:
stepFlexmix(winterActiv2 ~ 1, model = FLXMCmvbinary(),
  k = 1:4, nrep = 10, verbose = FALSE)

iter converged k k0      logLik       AIC       BIC       ICL
1     2      TRUE 1 1 -3656.137 7316.274 7328.260 7328.260
2    30      TRUE 2 2 -3438.491 6886.982 6916.948 7660.569
3    22      TRUE 3 3 -3438.490 6892.981 6940.927 10089.526
4    21      TRUE 4 4 -3438.490 6898.980 6964.907 10979.912
```

The output shows summary information for each of the four models fitted for different numbers of segments (`k = 1:4`). These four models are those resulting from the best of 10 restarts. The summary information consists of: the number of iterations of the EM algorithm until convergence (`iter`), whether or not the EM algorithm converged (`converged`), the number of segments in the fitted model (`k`), the number of segments initially specified (`k0`), the log-likelihood obtained (`logLik`), and the values for the information criteria (AIC, BIC and ICL). By default, package `flexmix` removes small segments when running the EM algorithm. Small segments can cause numeric problems in the estimation of the parameters because of the limited number of observations (consumers). We can add the argument `control = list(minprior = 0)` to the call of `stepFlexmix()` to avoid losing small segments. This argument specification ensures that `k` is equal to `k0`.

Results indicate EM algorithm convergence for all models. The number of segments in the final models are the same as the number used for initialisation. The log-likelihood increases strongly when going from one to two segments, but remains approximately the same for more segments. All information criteria except for the ICL suggest using a mixture with two segments. The best model with respect to the BIC results from:

```
R> best.winterActiv2.m14 <- getModel(winterActiv2.m14)
```

By default, the BIC value recommends a model. We can use the AIC by setting `which = "AIC"`. We can specify the number of segments with `which = "2"`. The following command returns basic information on this two-segment model:

```
R> best.winterActiv2.m14
```

```

Call:
stepFlexmix(winterActiv2 ~ 1, model = FLXMCmvbinary(),
k = 2, nrep = 10, verbose = FALSE)

Cluster sizes:
 1   2
1298 1663

convergence after 30 iterations

```

This basic information contains the number of consumers assigned to each segment and the number of iterations required to reach convergence.

The parameters of the segment-specific models are the probabilities of observing a 1 in each of the variables. These probabilities characterise the segments, and have the same interpretation as centroids in k -means clustering of binary data. They are used in the same way to create tables and figures of segment profiles, as discussed in detail in Step 6. We obtain the probabilities using:

```
R> p <- parameters(best.winterActiv2.m14)
R> p
```

	Comp.1	Comp.2
center.alpine.skiing	0.3531073	0.94334159
center.sight-seeing	0.6147303	0.04527384

Segment 1 (denoted as Comp.1) contains respondents with a high likelihood to go sight-seeing, and a low probability of going alpine skiing. Respondents in segment 2 (Comp.2) go alpine skiing, and are not interested in sight-seeing.

The expected table of frequencies given this fitted model results from:

```
R> pi <- prior(best.winterActiv2.m14)
R> pi

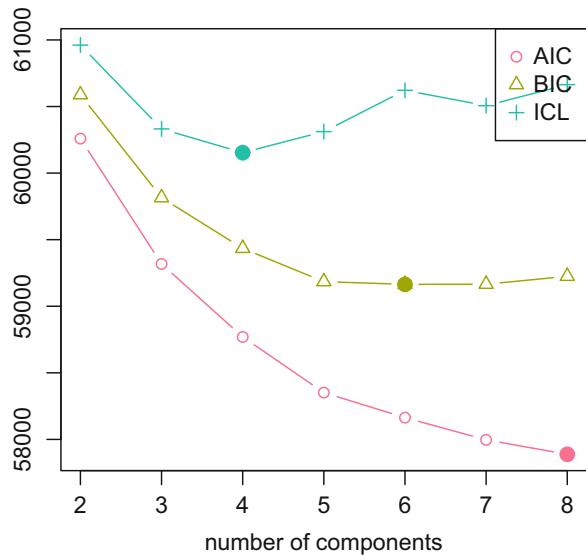
[1] 0.4435012 0.5564988

R> round(n * (pi[1] * expected(p[, "Comp.1"]) +
+ pi[2] * expected(p[, "Comp.2"])))

center.sight-seeing
center.alpine.skiing      0     1
                           0 416 526
                           1 1663 355
```

The table of expected frequencies is similar to the table of observed frequencies. Using the mixture model, the association between the two variables is explained by the segments. Within each segment, the two variables are not associated. But the fact that members of the segments differ in their vacation activity patterns, leads to the association of the two variables across all consumers.

Fig. 7.30 AIC, BIC and ICL values of mixtures of binary distributions for the winter vacation activities data set



Example: Austrian Winter Vacation Activities

We fit a mixture of binary distributions to the data set containing 27 winter activities. We vary the number of segments from 2 to 8, and use 10 random initialisations with the EM algorithm:

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

Figure 7.30 shows AIC, BIC and ICL curves for 2 to 8 segments, obtained by:

```
R> plot(winter.m28)
```

Figure 7.30 plots the number of market segments (components) along the x -axis, and the values of the information criteria along the y -axis. Lower values of information criteria are better. Inspecting the development of the values of all three information criteria in Fig. 7.30 leads to the following conclusions: ICL recommends 4 market segments (components); BIC recommends 6 segments, but displays a major decrease only up to 5 segments; and AIC suggests at least 8 market segments.

We choose the five-segment solution for closer inspection because it represents a compromise between the recommendations made by BIC and ICL:

```
R> winter.m5 <- getModel(winter.m28, "5")
R> winter.m5

Call:
stepFlexmix(winterActiv ~ 1, model = FLXMCmvbinary(),
```

```

k = 5, nrep = 10, verbose = FALSE)

Cluster sizes:
 1   2   3   4   5
912 414 200 218 1217

convergence after 67 iterations

```

The command `parameters(winter.m5)` extracts the fitted probabilities of the mixture model. Function `propBarchart` from package `flexclust` creates a chart similar to the segment profile plot discussed in Step 6.

Figure 7.31 shows the resulting plot. We can specify how we want to label the panels in the plot using the argument `strip.prefix`. In this example, we use the term “Segment” instead of “Cluster”.

```
R> propBarchart(winterActiv, clusters(winter.m5),
+ alpha = 1, strip.prefix = "Segment")
```

As can be seen, the results from the mixture of binary distributions are similar to those from bagged clustering, but not identical. The two largest segments of tourists (in this case segments 1 and 5) either engage in a range of activities including alpine skiing, going for walks, relaxing, shopping and going to the pool/sauna, or are primarily interested in alpine skiing. The health segment of tourists (using spas and health facilities) re-emerges as segment 4. Arriving at market segments with similar profiles when using these two distinctly different techniques, serves as validation of the solution, and gives confidence that these market segments are not entirely random.

7.3.2 Finite Mixtures of Regressions

Finite mixtures of distributions are similar to distance-based clustering methods and – in many cases – result in similar solutions. Compared to hierarchical or partitioning clustering methods, mixture models sometimes produce more useful, and sometimes less useful solutions. Finite mixtures of *regression models* (e.g., Wedel and Kamakura 2000; Bijmolt et al. 2004; Grün and Leisch 2007; Grün and Leisch 2008; Oppewal et al. 2010) offer a completely different type of market segmentation analysis.

Finite mixture of regression models assume the existence of a dependent target variable y that can be explained by a set of independent variables x . The functional relationship between the dependent and independent variables is considered different for different market segments. Figure 7.32 shows a simple artificial data set we will use to illustrate how finite mixtures of regressions work. The command `data("themepark", package = "MSA")` loads the data. The command `plot(pay ~ rides, data = themepark)` plots the data. Figure 7.32 shows the entrance fee consumers are willing to pay for a theme park

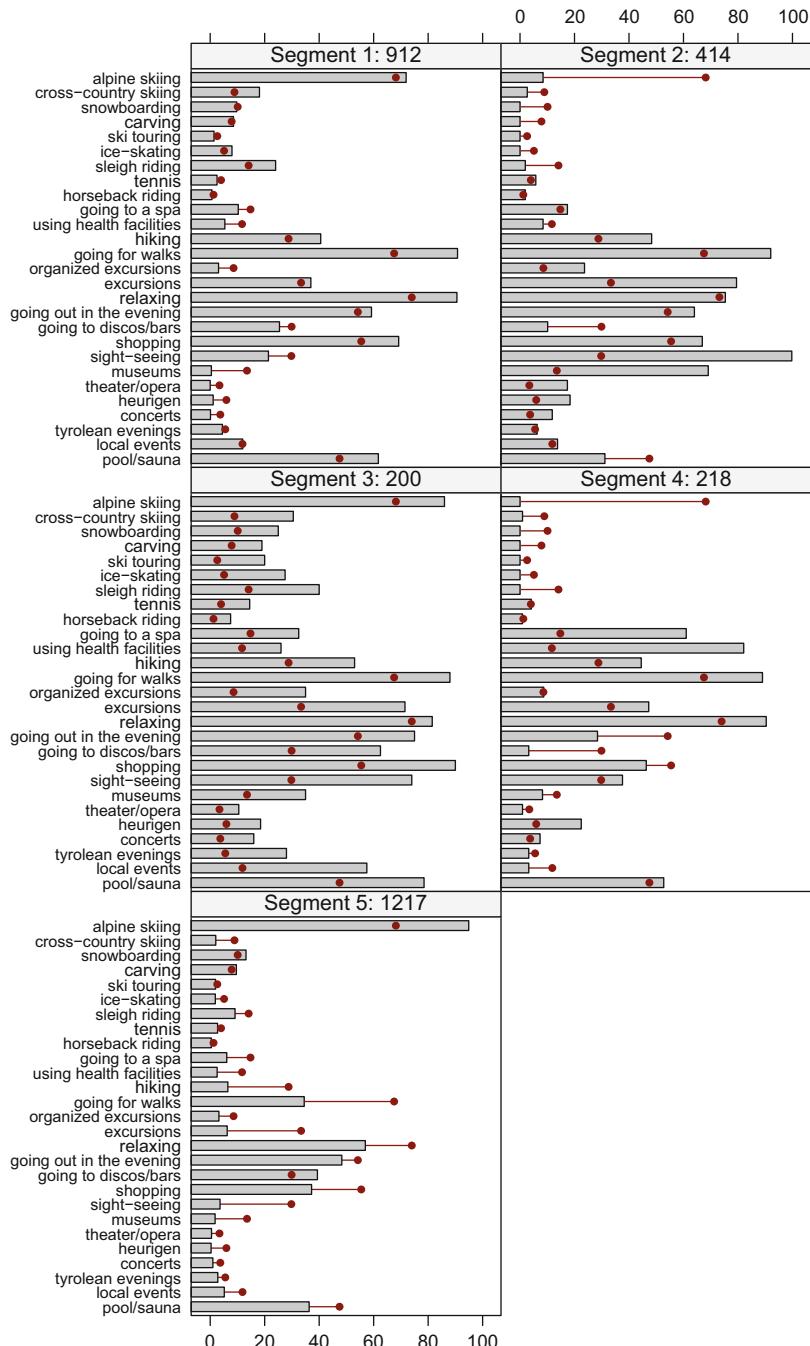
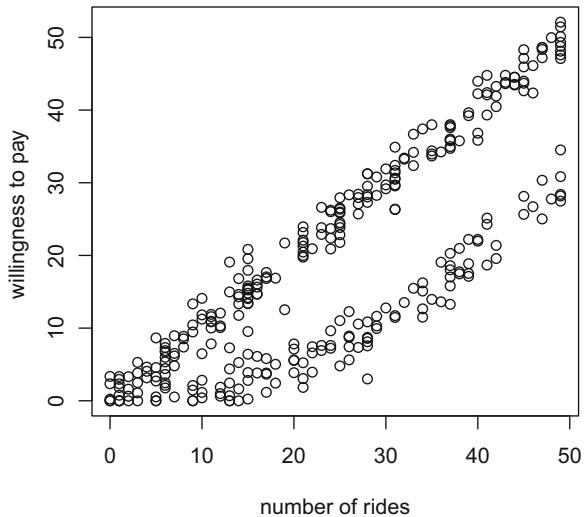


Fig. 7.31 Bar chart of segment-specific probabilities of the mixture of binary distributions fitted to the winter vacation activities data set

Fig. 7.32 Artificial theme park data set



in dependence of the number of rides available in the theme park. As can be seen in Fig. 7.32, two market segments are present in this data: the willingness to pay of the top segment increases linearly with the number of rides available. Members of this segment think that each ride is worth a certain fixed amount of money. The bottom segment does not share this view. Rather, members of this market segment are not willing to pay much money at all until a certain minimum threshold of rides is offered by a theme park. But their willingness to pay increases substantially if a theme park offers a large number of rides. Irrespective of the precise number of rides on offer in the theme park, the willingness to pay of members of the second segment is always lower than the willingness to pay of the first segment.

The artificial data set was generated using the following two linear regression models for the two segments:

$$\begin{aligned} \text{segment 1: } & y = x + \epsilon, \\ \text{segment 2: } & y = 0.0125x^2 + \epsilon, \end{aligned}$$

where x is the number of rides, y is the willingness to pay, and ϵ is normally distributed random noise with standard deviation $\sigma = 2$. In addition, y was ensured to be non-negative.

A linear regression model with the number of rides and the squared number of rides as regressors can be specified with the formula interface in R using:

```
R> pay ~ rides + I(rides^2)
```

Package **flexmix** allows fitting a finite mixture of two linear regression models. Because mixtures of regression models are the default in package **flexmix**, no model needs to be specified. The default `model = FLXMRglm()` is used.

Package **flexmix** allows calculating mixtures of linear regression models, as well as mixtures of generalised linear models (GLM) for logistic or Poisson regression. The following R command executes 10 runs of the EM algorithm with random initialisations. Only the correct number of segments $k = 2$ is used here, but selecting the number of segments using AIC, BIC or ICL works exactly like in the binary data example in Sect. 7.3.1.2.

```
R> library("flexmix")
R> set.seed(1234)
R> park.f1 <- stepFlexmix(pay ~ rides + I(rides^2),
+   data = themepark, k = 2, nrep = 10, verbose = FALSE)
R> park.f1

Call:
stepFlexmix(pay ~ rides + I(rides^2), data = themepark,
k = 2, nrep = 10, verbose = FALSE)

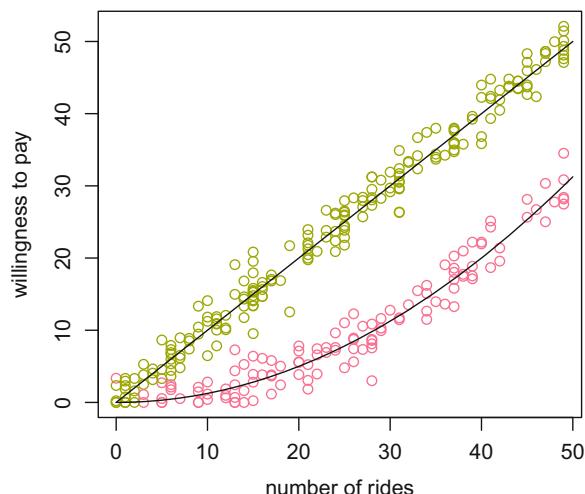
Cluster sizes:
 1   2
119 201

convergence after 20 iterations
```

The model formula `pay ~ rides + I(rides^2)` indicates that the number of rides and the squared number of rides are regressors. The same model formula specification can be used for a standard linear model fitted using function `lm()`. The only difference is that – in this example – two regression models are fitted simultaneously, and consumer (observation) membership to market segments (components) is unknown.

To assess to which market segments the mixture model assigns observations to, observations are plotted in a scatter plot colouring them by segment membership

Fig. 7.33 Scatter plot with observations coloured by segment membership from a two-segment mixture of linear regression models for the artificial theme park data



(see Fig. 7.33). Function `curve()` defines the true regression functions, and adds them to the plot using:

```
R> plot(pay ~ rides, data = themepark, col = clusters(park.f1),
+       xlab = "number of rides", ylab = "willingness to pay")
R> seg1 <- function(x) x
R> seg2 <- function(x) 0.0125 * x^2

R> curve(seg1, from = 0, to = 50, add = TRUE)
R> curve(seg2, from = 0, to = 50, add = TRUE)
```

The parameters estimated by the model are:

```
R> parameters(park.f1)
```

	Comp.1	Comp.2
coef.(Intercept)	1.60901610	0.3171846123
coef.rides	-0.11508969	0.9905130420
coef.I(rides^2)	0.01439438	0.0001851942
sigma	2.06263293	1.9899121188

Each segment has one regression coefficient for the intercept, for the linear term for the number of rides, and for the quadratic term for the number of rides; three estimates in total. The noise standard deviation `sigma` requires one additional estimate.

Fitting mixtures with the EM algorithm is as prone to label switching as any partitioning clustering method. Segment 1 and segment 2 in the description of the data generating process above now re-emerge as segment 2 and segment 1, respectively. This is obvious from the below summary of the fitted regression coefficients:

```
R> summary(refit(park.f1))

$Comp.1
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.6090161  0.6614587  2.4325  0.01499 *
rides      -0.1150897  0.0563449 -2.0426  0.04109 *
I(rides^2)  0.0143943  0.0010734 13.4104 < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.31718461 0.48268972  0.6571   0.5111
rides      0.99051304 0.04256232 23.2721 <2e-16 ***
I(rides^2) 0.00018516 0.00080704  0.2294   0.8185
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use the function `refit()` here because we want to see standard errors for the estimates. The EM algorithm generates point estimates, but does not indicate standard errors (the uncertainty of estimates) because it does not require this

information to obtain the point estimates. `refit()` takes the solution obtained with the EM algorithm, and uses a general purpose optimiser to obtain the uncertainty information.

The summary provides information separately for the two segments (referred to as Comp. 1 and Comp. 2). For each segment, we can see a summary table of the regression coefficients. Each coefficient is shown in one row. Column 1 contains the point estimate, column 2 the standard error, column 3 the test statistic of a z -test with the null hypothesis that the regression coefficient is equal to zero, and column 4 the corresponding p -value for this test. $< 2e-16$ indicates a p -value smaller than $2 \cdot 10^{-16}$. Asterisks indicate if the null hypothesis would be rejected at the significance level of 0.001 (***), 0.01 (**), 0.05 (*), and 0.1 (.).

Looking at the summary table, we see that all regression coefficients should be included in the model for segment 1 (Comp. 1) because the p -values are all smaller than 0.05. For the second market segment (Comp. 2) only the regression coefficient of the linear term (`rides`) needs to be included. This interpretation reflects correctly the nature of the artificial data set, except for label switching (segment 1 is Comp. 2 and segment 2 is Comp. 1).

Example: Australian Travel Motives

We illustrate finite mixtures of regressions using the Australian travel motives data set. We use the metric variables moral obligation score, NEP score, and environmental behaviour on vacation score. We extract these variables from the data set, and remove observations with missing values using:

```
R> data("vacmot", package = "flexclust")
R> envir <- vacmotdesc[, c("Obligation", "NEP",
+   "Vacation.Behaviour")]
R> envir <- na.omit(envir)
R> envir[, c("Obligation", "NEP")] <-
+   scale(envir[, c("Obligation", "NEP")])
```

We standardise the independent variables (moral obligation and NEP score) to have a mean of zero and a variance of one. We do this to improve interpretability and allow visualisation of effects in Fig. 7.34. The environmental behavioural score can be assumed to be influenced by the moral obligation respondents feel, and their attitudes towards the environment as captured by the NEP score.

We fit a single linear regression using:

```
R> envir.lm <- lm(Vacation.Behaviour ~ Obligation + NEP,
+   data = envir)
R> summary(envir.lm)

Call:
lm(formula = Vacation.Behaviour ~ Obligation + NEP,
  data = envir)

Residuals:
```

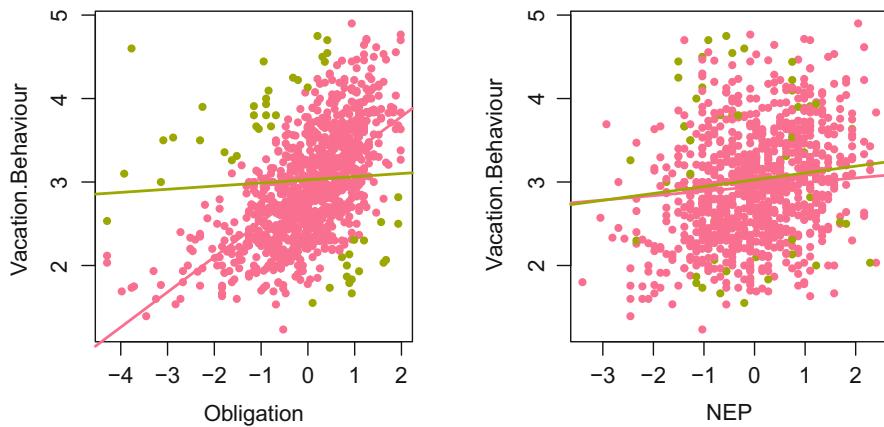


Fig. 7.34 Scatter plot with observations coloured by segment membership together with the segment-specific regression lines from a two-segment mixture of linear regressions fitted to the Australian vacation motives data set

```

      Min       1Q    Median      3Q      Max
-1.60356 -0.36512 -0.04501  0.34991  2.87038
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.96280   0.01821 162.680 < 2e-16 ***
Obligation  0.32357   0.01944  16.640 < 2e-16 ***
NEP          0.06599   0.01944   3.394 0.000718 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5687 on 972 degrees
of freedom
Multiple R-squared:  0.2775, Adjusted R-squared:  0.276
F-statistic: 186.7 on 2 and 972 DF,  p-value: < 2.2e-16

```

Results indicate that an increase in either moral obligation or the NEP score increases the score for environmental behaviour on vacation. But the predictive performance is modest with an R^2 value of 0.28. The R^2 value lies between zero and one, and indicates how much of the variance in the dependent variable is explained by the model; how close the predicted values are to the observed ones.

The association between vacation behaviour score and moral obligation and NEP score can be different for different groups of consumers. A mixture of linear regression models helps us investigate whether this is the case:

```

R> set.seed(1234)
R> envir.m15 <- stepFlexmix(Vacation.Behaviour ~ .,
+     data = envir, k = 1:4, nrep = 10, verbose = FALSE,
+     control = list(iter.max = 1000))

```

We increase the maximum number of iterations for the EM algorithm to 1000 using `control = list(iter.max = 1000)` to ensure convergence of the EM algorithm for all number of segments.

The best model is selected using the BIC:

```
R> envir.m2 <- getModel(envir.m15)
R> envir.m2

Call:
stepFlexmix(Vacation.Behaviour ~ ., data = envir,
control = list(iter.max = 1000), k = 2, nrep = 10,
verbose = FALSE)

Cluster sizes:
 1   2
928 47

convergence after 180 iterations
```

We select a mixture with two segments. The table of segment memberships indicates that the second segment is rather small.

```
R> summary(refit(envir.m2))

$Comp.1
      Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.944634  0.032669 90.1342  < 2e-16 ***
Obligation 0.418934  0.030217 13.8641  < 2e-16 ***
NEP         0.053489  0.027023  1.9794  0.04778 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
      Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.023214  0.139161 21.7246  <2e-16 ***
Obligation 0.018619  0.145845  0.1277  0.8984
NEP         0.082207  0.105744  0.7774  0.4369
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors for the fitted segment-specific parameters indicate that the associations between the dependent and independent variables are stronger for segment 1 than for the complete data set. This means that the predictive performance of the model is better for segment 1 than for the complete data set. For segment 2, neither moral obligation, nor NEP score allow predicting the environmental behaviour on vacation.

Scatter plots visualise the data together with the segmentation solution implied by the fitted model. Data points have different colours to indicate segment memberships. We add the segment-specific regression lines under the assumption that the other covariate has its average value of 0 (see Fig. 7.34):

```
R> par(mfrow = c(1, 2))
R> plot(Vacation.Behaviour ~ Obligation, data = envir,
+   pch = 20, col = clusters(envir.m2))
R> abline(parameters(envir.m2)[1:2, 1], col = 1, lwd = 2)
R> abline(parameters(envir.m2)[1:2, 2], col = 2, lwd = 2)
R> plot(Vacation.Behaviour ~ NEP, data = envir, pch = 20,
+   col = clusters(envir.m2))
R> abline(parameters(envir.m2)[c(1, 3), 1], col = 1, lwd = 2)
R> abline(parameters(envir.m2)[c(1, 3), 2], col = 2, lwd = 2)
```

We see in the left plot in Fig. 7.34 that the regression line for segment 1 (pink) has a steep slope. This means that there is a strong association between vacation behaviour and moral obligation. The regression line for segment 2 (green) is nearly horizontal, indicating no association. The right plot shows the association between vacation behaviour and NEP score. Here, neither of the market segments display a substantial association.

7.3.3 Extensions and Variations

Finite mixture models are more complicated than distance-based methods. The additional complexity makes finite mixture models very flexible. It allows using any statistical model to describe a market segment. As a consequence, finite mixture models can accommodate a wide range of different data characteristics: for metric data we can use mixtures of normal distributions, for binary data we can use mixtures of binary distributions. For nominal variables, we can use mixtures of multinomial distributions or multinomial logit models (see Sect. 9.4.2). For ordinal variables, several models can be used as the basis of mixtures (Agresti 2013). Ordinal variables are tricky because they are susceptible to containing response styles. To address this problem, we can use mixture models disentangling response style effects from content-specific responses while extracting market segments (Grün and Dolnicar 2016). In combination with conjoint analysis, mixture models allow to account for differences in preferences (Frühwirth-Schnatter et al. 2004).

An ongoing conversation in the segmentation literature (e.g. Wedel and Kamakura 2000) is whether differences between consumers should be modelled using a continuous distribution or through modelling distinct, well-separated market segments. An extension to mixture models can reconcile these positions by acknowledging that distinct segments exist, while members of the same segment can still display variation. This extension is referred to as mixture of mixed-effects models or heterogeneity model (Verbeke and Lesaffre 1996). It is used in the marketing and business context to model demand (Allenby et al. 1998).

If the data set contains repeated observations over time, mixture models can cluster the time series, and extract groups of similar consumers (for an overview using discrete data see Frühwirth-Schnatter 2011). Alternatively, segments can be extracted on the basis of switching behaviour of consumers between groups over time using Markov chains. This family of models is also referred to as dynamic

latent change models, and can be used to track changes in brand choice and buying decisions over time. In this case, the different brands correspond to the groups for each time point. Poulsen (1990) uses a finite mixture of Markov chains with two components to track new triers of a continuously available brand (A) over a one year period of time. The two segments differ both in the probability to buy brand A for the first time, and the probability to continue to do so afterwards. Similarly, Bockenholt and Langeheine (1996) model recurrent choices with a latent Markov model. Because several alternative brands are investigated, a multinomial choice model is formulated. Ramaswamy (1997) generalises this for the situation that new brands are introduced to an existing market such that the set of available choices changes over time. The application is on panel survey data for laundry detergents. Brangule-Vlagsma et al. (2002) also use a Markov switching model, but they use it to model changes in customer value systems, which in turn influence buying decisions.

Mixture models also allow to simultaneously include segmentation and descriptor variables. Segmentation variables are used for grouping, and are included in the segment-specific model as usual. Descriptor variables are used to model differences in segment sizes, assuming that segments differ in their composition with respect to the descriptor variables. If, for example, consumers in the segment interested in high-end mobile phones in the artificial mobile phone data set tend to be older and have a higher income, this is equivalent to the segment of consumers interested in high-end mobile phones being larger for older consumers and those with a higher income. The descriptor variables included to model the segment sizes are called *concomitant variables* (Dayton and Macready 1988). In package `flexmix`, concomitant variables can be included using the argument `concomitant`.

7.4 Algorithms with Integrated Variable Selection

Most algorithms focus only on extracting segments from data. These algorithms assume that each of the segmentation variables makes a contribution to determining the segmentation solution. But this is not always the case. Sometimes, segmentation variables were not carefully selected, and contain redundant or noisy variables. Pre-processing methods can identify them. For example, the filtering approach proposed by Steinley and Brusco (2008a) assesses the clusterability of single variables, and only includes variables above a certain threshold as segmentation variables. This approach outperforms a range of alternative variable selection methods (Steinley and Brusco 2008b), but requires metric variables. Variable selection for binary data is more challenging because single variables are not informative for clustering, making it impossible to pre-screen or pre-filter variables one by one.

When the segmentation variables are binary, and redundant or noisy variables can not be identified and removed during data pre-processing in Step 4, suitable segmentation variables need to be identified *during* segment extraction. A number of algorithms extract segments while – simultaneously – selecting suitable segmen-

tation variables. We present two such algorithms for binary segmentation variables: biclustering and the variable selection procedure for clustering binary data (VSBD) proposed by Brusco (2004). At the end of this section, we discuss an approach called factor-cluster analysis. In this two-step approach, segmentation variables are compressed into factors before segment extraction.

7.4.1 Biclustering Algorithms

Biclustering simultaneously clusters both consumers and variables. Biclustering algorithms exist for any kind of data, including metric and binary. This section focuses on the binary case where these algorithms aim at extracting market segments containing consumers who all have a value of 1 for a group of variables. These groups of consumers and variables together then form the *bicluster*.

The concept of biclustering is not new. Hartigan (1972) proposes several patterns for direct clustering of a data matrix. However, possibly due to the lack of available software, uptake of algorithms such as *biclustering*, *co-clustering*, or *two-mode clustering* was minimal. This changed with the advent of modern genetic and proteomic data. Genetic data is characterised by the large numbers of genes, which serve as variables for the grouping task. Humans, for example, have approximately 22,300 genes, which is more than a chicken with 16,700, but less than a grape with 30,400 (Pertea and Salzberg 2010). Traditional clustering algorithms are not useful in this context because many genes have no function, and most cell tasks are controlled by only a very small number of genes. As a consequence, getting rid of noisy variables is critically important. Biclustering experienced a big revival to address these challenges (e.g., Madeira and Oliveira 2004; Prelic et al. 2006; Kasim et al. 2017).

Several popular biclustering algorithms exist; in particular they differ in how a bicluster is defined. In the simplest case, a bicluster is defined for binary data as a set of observations with values of 1 for a subset of variables, see Fig. 7.35. Each row corresponds to a consumer, each column to a segmentation variable (in the example below: vacation activity). The market segmentation task is to identify tourists who all undertake *a subset* of all possible activities. In Fig. 7.35 an A marks a tourist that undertakes a specific vacation activity. An asterisk indicates that a tourist may or may not undertake this specific vacation activity. The challenge is to find large groups of tourists who have as many activities in common as possible.

The biclustering algorithm which extracts these biclusters follows a sequence of steps. The starting point is a data matrix where each row represents one consumer and each column represents a binary segmentation variable:

Step 1 First, rearrange rows (consumers) and columns (segmentation variables) of the data matrix in a way to create a rectangle with identical entries of 1s at the top left of the data matrix. The aim is for this rectangle to be as large as possible.

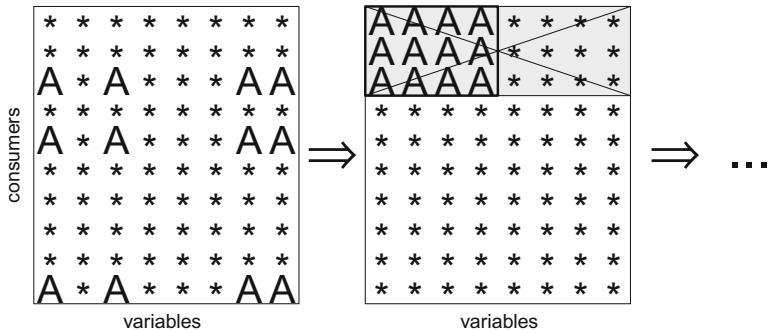


Fig. 7.35 Biclustering with constant pattern

- Step 2 Second, assign the observations (consumers) falling into this rectangle to one bicluster, as illustrated by the grey shading in Fig. 7.35. The segmentation variables defining the rectangle are active variables (A) for this bicluster.
- Step 3 Remove from the data matrix the rows containing the consumers who have been assigned to the first bicluster. Once removed, repeat the procedure from step 1 until no more biclusters of sufficient size can be located.

The algorithm designed to solve this task has control parameters – like minimum number of observations and minimum number of variables – that are necessary to form a bicluster of sufficient size.

This biclustering method has been proposed by Kaiser (2011) referring to it as repeated Bimax algorithm because step 1 can be solved with the Bimax algorithm proposed by Prelic et al. (2006). The Bimax algorithm is computationally very efficient, and allows to identify the largest rectangle corresponding to the global optimum, rather than returning a local optimum as other segment extraction algorithms do. Among the traditional market segmentation approaches, only standard hierarchical clustering implementations determine the globally best merge or split in each step, and therefore generate the same results across repetitions.

Biclustering is not one single very specific algorithm; rather it is a term describing a family of algorithms differing with respect to the properties of data they can accommodate, the extent of similarity between members of market segments required, and whether individual consumers can be assigned to only one or multiple market segments. A comprehensive overview of biclustering algorithms is provided by Madeira and Oliveira (2004), Kaiser and Leisch (2008) and Kaiser (2011). Different algorithms search for different patterns in biclusters. An example of such an alternative pattern – the constant column pattern – is shown in Fig. 7.36. Such a pattern could be used to identify consumers with identical socio-demographics, for example: all female (column with A's), aged 20–29 (column with B's), living in Europe (column with C's), and having a high school degree (column with D's). The same pattern could also be used to create a commonsense/data-driven segmentation where initially large groups of consumers with the same value in several socio-

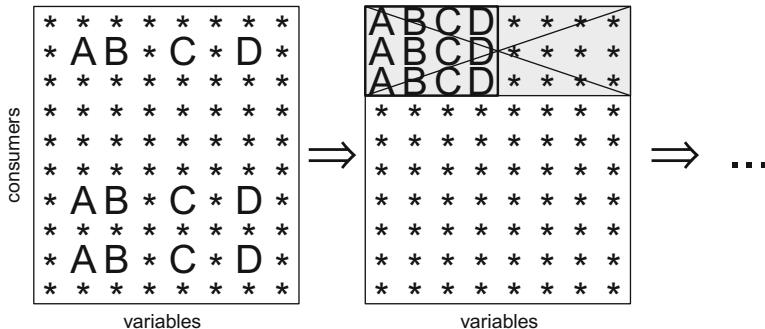


Fig. 7.36 Biclustering with constant column pattern

demographic variables are identified. Then, among those consumers, an interesting subsegment is extracted based on the vacation activity profile.

Biclustering is particularly useful in market segmentation applications with many segmentation variables. Standard market segmentation techniques risk arriving at suboptimal groupings of consumers in such situations. Biclustering also has a number of other advantages:

No data transformation: Typically, situations where the number of variables is too high are addressed by pre-processing data. Pre-processing approaches such as principal components analysis – combined with selecting only the first few components – reduce the number of segmentation variables by transforming the data. Any data transformation changes the information in the segmentation variables, thus risking that segmentation results are biased because they are not based on the original data. Biclustering does not transform data. Instead, original variables which do not display any systematic patterns relevant for grouping consumers are ignored.

Ability to capture niche markets: Because biclustering searches for identical patterns displayed by groups of consumers with respect to groups of variables, it is well suited for identifying niche markets. If a manager is specifically interested in niche markets, the control arguments for the biclustering algorithm should be set such that a high number of matches is required. This approach leads to smaller segments containing members who are very similar to one other. If the matching requirement is relaxed, larger and less homogeneous segments emerge.

Biclustering methods, however, do not group *all* consumers. Rather, they select groups of similar consumers, and leave ungrouped consumers who do not fit into any of the groups.

Example: Australian Vacation Activities

Imagine that a tourist destination wants to identify segments of tourists engaging in similar vacation activities. The data available is similar to that used in the bagged

clustering example, but this time it is binary information for 1003 adult Australian tourists about whether (coded 1) or not (coded 0) they engaged in each of 45 vacation activities during their last domestic vacation. Compared to the Austrian winter vacation activities data set, the number of segmentation variables is nearly twice as high, but the sample size is much smaller; only about one third of the size in the Austrian winter vacation activities data set. This sample size relative to the number of segmentation variables used is insufficient for segment extraction using most algorithms (Dolnicar et al. 2014). A detailed description of the data set is provided in Appendix C.3 and in Dolnicar et al. (2012). The fact that the list of vacation activities is so long complicates market segmentation analysis.

The repeated Bimax algorithm is implemented as method BCrepBimax in the R package `biiclust` (Kaiser and Leisch 2008). The Bimax algorithm is available as method BCBimax. The bicluster solution for this data with method BCrepBimax, a minimum of `minc = 2` activities (columns) and `minr = 50` observations (rows) per cluster can be obtained by:

```
R> library("bioclust")
R> data("ausActiv", package="MSA")
R> ausact.bic <- bioclust(x = ausActiv,
+   method = BCrepBimax,
+   minc = 2, minr = 50, number = 100, maxc = 100)
```

The value of 100 for the maximum number of biclusters (`number`) and maximum number of columns (`maxc`) in each cluster effectively means that no limit is set for both arguments.

We save the result to the hard drive. This allows loading the result from there, and avoiding re-computation when re-using this segmentation solution later.

```
R> save(ausact.bic, file = "ausact-bic.RData")
```

We visualise results using the bicluster membership plot generated by function `bioclustmember()`, see Fig. 7.37:

```
R> bioclustmember(x = ausActiv, bicResult = ausact.bic)
```

Each column in Fig. 7.37 represents one market segment. In total, 12 market segments are identified. Each row represents one of the vacation activities. Cells that are empty indicate that these variables are not useful to characterise this segment as an activity frequently undertaken by segment members. For example: the entire block of variables between THEATRE and SKIING can be ignored in terms of the interpretation of potential market segments because these activities do not characterise any of the market segments. Cells containing two dark outer boxes indicate that members of the segment in that particular row are very similar to one another with respect to their high engagement in that very vacation activity. For example, members of market segment 1 have in common that they like to visit industrial attractions (INDUSTRIAL). Members of segments 3 and 7 have in common that they like to visit museums (MUSEUM). Members of all segments except segments 7 and 12 share their interest in relaxation during their vacations (RELAXING).

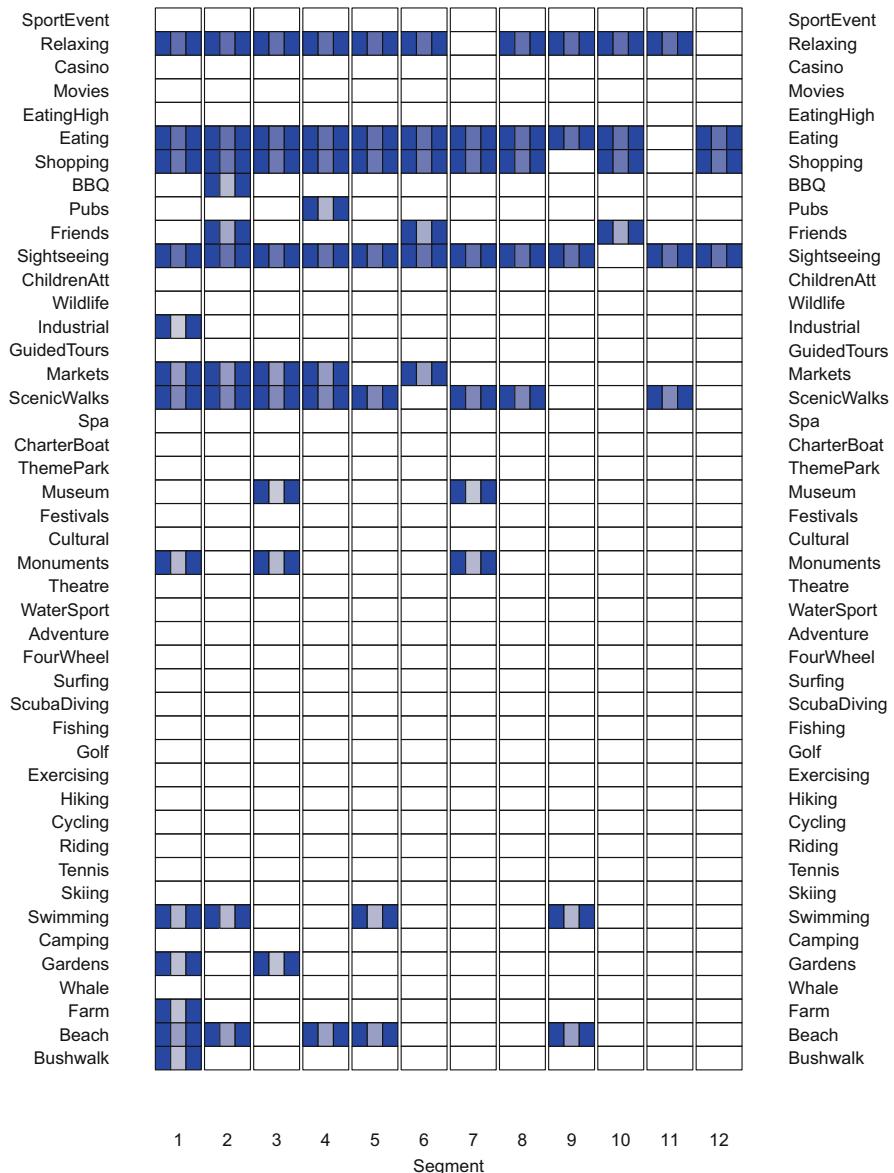


Fig. 7.37 Bicluster membership plot for the Australian vacation activities data set

Finally, the biclustering plot contains one more critical piece of information: how distinctly different members of one market segment are from the average tourist with respect to one specific vacation activity. This information is indicated by the shading of the box in the middle. The lighter that shading, the less does the total sample of

tourists engage in that vacation activity. The stronger the contrast between the two outer boxes and the inner box, the more distinct the market segment with respect to that vacation activity. For example, members of both segments 3 and 7 like to go to museums, but they differ strongly in this activity from the average tourist. Or, looking at segment 2: members of this segment relax, eat in reasonably priced restaurants, shop, go sightseeing, and go to markets, and on scenic walks. None of those vacation activities make them distinctly different from the average tourist. However, members of segment 2 also visit friends, do BBQs, go swimming, and enjoy the beach. These activities are not commonly shared among all tourists, and therefore describe segment 2 specifically.

Note that the segments presented here are slightly different from those reported in Dolnicar et al. (2012). The reason for this deviation is that the algorithm used and the corresponding R functions have been improved since the original analysis. The differences are minor, the variable characteristics for each one of the market segments are nearly identical.

7.4.2 Variable Selection Procedure for Clustering Binary Data (VSBD)

Brusco (2004) proposed a variable selection procedure for clustering binary data sets. His VSBD method is based on the k -means algorithm as clustering method, and assumes that not all variables available are relevant to obtain a good clustering solution. In particular, the method assumes the presence of masking variables. They need to be identified and removed from the set of segmentation variables. Removing irrelevant variables helps to identify the correct segment structure, and eases interpretation.

The procedure first identifies the best small subset of variables to extract segments. Because the procedure is based on the k -means algorithm, the performance criterion used to assess a specific subset of variables is the within-cluster sum-of-squares (the sum of squared Euclidean distances between each observation and their segment representative). This is the criterion minimised by the k -means algorithm. After having identified this subset, the procedure adds additional variables one by one. The variable added is the one leading to the smallest increase in the within-cluster sum-of-squares criterion. The procedure stops when the increase in within-cluster sum-of-squares reaches a threshold. The number of segments k has to be specified in advance. Brusco (2004) recommends calculating the Ratkowsky and Lance index (Ratkowsky and Lance 1978, see also Sect. 7.5.1) for the complete data with all variables to select the number of segments.

The algorithm works as follows:

Step 1 Select only a subset of observations with size $\phi \in (0, 1]$ times the size of the original data set. Brusco (2004) suggests to use $\phi = 1$ if the original data set contains less than 500 observations, $0.2 \leq \phi \leq 0.3$ if the number

of observations is between 500 and 2000 and $\phi = 0.1$ if the number of observations is at least 2000.

- Step 2 For a given number of variables V , perform an exhaustive search for the set of V variables that leads to the smallest within-cluster sum-of-squares criterion. The value for V needs to be selected small for the exhaustive search to be computationally feasible. Brusco (2004) suggests using $V = 4$, but smaller or larger values may be required depending on the number of clusters k , and the number of variables p . The higher the number of clusters, the larger V should be to capture the more complex clustering structure. The higher p , the smaller V needs to be to make the exhaustive search computationally feasible.
- Step 3 Among the remaining variables, determine the variable leading to the smallest increase in the within-cluster sum-of-squares value if added to the set of segmentation variables.
- Step 4 Add this variable if the increase in within-cluster sum-of-squares is smaller than the threshold. The threshold is δ times the number of observations in the subset divided by 4. δ needs to be in $[0, 1]$. Brusco (2004) suggests a default δ value of 0.5.

Brusco (2004) suggests 500 random initialisations in step 2, and 5000 random initialisations in step 3 for each run of the k -means algorithm. This recommendation is based on the use of the Forgy/Lloyd algorithm (Forgy 1965; Lloyd 1982). Using the more efficient Hartigan-Wong algorithm (Hartigan and Wong 1979) allows us to reduce the number of random initialisations. In the example below we use 50 random initialisations in step 2, and 100 random initialisations in step 3. The Hartigan-Wong algorithm is used by default by function `kmeans` in R.

Example: Australian Travel Motives

We illustrate the VSBD algorithm using the Australian travel motives data set:

```
R> data("vacmot", package = "flexclust")
```

We apply the algorithm to the complete data set when clustering the data set into 6 groups (`centers = 6`). The default settings with $\phi = 1$ (`phi`) and $V = 4$ (`initial.variables`) are used together with `nstart1 = 50`, the number of random initialisations in step 2, and `nstart2 = 100`, the number of random initialisations in step 3. The maximum number of variables (`max.variables`) is the number of available variables (default), and the stopping criterion is set to $\delta = 0.5$ (`delta = 0.5`).

```
R> set.seed(1234)
R> library("MSA")
R> vacmot.sv <- vsbd(vacmot, centers = 6, delta = 0.5)
```

Executing the command can take some time because the algorithm is computationally expensive due to the exhaustive search of the best subset of four variables.

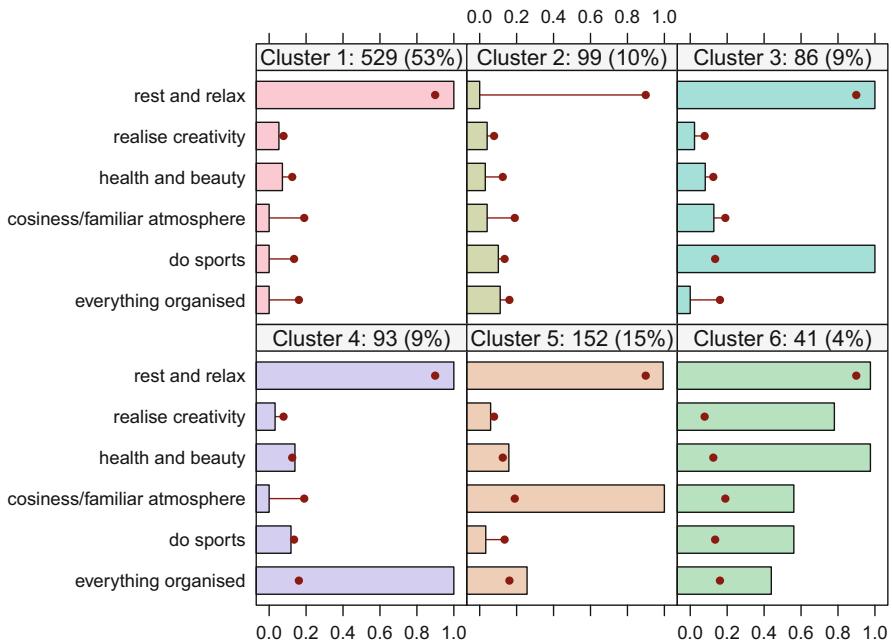


Fig. 7.38 Bar chart of cluster means obtained for the Australian travel motives data set after selecting variables with the VSBD algorithm

The VSBD procedure selects the following variables:

```
R> colnames(vacmot) [vacmot.sv]
[1] "rest and relax"
[2] "realise creativity"
[3] "health and beauty"
[4] "cosiness/familiar atmosphere"
[5] "do sports"
[6] "everything organised"
```

The original data set contained 20 variables. The VSBD algorithm selected only 6 variables. Using these variables, the final solution – together with the plot in Fig. 7.38 – results from:

```
R> library("flexclust")
R> vacmot.vsbd <- stepcclust(vacmot[, vacmot.sv], k = 6,
+ nrep = 10)
R> barchart(vacmot.vsbd)
```

The segmentation solution contains segments caring or not caring about rest and relaxation; the percentage of agreement with this motive within segments is either close to 100% or to 0% (segment 2). In addition, respondents in segment 3 agree that doing sports is a motive for them, while members of segment 4 want

everything organised. For members of segment 5 cosiness and a familiar atmosphere are important. To members of segment 6 the largest number of motives applies; they are the only ones caring about creativity and health and beauty. This result indicates that using the variable selection procedure generates a solution that is easy to interpret because only a small set of variables serve as segmentation variables, but each of them differentiates well between segments.

7.4.3 **Variable Reduction: Factor-Cluster Analysis**

The term *factor-cluster analysis* refers to a two-step procedure of data-driven market segmentation analysis. In the first step, segmentation variables are factor analysed. The raw data, the original segmentation variables, are then discarded. In the second step, the factor scores resulting from the factor analysis are used to extract market segments.

Sometimes this approach is conceptually legitimate. For example, if the empirical data results from a validated psychological test battery designed specifically to contain a number of variables which load onto factors, like IQ tests. In IQ tests, a number of items assess the general knowledge of a person. In this case a conceptual argument can be put forward that it is indeed legitimate to replace the original variables with the factor score for general knowledge. However, the factor scores should either be determined simultaneously when extracting the groups (for example using a model-based approach based on factor analyzers; McLachlan et al. 2003) or be provided separately and not determined in a data-driven way from the data where the presence of groups is suspected.

Validated psychological test batteries rarely serve as segmentation variables. More common is the case where factor-cluster analysis is used because the original number of segmentation variables is too high. According to the results from simulation studies by Dolnicar et al. (2014, 2016), a rule of thumb is that the number of consumers in a data set (sample size) should be at least 100 times the number of segmentation variables. This is not always easy to achieve, given that two thirds of applied market segmentation studies reviewed in Dolnicar (2002b) use between 10 and 22 variables. For 22 segmentation variables, the sample size should be at least 2200. Yet, most consumer data sets underlying the market segmentation analyses investigated in Dolnicar (2002a) contain fewer than 1000 consumers.

Running factor-cluster analysis to deal with the problem of having too many segmentation variables in view of their sample size lacks conceptual legitimisation and comes at a substantial cost:

Factor analysing data leads to a substantial loss of information. To illustrate this, we factor analyse all the segmentation variables used in this book, and report the number of extracted factors and the percentage of explained variance. We apply principal components analysis to the correlation matrix, and retain principal components with eigenvalues larger than 1, using the so-called Kaiser

criterion (Kaiser 1960). The reasoning for the Kaiser criterion is to keep only principal components that represent more information content than an average original variable.

The risk aversion data set (see Appendix C.1) contains six variables. When factor analysed, 1 factor is extracted, explaining 47% of the variability in the data. When using factor scores for segment extraction, 53% of the information is lost before segment extraction.

The Austrian winter vacation activities data set (see Appendix C.2) contains 27 variables. When factor analysed, 9 factors are extracted, explaining 51% of the variability in the data. If factor-cluster analysis is used, 49% of the information contained in the segmentation variables is lost before segment extraction.

The Australian vacation activities data set (see Appendix C.3) contains 45 variables. When factor analysed, 8 factors are extracted, explaining 50% of the variability in the data. In this case, half of the information contained in the raw data is sacrificed when segments are extracted using factor-cluster analysis.

Finally, the Australian travel motives data set (see Appendix C.4) contains 20 variables. When factor analysed, 7 factors are extracted, explaining 54% of the variability in the data. This means that discarding the original segmentation variables, and extracting segments on the basis of factor scores instead uses only 54% of the information collected from consumers.

Factor analysis transforms data. As a consequence of using a subset of resulting factors only, segments are extracted from a modified version of the consumer data, not the consumer data itself. Arabie and Hubert (1994) argue that factor-cluster analysis is an outmoded and statistically insupportable practice because data is transformed and, as a consequence, the nature of the data is changed before segment extraction. Similarly, Milligan (1996) concludes from experimental studies that market segments (clusters) derived from the factor score space do not represent market segments (clusters) derived from the raw segmentation variables well. Milligan recommends extracting segments from the space in which segments are postulated to exist. Typically this is the space of the original consumer data, the original segmentation variables.

Factors-cluster results are more difficult to interpret. Instead of obtaining the results for the original segmentation variables which directly reflect information about consumers contained in the data set, factor-cluster results need to be interpreted in factor space. Segment profiling using segment profile plots is easy when original consumer information is used. This is not the case when factors are the basis of profiling. Factors contain partial information from a range of variables, and therefore have no concrete meaning, making the translation process from segments to practical recommendations for the marketing mix when targeting a certain segment very difficult. Imagine, for example, a factor which represents the segmentation variables RELAXING AT THE BEACH, WINDSURFING, KITESURFING, and JETSKIING. If this factor is important for a market segment, should more jetski rentals be opened? Or should they not in order to facilitate relaxation at the beach?

An excellent conclusion of the above issues is offered by Sheppard (1996, p. 57): Cluster analysis on raw item scores, as opposed to factor scores, may produce more accurate or detailed segmentation as it preserves a greater degree of the original data. Sheppard (1996) discourages the use of factor-cluster analysis for market segmentation purposes, suggesting instead that the method may be useful for the purpose of developing an instrument for the entire population where homogeneity (not heterogeneity) among consumers is assumed.

In addition to the conceptual problems outlined above, empirical evidence suggests that factor-cluster analysis does not outperform cluster analysis using raw data. Using a series of artificial data sets of known structure, Dolnicar and Grün (2008) show that – even in cases where the artificial data was generated following a factor-analytic model, thus giving factor analysis an unfair advantage – factor-cluster analysis failed to outperform clustering of raw data in terms of identifying the correct market segment structure contained in the data.

7.5 Data Structure Analysis

Extracting market segments is inherently exploratory, irrespective of the extraction algorithm used. Validation in the traditional sense, where a clear optimality criterion is targeted, is therefore not possible. Ideally, validation would mean calculating different segmentation solutions, choosing different segments, targeting them, and then comparing which leads to the most profit, or most success in mission achievement. This is clearly not possible in reality because one organisation cannot run multiple segmentation strategies simultaneously just for the sake of determining which performs best.

As a consequence, the term validation in the context of market segmentation is typically used in the sense of assessing reliability or stability of solutions across repeated calculations (Choffrey and Lilien 1980; Doyle and Saunders 1985) after slightly modifying the data (Funkhouser 1983; Jurowski and Reich 2000; Calantone and Sawyer 1978; Hoek et al. 1996), or the algorithm (Esslemont and Ward 1989; Hoek et al. 1996). This approach is fundamentally different from validation using an external validation criterion. Throughout this book, we refer to this approach as stability-based data structure analysis.

Data structure analysis provides valuable insights into the properties of the data. These insights guide subsequent methodological decisions. Most importantly, stability-based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be revealed easily. If they do not, users and data analysts need to explore a large number of alternative solutions to identify the most useful segment(s) for the organisation.

If there is structure in the data, be it cluster structure or structure of a different kind, data structure analysis can also help to choose a suitable number of segments to extract.

We discuss four different approaches to *data structure analysis*: cluster indices, gorge plots, global stability analysis, and segment level stability analysis.

7.5.1 Cluster Indices

Because market segmentation analysis is exploratory, data analysts need guidance to make some of the most critical decisions, such as selecting the number of market segments to extract. So-called *cluster indices* represent the most common approach to obtaining such guidance. Cluster indices provide insight into particular aspects of the market segmentation solution. Which kind of insight, depends on the nature of the cluster index used. Generally, two groups of cluster indices are distinguished: internal cluster indices and external cluster indices.

Internal cluster indices are calculated on the basis of one single market segmentation solution, and use information contained in this segmentation solution to offer guidance. An example for an internal cluster index is the sum of all distances between pairs of segment members. The lower this number, the more similar members of the same segment are. Segments containing similar members are attractive to users.

External cluster indices cannot be computed on the basis of one single market segmentation solution only. Rather, they require another segmentation as additional input. The external cluster index measures the similarity between two segmentation solutions. If the correct market segmentation is known, the correct assignment of members to segments serves as the additional input. The correct segment memberships, however, are only known when artificially generated data is being segmented. When working with consumer data, there is no such thing as a correct assignment of members to segments. In such cases, the market segmentation analysis can be repeated, and the solution resulting from the second calculation can be used as additional input for calculating the external cluster index. A good outcome is if repeated calculations lead to similar market segments because this indicates that market segments are extracted in a stable way. The most commonly used measures of similarity of two market segmentation solutions are the Jaccard index, the Rand index and the adjusted Rand index. They are discussed in detail below.

7.5.1.1 Internal Cluster Indices

Internal cluster indices use a single segmentation solution as a starting point. Solutions could result from hierarchical, partitioning or model-based clustering methods. Internal cluster indices ask one of two questions or consider their combination: (1) how compact is each of the market segments? and (2) how well-separated are different market segments? To answer these questions, the notion of a distance measure between observations or groups of observations is required. In

addition, many of the internal cluster indices also require a segment representative or centroid as well as a representative for the complete data set.

A very simple internal cluster index measuring compactness of clusters results from calculating the sum of distances between each segment member and their segment representative. Then the sum of within-cluster distances W_k for a segmentation solution with k segments is calculated using the following formula where we denote the set of observations assigned to segment number h by S_h and their segment representative by \mathbf{c}_h :

$$W_k = \sum_{h=1}^k \sum_{\mathbf{x} \in S_h} d(\mathbf{x}, \mathbf{c}_h).$$

In the case of the k -means algorithm, the sum of within-cluster distances W_k decreases monotonically with increasing numbers of segments k extracted from the data (if the global optimum for each number of segments is found; if the algorithm is stuck in a local optimum, this may not be the case).

A simple graph commonly used to select the number of market segments for k -means clustering based on this internal cluster index is the scree plot. The scree plot visualises the sum of within-cluster distances W_k for segmentation solutions containing different numbers of segments k . Ideally, an *elbow* appears in the scree plot. An elbow results if there is a point (number of segments) in the plot where the differences in sum of within-cluster distances W_k show large decreases before this point and only small decreases after this point. The scree plot for the artificial mobile phone data set (first introduced in Sect. 7.2.3.1 and visualised in Fig. 7.9) is given in Fig. 7.12. This data set contains three distinct market segments. In the scree plot a distinct elbow is visible because the within-cluster distances have distinct drops up to three segments and only small decreases after this point, thus correctly guiding the data analyst towards extracting three market segments. In consumer data, elbows are not so easy to find in scree plots, as can be seen in Fig. 7.13 for the tourist risk taking data set, and in Fig. A.2 for the fast food data set. In both these scree plots the sum of within-cluster distances W_k slowly drops as the number of segments increases. No distinct elbow offers guidance to the data analyst. This is not an unusual situation when working with consumer data.

A slight variation of the internal cluster index of the sum of within-cluster distances W_k is the Ball-Hall index W_k/k . This index was proposed by Ball and Hall (1965) with the aim of correcting for the monotonous decrease of the internal cluster index with increasing numbers of market segments. The Ball-Hall index W_k/k achieves this by dividing the sum of within-cluster distances W_k by the number of segments k .

The internal cluster indices discussed so far focus on assessing the aspect of similarity (or homogeneity) of consumers who are members of the same segment, and thus the compactness of the segments. Dissimilarity is equally interesting. An optimal market segmentation solution contains market segments that are very different from one another, and contain very similar consumers. This idea is

mathematically captured by another internal cluster index based on the weighted distances between centroids (cluster centres, segment representative) B_k :

$$B_k = \sum_{h=1}^k n_h d(\mathbf{c}_h, \bar{\mathbf{c}})$$

where $n_h = |\mathcal{S}_h|$ is the number of consumers in segment \mathcal{S}_h , and $\bar{\mathbf{c}}$ is the centroid of the entire consumer data set (when squared Euclidean distance is used this centroid is equivalent to the mean value across all consumers; when Manhattan distance is used it is equivalent to the median).

A combination of the two aspects of compactness and separation is mathematically captured by other internal cluster indices which relate the sum of within-cluster distances W_k to the weighted distances between centroids B_k . If natural market segments exist in the data, W_k should be small and B_k should be large. Relating these two values can be very insightful in terms of guiding the data analyst to choose a suitable number of segments. W_k and B_k can be combined in different ways. Each of these alternative approaches represents a different internal cluster index.

The Ratkowsky and Lance index (Ratkowsky and Lance 1978) is recommended by Brusco (2004) for use with the VSBD procedure for variable selection (see Sect. 7.4.2). The Ratkowsky and Lance index is based on the squared Euclidean distance, and uses the average value of the observations within a segment as centroid. The index is calculated by first determining, for each variable, the sum of squares between the segments divided by the total sum of squares for this variable. These ratios are then averaged, and divided by the square root of the number of segments. The number of segments with the maximum Ratkowsky and Lance index value is selected.

Many other internal cluster indices have been proposed in the literature since Ball and Hall (1965). The seminal paper by Milligan and Cooper (1985) compares a large number of indices in a series of simulation experiments using artificial data. The best performing index in the simulation study by Milligan and Cooper (1985) is the one proposed by Calinski and Harabasz (1974):

$$CH_k = \frac{B_k/(k-1)}{W_k/(n-k)},$$

where n is equal to the number of consumers in the data set. The recommended number of segments has the highest value of CH_k .

Many internal cluster indices are available in R. Function `cluster.stats()` in package `fpc` (Hennig 2015) automatically returns a set of internal cluster indices. Package `clusterSim` (Walesiak and Dudek 2016) allows to request individual internal cluster indices. A very comprehensive list of 30 internal indices is available in package `NbClust` (Charrad et al. 2014). For objects returned by functions in package `flexclust`, the Calinski-Harabasz index can be computed using function `chIndex()`.

Calculating internal cluster indices is valuable as it comes at no cost to the data analyst, yet may reveal interesting aspects of market segmentation solutions. It is possible, however, given that consumer data typically do not contain natural market segments, that internal cluster indices fail to provide much guidance to the data analyst on the best number of segments to extract. In such situations, external cluster indices and global and segment-specific stability analysis are particularly useful.

7.5.1.2 External Cluster Indices

External cluster indices evaluate a market segmentation solution using additional external information; they cannot be calculated using only the information contained in one market segmentation solution. A range of different additional pieces of information can be used. The true segment structure – if known – is the most valuable additional piece of information. But the true segment structure of the data is typically only known for artificially generated data. The true segment structure of consumer data is never known. When working with consumer data, the market segmentation solution obtained using a repeated calculation can be used as additional, external information. The repeated calculation could use a different clustering algorithm on the same data; or it could apply the same algorithm to a variation of the original data, as discussed in detail in Sect. 7.5.3.

A problem when comparing two segmentation solutions is that the labels of the segments are arbitrary. This problem of invariance of solutions when labels are permuted is referred to as *label switching* (Redner and Walker 1984). One way around the problem of label switching is to focus on whether pairs of consumers are assigned to the same segments repeatedly (irrespective of segment labels), rather than focusing on the segments individual consumers are assigned to. Selecting any two consumers, the following four situations can occur when comparing two market segmentation solutions \mathcal{P}_1 and \mathcal{P}_2 :

- *a*: Both consumers are assigned to the same segment twice.
- *b*: The two consumers are in the same segment in \mathcal{P}_1 , but not in \mathcal{P}_2 .
- *c*: The two consumers are in the same segment in \mathcal{P}_2 , but not in \mathcal{P}_1 .
- *d*: The two consumers are assigned to different market segments twice.

To differentiate those four cases, it is not necessary to know the segment labels. These cases are invariant to specific labels assigned to segments. Across the entire data set containing n consumers, $n(n - 1)/2$ pairs of consumers can be selected. Let a , b , c and d represent the number of pairs where each of the four situations outlined above applies. Thus $a + b + c + d = n(n - 1)/2$. If the two segmentation solutions are very similar, a and d will be large and b and c will be small. The index proposed by Jaccard (1912) is based on this observation, but uses only a , b and c while dropping d :

$$J = \frac{a}{a + b + c}.$$

Jaccard did not propose this index for market segmentation analysis. Rather, he was interested in comparing similarities of certain alpine regions in relation to plant species found. But the mathematical problem is the same. The Jaccard index takes values in $[0, 1]$. A value of $J = 0$ indicates that the two market segmentation solutions are completely different. A value of $J = 1$ means that the two market segmentation solutions are identical.

Rand (1971) proposed a similar index based on all four values a, b, c and d :

$$R = \frac{a + d}{a + b + c + d}.$$

The Rand index also takes values in $[0, 1]$; the index values have the same interpretation as those for the Jaccard index, but the Rand index includes d .

Both the Jaccard index and the Rand index share the problem that the absolute values (ranging between 0 and 1) are difficult to interpret because minimum values depend on the size of the market segments contained in the solution. If, for example, one market segmentation solution contains two segments: segment 1 with 80% of the data, and segment 2 with 20% of the data. And a second market segmentation solution also results in an 80:20 split, but half of the members of the small segment were members of the large segment in the first segmentation solution, one would expect a similarity measure of these two segmentation solutions to indicate low values. But because – in each of the two solutions – the large segment contains so many consumers, 60% of them will still be allocated to the same large segment, leading to high Rand and Jaccard index values. Because – in this case – at least 60% of the data are in the large segment for both segmentation solutions, neither the value for the Jaccard index, nor the value for the Rand index can ever be 0.

The values of both indices under random assignment to segments with their size fixed depend on the sizes of the extracted market segments. To solve this problem, Hubert and Arabie (1985) propose a general correction for agreement by chance given segment sizes. This correction can be applied to any external cluster index. The expected index value assuming independence is the value the index takes on average when segment sizes are fixed, but segment membership is assigned to the observations completely at random to obtain each of the two segmentation solutions. The proposed correction has the form

$$\frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}$$

such that a value of 0 indicates the level of agreement expected by chance given the segment sizes, while a value of 1 indicates total agreement. The result of applying the general correction proposed by Hubert and Arabie (1985) to the Rand index is the so-called *adjusted Rand index*.

In R, function `comPart()` from package `flexclust` computes the Jaccard index, the Rand index and the adjusted Rand index. The adjusted Rand index is critically important to the resampling-based data structure analysis approach recommended in Sects. 7.5.3 and 7.5.4.

7.5.2 Gorge Plots

A simple method to assess how well segments are separated, is to look at the distances of each consumer to all segment representatives. Let d_{ih} be the distance between consumer i and segment representative (centroid, cluster centre) h . Then

$$s_{ih} = \frac{e^{-d_{ih}^\gamma}}{\sum_{l=1}^k e^{-d_{il}^\gamma}}$$

can be interpreted as the similarity of consumer i to the representative of segment h , with hyper parameter γ controlling how differences in distance translate into differences in similarity. These similarities are between 0 and 1, and sum to 1 for each consumer i over all segment representatives h , $h = 1, \dots, k$.

For partitioning methods, segment representatives and distances between consumers and segment representatives are directly available. For model-based methods, we use the probability of a consumer i being in segment h given the consumer data, and the fitted mixture model to assess similarities. In the mixture of normal distributions case, these probabilities are close to the similarities obtained with Euclidean distance and $\gamma = 2$ for k -means clustering. Below we use $\gamma = 1$ because it shows more details, and led to better results in simulations on artificial data. The parameter can be specified by the user in the R implementation.

Similarity values can be visualised using *gorge plots*, *silhouette plots* (Rousseeuw 1987), or *shadow plots* (Leisch 2010). We illustrate the use of gorge plots using the three artificial data sets introduced in Table 2.3. The plots in the middle column of Fig. 7.39 show the gorge plots for the three-segment solutions extracted using k -means partitioning clustering for these data sets. Each gorge plot contains histograms of the similarity values s_{ih} separately for each segment. The x -axis plots similarity values. The y -axis plots the frequency with which each similarity value occurs. If the similarity values are the result of distance-based segment extraction methods, high similarity values indicate that a consumer is very close to the centroid (the segment representative) of the market segment. Low similarity values indicate that the consumer is far away from the centroid. If the similarity values are the result of model-based segment extraction methods, high similarity values indicate that a consumer has a high probability of being a member of the market segment. Low similarity values indicate low probability of segment membership.

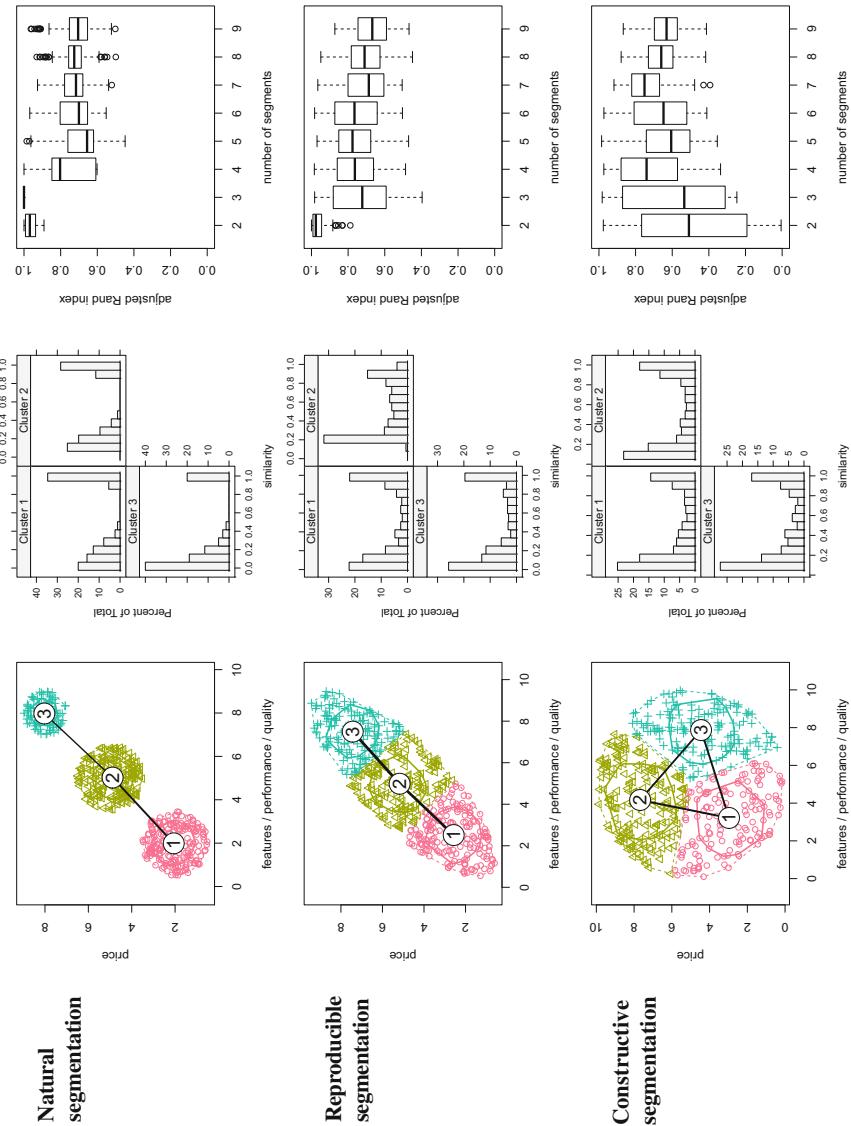


Fig. 7.39 Segment separation plots (*left*), and gorgé plots (*middle*) for the three-segment solutions, and global stability boxplots (*right*) for 2–9 segments for the three artificial data sets.

If natural, well-separated market segments are present in the data, we expect the gorge plot to contain many very low and many very high values. This is why this plot is referred to as gorge plot. Optimally, it takes the shape of a gorge with a peak to the left and a peak to the right.

Figure 7.39 shows prototypical gorge plots for the three-segment solutions extracted from the data sets used to illustrate the three concepts of market segmentation (see also Table 2.3): natural (top row of Fig. 7.39), reproducible (middle row) and constructive segmentation (bottom row). Looking at the natural clustering case with three clearly separated segments: the gorge plot shows a close to perfect gorge, pointing to the fact that most consumers are either close to their segment representative or far away from the representatives of other market segments. The gorge is much less distinct for the reproducible and the constructive clustering cases where many consumers sit in the middle of the plot, indicating that they are neither very close to their segment representative, nor very far away from the segment representatives of other clusters.

Figure 7.39 only contains gorge plots for the three-segment solutions. For a real market segmentation analysis, gorge plots have to be generated and inspected for every number of segments. Producing and inspecting a large number of gorge plots is a tedious process, and has the disadvantage of not accounting for randomness in the sample used. These disadvantages are overcome by stability analysis, which can be conducted at the global or segment level.

7.5.3 Global Stability Analysis

An alternative approach to data structure analysis that can be used for both distance- and model-based segment extraction techniques is based on resampling methods. Resampling methods offer insight into the stability of a market segmentation solution across repeated calculations. To assess the global stability of any given segmentation solution, several new data sets are generated using resampling methods, and a number of segmentation solutions are extracted.

Then the stability of the segmentation solutions across repeated calculations is compared. The solution which can best be replicated is chosen. One such resampling approach is described in detail in this section. Others have been proposed by Breckenridge (1989), Dudoit and Fridlyand (2002), Grün and Leisch (2004), Lange et al. (2004), Tibshirani and Walther (2005), Gana Dresen et al. (2008), and Maitra et al. (2012).

To understand the value of resampling methods for market segmentation analysis, it is critical to accept that consumer data rarely contain distinct, well-separated market segments like those in the artificial mobile phone data set. In the worst case, consumer data can be totally unstructured. Unfortunately, the structure of any given empirical data set is not known in advance.

Resampling methods – combined with many repeated calculations using the same or different algorithms – provide critical insight into the structure of the data. It is helpful, before using resampling methods, to develop a systematics of

data structures that might be discovered, and discuss the implications of those data structures on the way market segmentation analysis is conducted.

Conceptually, consumer data can fall into one of three categories: rarely, naturally existing, distinct, and well-separated market segments exist. If *natural segments* exist in the data, these are easy to identify with most extraction methods. The resulting segments can safely be used by the organisation as the basis of long-term strategic planning, and the development of a customised marketing mix.

A second possibility is that data is entirely unstructured, making it impossible to reproduce any market segmentation solution across repeated calculations. In this worst case scenario, the data analyst must inform the user of the segmentation solution of this fact because it has major implications on how segments are extracted. If data is truly unstructured, and an organisation wishes to pursue a market segmentation strategy, managerially useful market segments have to be constructed. If the segmentation is *constructive*, the role of the data analyst is to offer potentially interesting segmentation solutions to the user, and assist them in determining which of the artificially created segments is most useful to them.

Of course, there is always a middle option between the worst case and the best case scenario. Consumer data can lack distinct, well-separated natural clusters, while not being entirely unstructured. In this case, the existing structure can be leveraged to extract artificially created segments that re-emerge across repeated calculations. This case is referred to as *reproducible segmentation*.

Global stability analysis helps determine which of the concepts applies to any given data set (Dolnicar and Leisch 2010). Global stability analysis acknowledges that both the sample of consumers, and the algorithm used in data-driven segmentation introduce randomness into the analysis. Therefore, conducting one single computation to extract market segments generates nothing more than one of many possible solutions.

The problem of sample randomness has been discussed in early work on market segmentation. Haley (1985), who is credited as being the father of benefit segmentation, recommends addressing the problem by dividing the sample of respondents into subsamples, and extracting market segments independently for each of the subsamples. Then, segmentation variables are correlated across segments from different solutions to identify reproducible segments. Haley (1985) notes that this approach is also useful in informing the decision how many segments to extract from the data, although he acknowledges that the final choice as to the number of segments rests heavily on the judgement of the researchers making the decision (p. 224).

The increase in computational power since Haley's recommendation makes available more efficient new approaches to achieve the same aim. Dolnicar and Leisch (2010) recommend using bootstrapping techniques. Bootstrapping generates a number of new data sets by drawing observations with replacement from the original data. These new data sets can then be used to compute replicate segmentation solutions for different numbers of segments. Computing the similarity between the resulting solutions for the same number of clusters provides insight into whether natural segments exist in the data (in which case all replications will lead to

essentially the same solution), whether reproducible segments exist (in which case similar segments will emerge, indicating that there is some data structure, but no cluster structure), or whether segments are being constructed artificially (in which case replications of segment extraction will lead to different results every time).

In addition, the results from global stability analysis assist in determining the most suitable number of segments to extract from the data. Numbers of segments that allow the segmentation solution in its entirety to be reproduced in a stable manner across repeated calculations are more attractive than numbers of segments leading to different segmentation solutions across replications.

Dolnicar and Leisch (2010) recommend the following steps:

1. Draw b pairs of bootstrap samples ($2b$ bootstrap samples in total) from the sample of consumers, including as many cases as there are consumers in the original data set ($b = 100$ bootstrap sample pairs works well).
2. For each of the $2b$ bootstrap samples, extract $2, 3, \dots, k$ market segments using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model). The maximum number of segments k needs to be specified.
3. For each pair of bootstrap samples b and number of segments k , compute the adjusted Rand index (Hubert and Arabie 1985) or another external cluster index (see Sect. 7.5.1) to evaluate how similar the two segmentation solutions are. This results in b adjusted Rand indices (or other external cluster index values) for each number of segments.
4. Create and inspect boxplots to assess the global reproducibility of the segmentation solutions. For the adjusted Rand index, many replications close to 1 indicate the existence of reproducible clusters, while many replications close to 0 indicate the artificial construction of clusters.
5. Select a segmentation solution, and describe resulting segments. Report on the nature of the segments (natural, reproducible, or constructive).

We first illustrate the procedure using the artificial mobile phone data set containing three distinct, well-separated natural segments. The following command fully automates the bootstrapping procedure, and can distribute calculations to enable parallel processing. The simple artificial example below takes approximately 80 seconds on an Intel Xeon E5 2.4GHz CPU, but only 5 seconds when running 40 R processes in parallel using the same CPU. There are some fixed communication overheads to start the 40 child processes and collect their results, hence the time needed is more than the theoretical value of $80/40 = 2$ seconds. For more complex examples with higher-dimensional and larger data sets, the communication overhead is much smaller in relation to the actual computing time. Details on distributing computational tasks are provided on the help page for function `bootFlexclust()` which can be accessed in R using `help("bootFlexclust")`. The following command applies the bootstrap procedure for $k = 2$ to 9 segments, using function `cclust` as segmentation algorithm with `nrep = 10` random restarts:

```
R> set.seed(1234)
R> PF3.b29 <- bootFlexclust(PF3, k = 2:9, FUN = "cclust",
```

```
+ nrep = 10)
R> summary(PF3.b29)

Call:
bootFlexclust(x = PF3, k = 2:9, FUN = "cclust", nrep = 10)

Summary of Rand Indices:
          2           3           4           5
Min. :0.89   Min. :1   Min. :0.60   Min. :0.45
1st Qu.:0.94 1st Qu.:1   1st Qu.:0.61   1st Qu.:0.62
Median :0.97 Median :1   Median :0.80   Median :0.66
Mean   :0.96 Mean   :1   Mean   :0.75   Mean   :0.69
3rd Qu.:0.99 3rd Qu.:1   3rd Qu.:0.85   3rd Qu.:0.76
Max.  :1.00 Max.  :1   Max.  :1.00   Max.  :0.99
          6           7           8
Min. :0.55   Min. :0.52   Min. :0.50
1st Qu.:0.65 1st Qu.:0.68 1st Qu.:0.69
Median :0.70 Median :0.72   Median :0.73
Mean   :0.73 Mean   :0.73   Mean   :0.72
3rd Qu.:0.80 3rd Qu.:0.78 3rd Qu.:0.76
Max.  :0.97 Max.  :0.93   Max.  :0.93
          9
Min. :0.50
1st Qu.:0.65
Median :0.70
Mean   :0.72
3rd Qu.:0.75
Max.  :0.96
```

A parallel boxplot of the adjusted Rand indices is shown in the top right panel of Fig. 7.39. The boxplot can be obtained by:

```
R> boxplot(PF3.b29, ylim = c(0.2, 1),
+   xlab = "number of segments",
+   ylab = "adjusted Rand index")
```

As can be seen from both the numeric output and the global stability boxplot in the top right corner of Fig. 7.39 for the artificial mobile phone data set: using the correct number of three market segments always results in the same partition. All adjusted Rand indices are equal to 1 for three segments. Using fewer or more segments decreases the global stability of the segmentation solution in its entirety. This happens because the three natural segments either have to be forced into two clusters, or because the three natural segments have to be split up to generate more than three segments. Both the merger and the split is artificial because the resulting segments do not reflect the actual data structure. As a consequence, the results are not stable. The global stability boxplot indicates that, in this case, there are three natural clusters in the data. Of course – for the simple two-dimensional artificial mobile phone data set – this can easily be inferred from the top left corner in Fig. 7.39. But such a simple visual inspection is not possible for higher-dimensional data.

Looking at the global stability boxplots for the reproducible and constructive segmentation cases in Fig. 7.39 makes it obvious that no single best solution exists. One could argue that the two-segment solution for the elliptic data in the middle row is very stable, but two market segments need to be interpreted with care as they often reflect nothing more than a split of respondents in high and low response or behavioural patterns. Such high and low patterns are not very useful for subsequent marketing action.

For higher-dimensional data – where it is not possible to simply plot the data to determine its structure – it is unavoidable to conduct stability analysis to gain insight into the likely conceptual nature of the market segmentation solution. The study by Ernst and Dolnicar (2018) – which aimed at deriving a rough estimate of how frequently natural, reproducible and constructive segmentation is possible in empirical data – offered the following guidelines for assessing global stability boxplots based on the inspection of a wide range of empirical data sets:

- Indicative of natural segments are global stability boxplots with high stability and low variance of the overall market segmentation solution for at least a limited range of numbers of segments, and a distinct drop in global stability for all other numbers of segments.
- Indicative of reproducible segmentation are global stability boxplots – which starting from a reasonable high stability – show a gradual decline in the global stability of the market segmentation solution with increasing numbers of segments.
- Indicative of constructive segmentation are stability boxplots which display near-constant low stability across the overall market segmentation solutions for all numbers of segments.

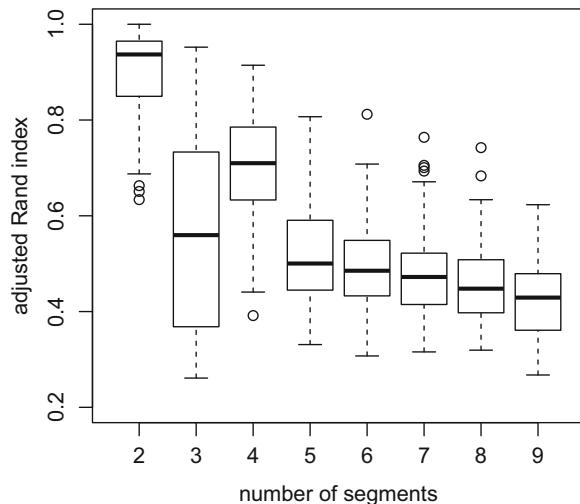
Example: Tourist Risk Taking

We illustrate global stability analysis using the data on risk taking behaviours by tourists.

```
R> data("risk", package = "MSA")
R> set.seed(1234)
R> risk.b29 <- bootFlexclust(risk, k = 2:9,
+     FUN = "cclust", nrep = 10)
```

As can be seen in the global stability boxplot in Fig. 7.40, the two- and the four-segment solutions display high levels of global stability compared to the other numbers of segments. The two-segment solution splits consumers in low and high risk takers. The four-segment solution is more profiled and may therefore contain a useful target segment for an organisation. It contains one market segment characterised by taking recreational risks, but not health risks; and a second segment that takes health, financial and safety risks, but not recreational, career or social risks. The first of those two may well represent an attractive target segment for

Fig. 7.40 Global stability boxplot using k -means clustering for 2-9 segments for the tourist risk taking data set



a tourism destination specialising in action packed adventure activities, such as bungee jumping, skydiving or paragliding.

For data analysts preferring graphical user interfaces to the command line, the complete bootstrapping procedure for global segment level stability analysis is integrated into the R Commander (Fox 2017) point-and-click interface to R in the extension package `RcmdrPlugin.BCA` (Putler and Krider 2012).

The stability analysis presented in this section assesses the *global* stability of the *entire* segmentation solution. In case of the four-segment solution it assesses the stable recovery of *all four* segments. This is a very useful approach to learn about the segmentation concept that needs to be followed. It also provides valuable guidance for selecting the number of segments to extract. However, global stability does not provide information about the stability of *each one of the segments individually* in the four-segment solution. Segment level stability is important information for an organisation because, after all, the organisation will never target a *complete* segmentation solution. Rather, it will target *one* segment or a small number of segments contained in a market segmentation solution. An approach to assessing segment level stability is presented next.

7.5.4 Segment Level Stability Analysis

Choosing the *globally* best segmentation solution does not necessarily mean that this particular segmentation solution contains the *single best* market segment. Relying on global stability analysis could lead to selecting a segmentation solution with suitable global stability, but without a single highly stable segment. It is recommendable, therefore, to assess not only *global* stability of alternative mar-

ket segmentation solutions, but also *segment level* stability of market segments contained in those solutions to protect against discarding solutions containing interesting individual segments from being prematurely discarded. After all, most organisations only need one single target segment.

7.5.4.1 Segment Level Stability Within Solutions (SLS_W)

Dolnicar and Leisch (2017) propose to assess segmentation solutions based on an approach that determines stability separately for each segment, rather than for the entire market segmentation solution. This prevents an overall bad market segmentation solution (containing one suitable market segment) from being discarded. Many organisations want to only target one segment; one suitable market segment is all they need to secure their survival and competitive advantage.

The criterion of *segment level stability within solutions* (SLS_W) is similar to the concept of global stability (see Sect. 7.5.3). The difference is that stability is computed at segment level, allowing the detection of one highly stable segment (for example a potentially attractive niche market) in a segmentation solution where several or even all other segments are unstable.

Segment level stability within solutions (SLS_W) measures how often a market segment with the same characteristics is identified across a number of repeated calculations of segmentation solutions with the *same* number of segments. It is calculated by drawing several bootstrap samples, calculating segmentation solutions independently for each of those bootstrap samples, and then determining the maximum agreement across all repeated calculations using the method proposed by Hennig (2007). Details are provided in Leisch (2015) and Dolnicar and Leisch (2017).

Hennig (2007) recommends the following steps:

1. Compute a partition of the data (a market segmentation solution) extracting k segments $\mathcal{S}_1, \dots, \mathcal{S}_k$ using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model).
2. Draw b bootstrap samples from the sample of consumers including as many cases as there are consumers in the original data set ($b = 100$ bootstrap samples works well).
3. Cluster all b bootstrap samples into k segments. Based on these segmentation solutions, assign the observations in the original data set to segments $\mathcal{S}_1^i, \dots, \mathcal{S}_k^i$ for $i = 1, \dots, b$.
4. For each bootstrap segment $\mathcal{S}_1^i, \dots, \mathcal{S}_k^i$, compute the maximum agreement with the original segments $\mathcal{S}_1, \dots, \mathcal{S}_k$ as measured by the Jaccard index:

$$s_h^i = \max_{1 \leq h' \leq k} \frac{|\mathcal{S}_h \cap \mathcal{S}_{h'}^i|}{|\mathcal{S}_h \cup \mathcal{S}_{h'}^i|}, \quad 1 \leq h \leq k.$$

The Jaccard index is the ratio between the number of observations contained in both segments, and the number of observations contained in at least one of the two segments.

5. Create and inspect boxplots of the s_h^i values across bootstrap samples to assess the segment level stability within solutions (SLS_W). Segments with higher segment level stability within solutions (SLS_W) are more attractive.

To demonstrate the procedure, consider the artificial mobile phone data set from Sect. 7.2.3. Three distinct and well-separated segments are known to exist in this data because the data was artificially generated. If – in the process of data-driven market segmentation – three segments are extracted, the correct segments emerge, and segment level stability within solutions (SLS_W) is very high. If the data are clustered into more than three segments, one of the larger natural segments is split up. This split is not stable, manifesting in a low segment level stability within solutions (SLS_W) for at least some segments. In the following, we inspect segment level stability within solutions for the six-segment solution.

To illustrate this with the artificial mobile phone data set, the data first needs to be loaded. We then cluster the data into three to eight segments. We will also use this data set to illustrate the methods in Sect. 7.5.4.2. At that point we will need all segmentation solutions from three to eight segments, and we will need all segments to be consistently labelled across segmentation solutions. Consistent labelling is achieved using function `relabel`. Finally we save the three- and six-cluster solutions into individual objects:

```
R> library("flexclust")
R> set.seed(1234)
R> PF3 <- priceFeature(500, which = "3clust")
R> PF3.k38 <- stepcclust(PF3, k = 3:8, nrep = 10)
R> PF3.k38 <- relabel(PF3.k38)
R> PF3.k3 <- PF3.k38[["3"]]
R> PF3.k6 <- PF3.k38[["6"]]
```

Figure 7.41 shows the segmentation solutions for three and six segments. Assessing the *global* stability of the two segmentation solutions (as discussed in Sect. 7.5.3) reveals that the three-segment solution is much more stable than the six-segment solution. This is evident from inspecting the top right hand plot of Fig. 7.39: if three segments are extracted, the same segmentation solution is obtained for each bootstrap sample; stability values are always equal to 1, and the box in the boxplot is a horizontal line. Stability values are lower and more variable if six segments are extracted.

To assess segment level stability within solutions (SLS_W), we use the following R commands:

```
R> PF3.r3 <- slswFlexclust(PF3, PF3.k3)
R> PF3.r6 <- slswFlexclust(PF3, PF3.k6)
```

R function `slswFlexclust()` from package `flexclust` takes as input the original data `PF3` to create bootstrap samples. Then, segment level stability within solutions (SLS_W) is calculated for the three-segment solution (`PF3.k3`) and

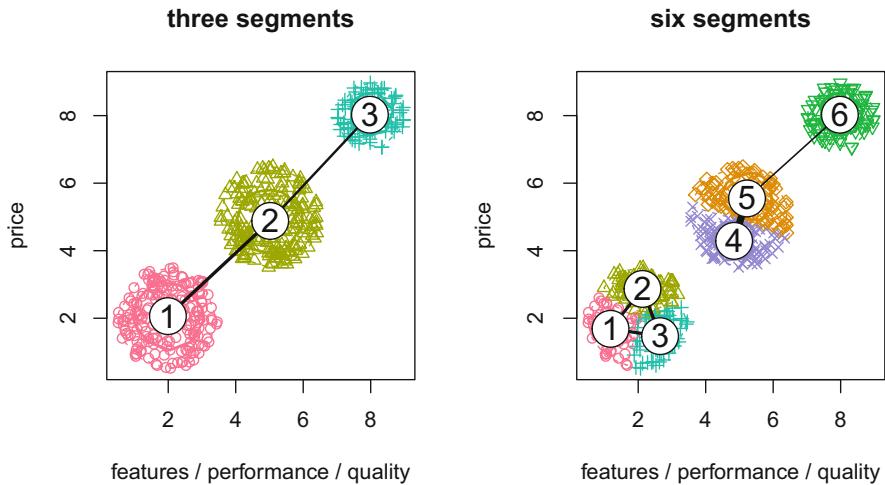


Fig. 7.41 Artificial mobile phone data set with three and six segments extracted

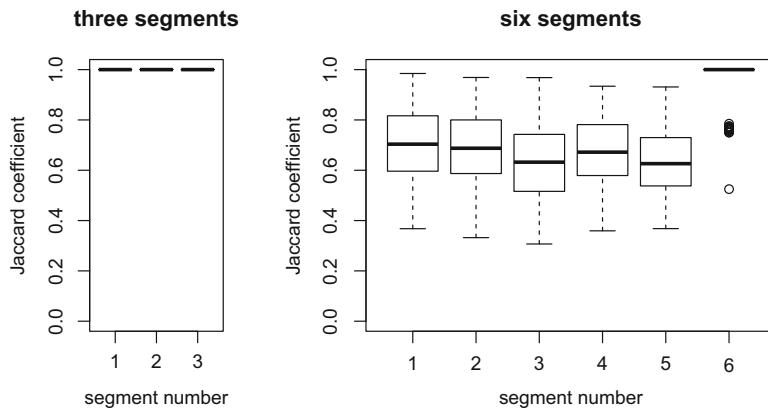


Fig. 7.42 Segment level stability within solutions (SLS_W) plot for the artificial mobile phone data set with three and six segments extracted

the six-segment solution (PF3.k6). `s1swFlexclust` implements the stepwise procedure described above slightly differently. `s1swFlexclust` draws pairs of bootstrap samples, and returns the average agreement measured by the average Jaccard index for each pair.

We obtain boxplots showing the segment level stability within solutions (SLS_W) (Fig. 7.42) using `plot(PF3.r3)` and `plot(PF3.r6)`. As can be seen, all three segments contained in the three-segment solution have the maximal stability of 1. The boxes in Fig. 7.42 therefore do not look like boxes at all. Rather, they present as thick horizontal lines at value 1. For the artificially generated mobile phone data set this is not surprising; the data set contain three distinct and well-separated segments.

Looking at the segment level stability within solutions (SLS_W) for the six-segment solution on the right side of Fig. 7.42 indicates that only segment 6 in this solution is very stable. The other segments are created by randomly splitting up the two market segments not interested in high-end mobile phones. The fact that market segments not interested in expensive mobile phones with many features are not extracted in a stable way is irrelevant to a manufacturer of premium mobile phones. Such a manufacturer is only interested in the correct identification of the high-end segment because this is the segment that will be targeted. This one segment may be all that such a mobile phone manufacturer needs to survive and maximise competitive advantage.

This insight is only possible if segment level stability within solutions (SLS_W) is assessed. If the segmentation solution would have only been chosen based on the inspection of the global stability boxplot in Fig. 7.39, the six-segment solution would have been discarded.

For two-dimensional data (like the mobile phone data set), data structure – and with it the correctness of a market segmentation solution – is seen by simply taking a quick look at a scatter plot of the actual data. Typical consumer data, however, is not two-dimensional; it is multi-dimensional. Each segmentation variable represents one dimension. The Australian vacation activities data set used in Sect. 7.4.1, for example, contains 45 segmentation variables. The data space, therefore, is 45-dimensional, and cannot be plotted in the same way as the simple mobile phone data set. Analysing data structure thoroughly when extracting market segments is therefore critically important.

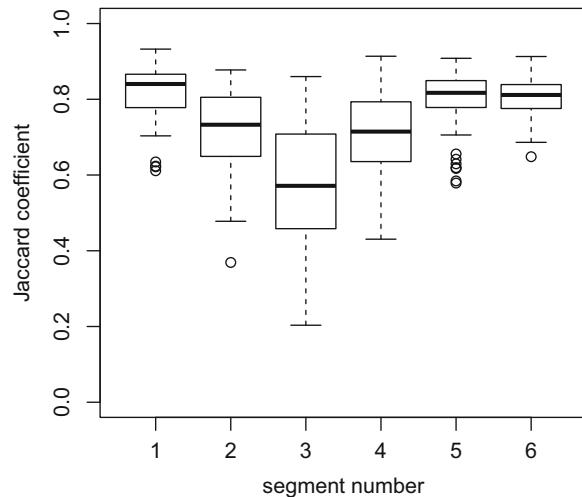
Example: Australian Travel Motives

To illustrate the use of segment level stability within solutions (SLS_W) on real consumer data, we use the data containing 20 travel motives of 1000 Australian residents presented in Step 4 (see Appendix C.4). We load the data set (available in package `flexclust`) into R using:

```
R> library("flexclust")
R> data("vacmot", package = "flexclust")
```

When the data was segmented for the first time (in Dolnicar and Leisch 2008), several clustering algorithms and numbers of clusters were tried. The data set does not contain natural segments. As a consequence, the clustering algorithm will impose structure on the segments extracted from the data. Selecting a suitable algorithm is therefore important. The neural gas algorithm (Martinetz and Schulten 1994) delivered the most interesting segmentation solution for six clusters. Unfortunately the seed used for the random number generator has been lost in the decade since the first analysis, hence the result presented here deviates slightly from that reported in Dolnicar and Leisch (2008). Nevertheless, all six segments re-emerge in the new partition, but with different segment numbering, and slightly different centroid values and segment sizes.

Fig. 7.43 Segment level stability within solutions (SLS_W) plot of the six-segment solution for the Australian travel motives data set



We obtain a series of segmentation solutions ranging from three to eight segments by using neural gas clustering (argument `method = "neuralgas"` with `nrep = 20` random restarts):

```
R> set.seed(1234)
R> vacmot.k38 <- stepcclust(vacmot, k = 3:8,
+   method = "neuralgas", nrep = 20, save.data = TRUE,
+   verbose = FALSE)
R> vacmot.k38 <- relabel(vacmot.k38)
```

Because these segmentation solutions will be reused as examples in Steps 5 and 7, we integrate the original data set into the cluster object by setting `save.data = TRUE`. In addition, `verbose = FALSE` avoids printing of progress information of the calculations to the console. Finally, we save the entire series of segmentation solutions to the hard drive:

```
R> vacmot.k6 <- vacmot.k38[["6"]]
R> save(vacmot.k38, vacmot.k6,
+   file = "vacmot-clusters.RData")
```

Next, we assess segment level stability within solutions (SLS_W) for the six-segment solution. In addition to the data set `vacmot`, and the fitted partition `vacmot.k6`, we need to specify that the neural gas method of function `cclust()` is used:

```
R> vacmot.r6 <- slswFlexclust(vacmot, vacmot.k6,
+   method = "neuralgas", FUN = "cclust")
```

Figure 7.43 shows the resulting boxplot. Segments with the highest segment level stability within solutions (SLS_W) are segments 1, 5 and 6, followed by 2 and 4. Segments 1 and 5 will be identified as likely response style segments in Step 6. This means that the pattern of responses by members of these segments may be caused by the way they interact with the answer format offered to them in the survey, rather

than reflecting their responses to the content. Segment 6 – which is not suspicious in terms of response style bias – is also very stable, and displays an interesting profile (discussed in Step 6). Making segment 6 even more interesting is the fact that members display characteristic descriptor variables (discussed in Step 7). Segment 3 represents tourists interested in the lifestyle of the local people, caring about unspoilt nature, wishing to maintain unspoilt surroundings, and wanting to intensely experience nature. They do not want entertainment facilities, and they have no desire for luxury or to be spoilt. Only segment 3 emerges as being very unstable when inspecting the segment level stabilities provided in Fig. 7.43. The reason for this high level of instability will become obvious in the next section where we gain insight into the stability of segments across solutions with *different* numbers of segments.

7.5.4.2 Segment Level Stability Across Solutions (SLS_A)

The second criterion of stability at segment level proposed by Dolnicar and Leisch (2017) is referred to as *segment level stability across solutions* (SLS_A). The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing *different* numbers of segments. High values of segment level stability across solutions (SLS_A) serve as indicators of market segments occurring naturally in the data, rather than being artificially created. Natural segments are more attractive to organisations because they actually exist, and no managerial judgement is needed in the artificial construction of segments.

Let $\mathcal{P}_1, \dots, \mathcal{P}_m$ be a series of m partitions (market segmentation solutions) with $k_{\min}, k_{\min} + 1, k_{\min} + 2, \dots, k_{\max}$ segments, where $m = k_{\max} - k_{\min} + 1$. The minimum and maximum number of segments of interest (k_{\min} and k_{\max}) have to be specified by the user of the market segmentation analysis in collaboration with the data analyst.

Segment level stability across solutions (SLS_A), can be calculated in combination with any algorithm which extracts segments. However, for hierarchical clustering, segment level stability across solutions will reflect the fact that a sequence of nested partitions is created. If partitioning methods (k -means, k -medians, neural gas, ...) or finite mixture models are used, segmentation solutions are determined separately for each number of segments k . A common problem with these methods, however, is that the segment labels are random and depend on the random initialisation of the extraction algorithm (for example the segment representatives which are randomly drawn from the data at the start). To be able to compare market segmentation solutions, it is necessary to identify which segments in each of the solutions with neighbouring numbers of segments ($\mathcal{P}_i, \mathcal{P}_{i+1}$) are similar to each other and assign consistent labels. The difference in number of segments complicates this task. A way around this problem is to first sort the segments in \mathcal{P}_1 using any heuristic, then renumber \mathcal{P}_2 such that segments that are similar to segments in \mathcal{P}_1 get suitable numbers assigned as labels, etc.

Based on this idea, Dolnicar and Leisch (2017) propose an algorithm to *relabel series of partitions (segmentation solutions)*, which is implemented in function

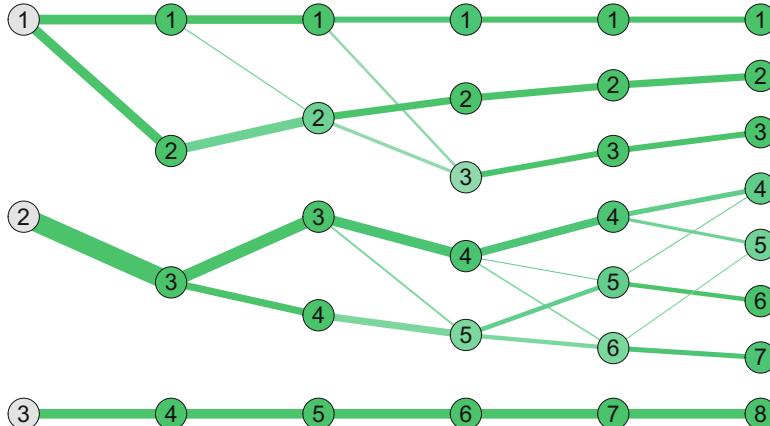


Fig. 7.44 Segment level stability across solutions (SLS_A) plot for the artificial mobile phone data set for three to eight segments

`relabel()` in package `flexclust`. This function was used on pages 168 and 171 to renumber segmentation solutions. Once segments are suitably labelled, a segment level stability across solutions (SLS_A) plot can be created.

We use the artificial mobile phone data set to illustrate the usefulness of segment level stability across solutions (SLS_A) as guidance for the data analyst. We create the segment level stability across solutions (SLS_A) plot in Fig. 7.44 using the command `slsaplot(PF3.k38)` from package `flexclust`. This plot shows the development of each segment across segmentation solutions with different numbers of segments.

Each column in the plot represents a segmentation solution with a specific number of segments. The number of segments extracted increases from left to right. The column on the far left represents the segmentation solution with three segments. The column on the far right represents the segmentation solution with eight segments. The lines between segments indicate movements of segment members between segments. Thick lines between two segments indicate that many segment members are retained (despite the number of segments increasing). Thick lines represent stubborn market segments, market segments which re-occur across segmentation solutions, and therefore are more likely to represent natural segments. Segments which have many lines coming in from the left and branching into many lines to their right, suffer from changing segment membership across calculations with different numbers of segments. Such segments are more likely to be artificially created during the segment extraction process.

For the artificial mobile phone data set containing three distinct market segments, the segment level stability across solutions (SLS_A) plot offers the following insights: segment 3 in the three-segment solution remains totally unchanged across segmentation solutions with different numbers of segments. Segment 3 is the high-end mobile phone market segment. Segments 1 and 2 in the three-segment solution are split up into more and more subsegments as the number of market segments in

the segmentation solution increases. The segment level stability across solutions (SLS_A) plot confirms what is seen to happen in the right chart in Fig. 7.41: if more than three segments are extracted from the mobile phone data set, the high-end segment continues to be identified correctly. The other two (larger) segments gradually get subdivided.

So far all interpretations of segment level stability across solutions (SLS_A) were based on visualisations only. The measure of entropy (Shannon 1948) can be used as a numeric indicator of segment level stability across solutions (SLS_A). Let p_j be the percentage of consumers segment \mathcal{S}_l^i (segment l) in partition (segmentation solution) \mathcal{P}_i recruits from each segment \mathcal{S}_j^{i-1} in partition (segmentation solution) \mathcal{P}_{i-1} , with $j = 1, \dots, k_{i-1}$. One extreme case is if one value p_{j^*} is equal to 1 and all others are equal to 0. In this case segment \mathcal{S}_l^i recruits all its members from segment $\mathcal{S}_{j^*}^{i-1}$ in the smaller segmentation solution; it is identical in both solutions and maximally stable. The other extreme case is that the p_j 's are all the same, that is, $p_j = 1/k_{i-1}$ for $j = 1, \dots, k_{i-1}$. The new segment \mathcal{S}_l^i recruits an equal share of consumers from each segment in the smaller segmentation solution; the segment has minimal stability.

Entropy is defined as $-\sum p_j \log p_j$ and measures the uncertainty in a distribution. Maximum entropy is obtained for the uniform distribution with $p_j = 1/k$; the entropy is then $-\sum(1/k) \log(1/k) = \log(k)$. The minimum entropy is 0 and obtained if one p_j is equal to 1. Numerical stability $SLS_A(\mathcal{S}_l^i)$ of segment l in the segmentation solution with k_i segments is defined by

$$SLS_A(\mathcal{S}_l^i) = 1 - \frac{\sum_{j=1}^{k_{i-1}} p_j \log p_j}{\log(k_{i-1})}.$$

A value of 0 indicates minimal stability and 1 indicates maximal stability.

The numeric segment level stability across solutions (SLS_A) values for each segment in each segmentation solution is used in Fig. 7.44 to colour the nodes and edges. In Fig. 7.44, green is uniform across the plot because all new segments are created by splitting an existing segment into two. Each segment in the larger segmentation solution only has one single parent in the smaller partition, hence low entropy and high stability.

Example: Australian Travel Motives

Figure 7.45 contains the segment level stability across solutions (SLS_A) plot for the Australian travel motives data set. The segmentation solutions were saved for later re-use on page 171, and the plot results from `s1splot(vacmot.k38)`.

The numeric segment level stability across solution (SLS_A) values for each segment in each segmentation solution used to colour nodes and edges indicate that the segments in the top and bottom rows do not change much from left to right. The corresponding nodes and edges are all solid green. The only exception is the jump from four to five segments, where some members are recruited from other segments

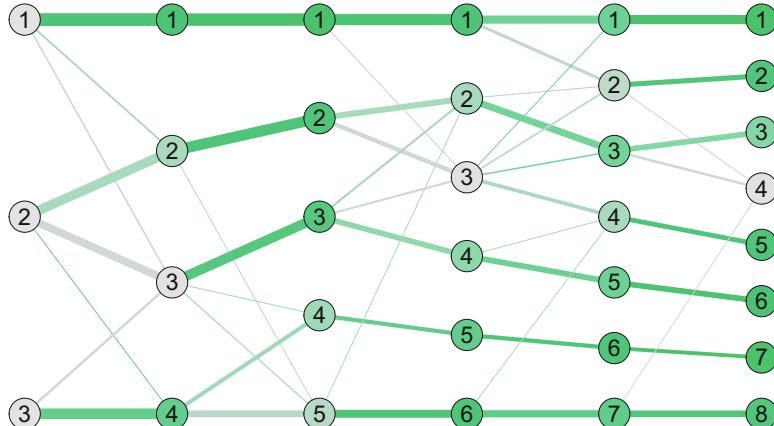


Fig. 7.45 Segment level stability across solutions (SLS_A) plot for the Australian travel motives data set for three to eight segments

by segment 5 in the five-segment solution. The opposite is true for segment 3 in the six-segment solution. Segment 3 recruits its members almost uniformly from segments 1, 2 and 3 in the five-segment solution; the corresponding node and edges are all light grey.

From Fig. 7.45 the segment labelled segment 1 in each segmentation solution emerges as the segment with the highest average segment level stability across solutions (SLS_A) value over all segmentation solutions. However – upon inspection of the profile of this particular segment (Fig. 8.2) – it becomes clear that it may represent (at least partially) a response style segment. Response bias is displayed by survey respondents who have a tendency to use certain response options, irrespective of the question asked. But an average high segment level stability across solutions (SLS_A) value driven by a response style does not make a market segment attractive as a potential target segment. The segment with the second highest segment level stability across solutions (SLS_A) value in Fig. 7.45 is segment 6 in the six-segment solution. This particular segment hardly changes at all between the six- and the eight-segment solutions. Note that, in the eight-segment solution, segment 6 is renamed segment 8. Looking at the segment profile plot in Fig. 8.2, it can be seen that members of this segment are tourists interested in the lifestyle of locals, and caring deeply about nature.

From Fig. 7.45 it also becomes obvious why segment 3 in the six-segment solution demonstrates low segment level stability within solution (SLS_W). Segment 3 emerges as an entirely new segment in the six-segment solution by recruiting members from several segments contained in the five-segment solution. Then, segment 3 immediately disappears again in the seven-segment solution by distributing its members across half of the segments in the seven-segment solution. It is safe to conclude that segment 3 is not a natural segment. Rather, it represents

a grouping of consumers the algorithm was forced to extract because we asked for six segments.

Two key conclusions can be drawn from the segment level stability across solutions (SLS_A) plot in Fig. 7.45: seriously consider segment 6 in the six-segment solution as a potential target segment because it shows all signs of a naturally existing market segment. Do not consider targeting segment 3. It is an artefact of the analysis.

7.6 Step 5 Checklist

Task	Who is responsible?	Completed?
Pre-select the extraction methods that can be used given the properties of your data.		<input type="checkbox"/>
Use those suitable extraction methods to group consumers.		<input type="checkbox"/>
Conduct global stability analyses and segment level stability analyses in search of promising segmentation solutions and promising segments.		<input type="checkbox"/>
Select from all available solutions a set of market segments which seem to be promising in terms of segment-level stability.		<input type="checkbox"/>
Assess those remaining segments using the knock-out criteria you have defined in Step 2.		<input type="checkbox"/>
Pass on the remaining set of market segments to Step 6 for detailed profiling.		<input type="checkbox"/>

References

- Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken
- Akaike H (1987) Factor analysis and AIC. *Psychometrika* 52(3):317–332
- Allenby GM, Arora N, Ginter JL (1998) On the heterogeneity of demand. *J Mark Res* 35(3): 384–389
- Arabie P, Hubert L (1994) Cluster analysis in marketing research. In: Bagozzi R (ed) Advanced methods of marketing research. Blackwell, Cambridge, pp 160–189
- Ball G, Hall D (1965) ISODATA, a novel method of data analysis and pattern classification. Tech. Rep. NTIS No. AD 699616, Stanford Research Institute, Menlo Park
- Bhatnagar A, Ghose S (2004) A latent class segmentation analysis of e-shoppers. *J Bus Res* 57(7):758–767
- Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell* 22(7):719–725
- Bijmolt THA, Paas LJ, Vermunt JK (2004) Country and consumer segmentation: multi-level latent class analysis of financial product ownership. *Int J Res Mark* 21(4):323–340

- Bockenholt U, Langeheine UR (1996) Latent change in recurrent choice data. *Psychometrika* 61(2):285–301
- Boztug Y, Reutterer T (2008) A combined approach for segment-specific market basket analysis. *Eur J Oper Res* 187(1):294–312
- Brangule-Vlagsma K, Pieters RGM, Wedel M (2002) The dynamics of value segments: modelling framework and empirical illustration. *Int J Res Mark* 19(3):267–285
- Breckenridge JN (1989) Replicating cluster analysis: method, consistency, and validity. *Multivar Behav Res* 24(2):147–161
- Brusco M (2004) Clustering binary data in the presence of masking variables. *Psychol Methods* 9(4):510–523
- Calantone RJ, Sawyer AG (1978) The stability of benefit segments. *J Mark Res* 15(3):395–404
- Calinski RB, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat* 3:1–27
- Campbell C, Ferraro C, Sands S (2014) Segmenting consumer reactions to social network marketing. *Eur J Mark* 48(3/4):432–452
- Casella G, Berger RL (2010) Statistical inference, 2nd edn. Cengage Learning, Stamford, Conn
- Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Software* 61(6):1–36
- Choffrey JM, Lilien GL (1980) Industrial market segmentation. In: Choffrey M, Lilien GL (eds) Marketing planning for new industrial products. Wiley, New York, pp 74–91
- Dayton CM, Macready GB (1988) Concomitant-variable latent-class models. *J Am Stat Assoc* 83(401):173–178
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38
- Dolnicar S (2002a) Activity-based market sub-segmentation of cultural tourists. *J Hosp Tour Manag* 9(2):94–105
- Dolnicar S (2002b) Review of data-driven market segmentation in tourism. *J Travel Tour Mark* 12(1):1–22
- Dolnicar S, Grün B (2008) Challenging “factor-cluster segmentation”. *J Travel Res* 47(1):63–71
- 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 (2004) Segmenting markets by bagged clustering. *Australas Mark J* 12(1): 51–65
- Dolnicar S, Leisch F (2008) An investigation of tourists’ patterns of obligation to protect the environment. *J Travel Res* 46:381–391
- Dolnicar S, Leisch F (2010) Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Mark Lett* 21:83–101
- Dolnicar S, Leisch F (2014) Using graphical statistics to better understand market segmentation solutions. *Int J Mark Res* 56(2):97–120
- Dolnicar S, Leisch F (2017) Using segment level stability to select target segments in data-driven market segmentation studies. *Mark Lett* 28(3):423–436
- Dolnicar S, Kaiser S, Lazarevski K, Leisch F (2012) Bioclustering – overcoming data dimensionality problems in market segmentation. *J Travel Res* 51(1):41–49
- Dolnicar S, Grün B, Leisch F, Schmidt K (2014) Required sample sizes for data-driven market segmentation analyses in tourism. *J Travel Res* 53(3):296–306
- Dolnicar S, Grün B, Leisch F (2016) Increasing sample size compensates for data problems in segmentation studies. *J Bus Res* 69:992–999
- Doyle P, Saunders J (1985) Market segmentation and positioning in specialized industrial markets. *J Mark* 49(2):24–32
- Dubossarsky E, Tyshetskiy Y (2015) autoencoder: sparse autoencoder for automatic learning of representative features from unlabeled data. <https://CRAN.R-project.org/package=autoencoder>, R package version 1.1
- Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 3(7):1–21
- Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman & Hall, New York

- Ernst D, Dolnicar S (2018) How to avoid random market segmentation solutions. *J Travel Res.* 57(1):69–82.
- Esslemont DH, Ward T (1989) The stability of segmentation solutions in a commercial survey. *N Z J Bus* 10(2):89–95
- Ferraro M, Giordani P (2015) A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets Syst* 279:1–16
- Forgy EW (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21:768–769
- Fox J (2017) Using the R Commander: a point-and-click interface for R. Chapman & Hall/CRC Press, Boca Raton
- Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* 41:578–588
- Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc* 97(458):611–631
- Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report No. 597, Department of Statistics, University of Washington
- Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer series in statistics. Springer, New York
- Frühwirth-Schnatter S (2011) Panel data analysis: a survey on model-based clustering of time series. *Adv Data Anal Classif* 5(4):251–280
- Frühwirth-Schnatter S, Tüchler R, Otter T (2004) Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. *Int J Res Mark* 21(3):285–297
- Funkhouser GR (1983) A note on the reliability of certain clustering algorithms. *J Mark Res* 20(1):99–102
- Gana Dresen I, Boes T, Huesing J, Neuhaeuser M, Joeckel KH (2008) New resampling method for evaluating stability of clusters. *BMC Bioinf* 9(1):42
- Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27(4):857–871
- Grün B, Dolnicar S (2016) Response-style corrected market segmentation for ordinal data. *Mark Lett* 27:729–741
- Grün B, Leisch F (2004) Bootstrapping finite mixture models. In: Antoch J (ed) Compstat 2004 – Proceedings in computational statistics. Physica Verlag, Heidelberg, pp 1115–1122
- Grün B, Leisch F (2007) Fitting finite mixtures of generalized linear regressions in R. *Comput Stat Data Anal* 51(11):5247–5252
- Grün B, Leisch F (2008) FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J Stat Softw* 28(4):1–35
- Hajibaba H, Karlsson L, Dolnicar S (2017) Residents open their homes to tourists when disaster strikes. *J Travel Res* 56(8):1065–1078
- Haley RI (1985) Developing effective communications strategy – a benefit segmentation approach. Wiley, New York
- Hartigan JA (1972) Direct clustering of a data matrix. *J Am Stat Assoc* 67(337):123–129
- Hartigan JA, Wong MA (1979) A k-means clustering algorithm. *Appl Stat* 28:100–108
- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction. Springer, New York
- Hennig C (2007) Cluster-wise assessment of cluster stability. *Comput Stat Data Anal* 52:258–271
- Hennig C (2015) fpc: flexible procedures for clustering. <https://CRAN.R-project.org/package=fpc>, R package version 2.1-10
- Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *J R Stat Soc C* 62(3):309–369
- Hoek J, Gendall P, Esslemont D (1996) Market segmentation: a search for the holy grail? *J Mark Pract: Appl Mark Sci* 2(1):25–34
- Hornik K (2005) A CLUE for CLUster Ensembles. *J Stat Softw* 14(12): 1–25

- Hruschka H, Natter M (1999) Comparing performance of feedforward neural nets and k-means for cluster-based market segmentation. *Eur J Oper Res* 114(2):346–353
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218
- IBM Corporation (2016) IBM SPSS Statistics 24. IBM Corporation, Armonk. <http://www.ibm.com/software/analytics/spss/>
- Jaccard P (1912) The distribution of the flora in the alpine zone. *New Phytolog* 11(2):37–50
- Jurowski C, Reich AZ (2000) An explanation and illustration of cluster analysis for identifying hospitality market segments. *J Hosp Tour Res* 24(1):67–91
- Kaiser HF (1960) The application of electronic computers to factor analysis. *Educ Psychol Meas* 20(1):141–151
- Kaiser S (2011) Biclustering: methods, software and application. Ph.D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München, Munich. <https://edoc.ub.uni-muenchen.de/13073/>
- Kaiser S, Leisch F (2008) A toolbox for bicluster analysis in R. In: Brito P (ed) Compstat 2008 – Proceedings in computational statistics. Physica Verlag, Heidelberg, pp 201–208
- Kasim A, Shkedy Z, Kaiser S, Hochreiter S, Talloen W (eds) (2017) Applied biclustering methods for big and high-dimensional data using R. Chapman and Hall/CRC, Boca Raton
- Kemperman ADA, Timmermanns HJP (2006) Preferences, benefits and park visits: a latent class segmentation analysis. *Tour Anal* 11:221–230
- Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biolog Cybern* 43(1):59–69
- Kohonen T (2001) Self-organizing maps. Springer series in information sciences. Springer, Berlin
- Lance GN, Williams WT (1967) A general theory of classification sorting strategies I. Hierarchical systems. *Comput J* 9:373–380
- Lange T, Roth V, Braun ML, Buhmann JM (2004) Stability-based validation of clustering solutions. *Neural Comput* 16(6):1299–1323
- Leisch F (1998) Ensemble methods for neural clustering and classification. Ph.D. thesis, Technische Universität Wien, Vienna
- Leisch F (1999) Bagged clustering. Working paper 51, SFB “Adaptive Information Systems and Modeling in Economics and Management Science”
- Leisch F (2004) FlexMix: a general framework for finite mixture models and latent class regression in R. *J Stat Softw* 11(8):1–18
- Leisch F (2006) A toolbox for k-centroids cluster analysis. *Comput Stat Data Anal* 51(2):526–544
- Leisch F (2010) Neighborhood graphs, stripes and shadow plots for cluster visualization. *Stat Comput* 20(4):457–469
- Leisch F (2015) Resampling methods for exploring cluster stability. In: Hennig C, Meila M, Murtagh F, Rocci R (eds) Handbook of cluster analysis. Handbooks of modern statistical methods. Chapman and Hall/CRC, Boca Raton, pp 637–652
- Lloyd SP (1982) Least squares quantization in PCM. *IEEE Trans Inf Theory* 28:128–137
- MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J (eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability
- Madeira S, Oliveira A (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 1(1):24–45
- Maechler M, Rousseeuw P, Struyf A, Hubert M (2017) cluster: “Finding Groups in Data” – cluster analysis extended Rousseeuw et al. <https://CRAN.R-project.org/package=cluster>, R package version 2.0.6
- Maitra R, Melnykov V, Lahiri SN (2012) Bootstrapping for significance of compact clusters in multidimensional datasets. *J Am Stat Assoc* 107(497):378–392
- Martinetz T, Schulten K (1994) Topology representing networks. *Neural Netw* 7(5):507–522
- Martinetz TM, Berkovich SG, Schulten KJ (1993) “Neural-gas” network for vector quantization and its application to time-series prediction. *IEEE Trans Neural Netw* 4(4):558–569
- Mazanec JA (1999) Simultaneous positioning and segmentation analysis with topologically ordered feature maps: a tour operator example. *J Retail Consum Serv* 6(4):219–235

- McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
- McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
- McLachlan GJ, Peel D, Bean RW (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Comput Stat Data Anal* 41(3–4):379–388
- Milligan GW (1996) Clustering validation: results and implications for applied analyses. In: Arabie P, Hubert LJ (eds) Clustering and classification. World Scientific Publications, River Edge
- Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2):159–179
- Mohamed M, Higgins C, Ferguson M, Kanaroglou P (2016) Identifying and characterizing potential electric vehicle adopters in Canada: a two-stage modelling approach. *Transp Policy* 52:100–112
- Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J Classif* 31(3):274–295
- Natter M (1999) Conditional market segmentation by neural networks: a Monte-Carlo study. *J Retail Consum Serv* 6(4):237–248
- Okazaki S (2006) What do we know about mobile internet adopters? A cluster analysis. *Inf Manag* 43:127–141
- Oppewal H, Paas LJ, Crouch GI, Huybers T (2010) Segmenting consumers based on how they spend a tax rebate: an analysis of the Australian stimulus payment. *J Econ Psychol* 31(4): 510–519
- Pertea M, Salzberg SL (2010) Between a chicken and a grape: estimating the number of human genes. *Genome Biol* 11(5):1–7
- Poulsen CS (1990) Mixed Markov and latent Markov modelling applied to brand choice behaviour. *Int J Res Mark* 7:5–19
- Prayag G, Disegna M, Cohen SA, Yan HG (2015) Segmenting markets by bagged clustering: young Chinese travelers to western Europe. *J Travel Res* 54(2):234–250
- Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22(9):1122–1129
- Putler DS, Krider RE (2012) Customer and business analytics: applied data mining for business decision making using R. Chapman&Hall/CRC, Boca Raton, FL
- Ramaswamy V (1997) Evolutionary preference segmentation with panel survey data: an application to new products. *Int J Res Mark* 14(1):57–89
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66(336):846–850
- Ratkowsky DA, Lance GN (1978) A criterion for determining the number of groups in a classification. *Aust Comput J* 10(3):115–117
- Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev* 26(2):195–239
- Reutterer T, Natter M (2000) Segmentation-based competitive analysis with MULTICLUS and topology representing networks. *Comput Oper Res* 27(11):1227–1247
- Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge, UK
- Ritchie BW, Chien PM, Sharifpour M (2017) Segmentation by travel related risks: an integrated approach. *J Travel Tour Mark* 34(2):274–289
- Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math* 20:53–65
- Schwarz GE (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Sheppard AG (1996) The sequence of factor analysis and cluster analysis: differences in segmentation and dimensionality through the use of raw and factor scores. *Tour Anal* 1:49–57
- SPSS (2001) The SPSS TwoStep cluster component. SPSS Inc., Chicago, Technical report TSCWP-0101

- Steinley D, Brusco MJ (2007) Initializing k-means batch clustering: a critical evaluation of several techniques. *J Classif* 24:99–121
- Steinley D, Brusco MJ (2008a) A new variable weighting and selection procedure for k -means cluster analysis. *Multivar Behav Res* 43(1):77–108
- Steinley D, Brusco MJ (2008b) Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika* 73(1):125–144
- Thorndike RL (1953) Who belongs in the family? *Psychometrika* 18:267–276
- Tibshirani R, Walther G (2005) Cluster validation by prediction strength. *J Comput Graph Stat* 14(3):511–528
- Tkaczynski A, Rundle-Thiele SR, Prebensen NK (2015) Segmenting potential nature-based tourists based on temporal factors: the case of Norway. *J Travel Res* 54(2):251–265
- Verbeke G, Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *J Am Stat Assoc* 91(433):217–221
- Walesiak M, Dudek A (2016) clusterSim: searching for optimal clustering procedure for a data set. <https://CRAN.R-project.org/package=clusterSim>, R package version 0.45-1
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244
- Wedel M, Kamakura W (2000) Market segmentation – conceptual and methodological foundations. Kluwer Academic, Boston
- Wehrens R, Buydens LMC (2007) Self- and super-organising maps in R: the kohonen package. *J Stat Softw* 21(5):1–19

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 8

Step 6: Profiling Segments



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

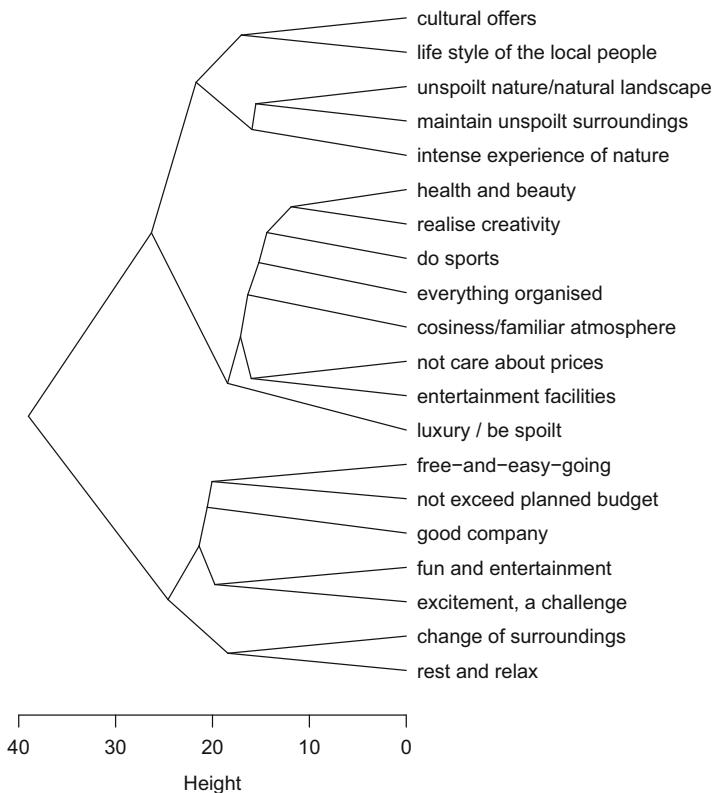


Fig. 8.1 Hierarchical clustering of the segmentation variables of the Australian travel motives data set using Ward's method

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.

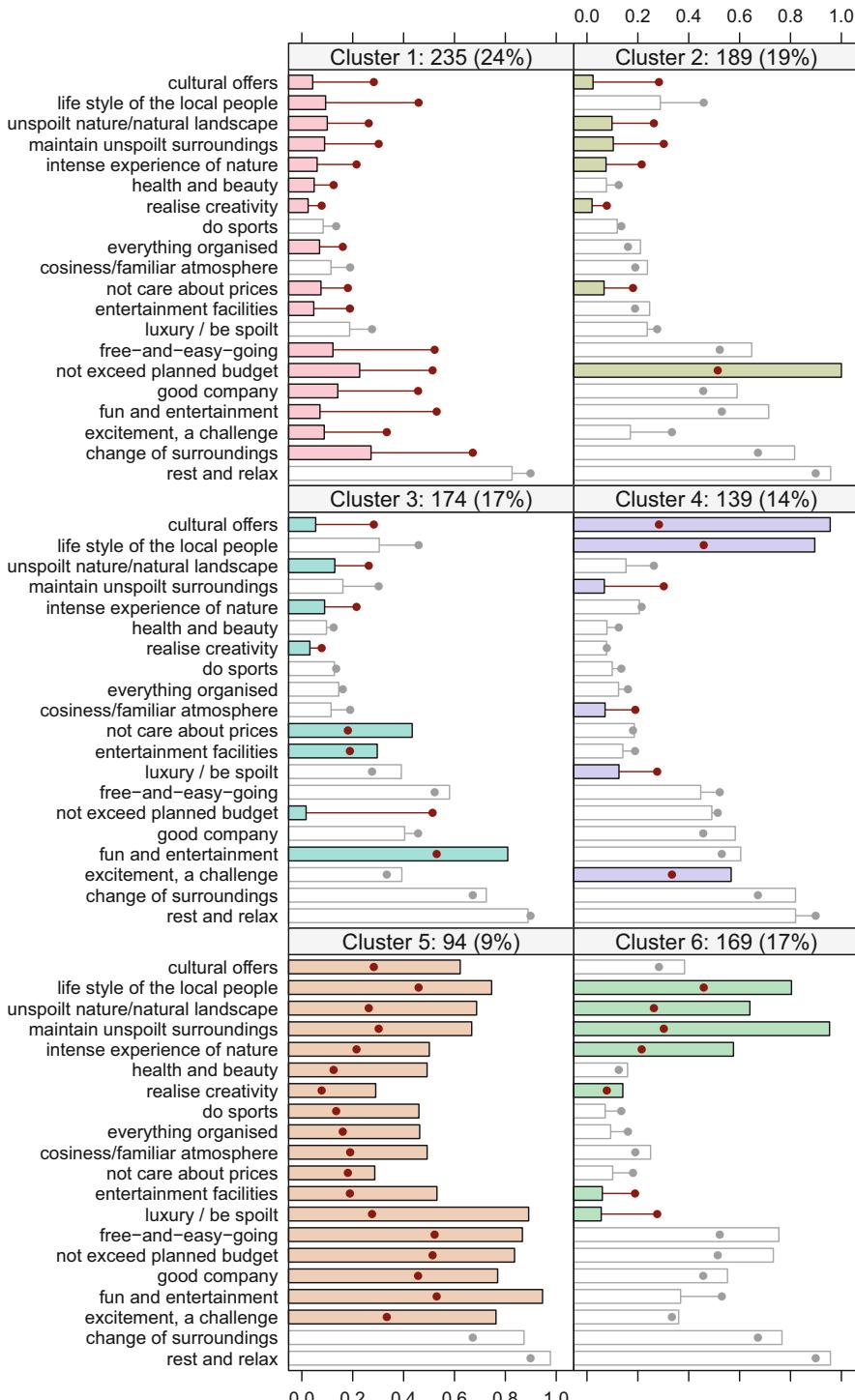


Fig. 8.2 Segment profile plot for the six-segment solution of the Australian travel motives data set

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.

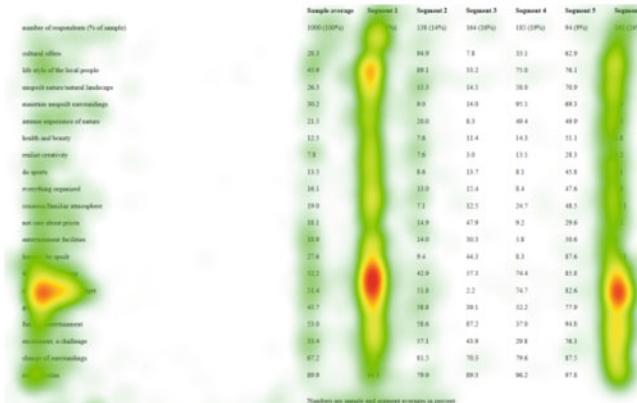
An eye tracking study conducted by Nazila Babakhani as part of her PhD studies investigated differences in people's ability to interpret complex data analysis results from market segmentation studies presented in traditional tabular versus graphical statistics format. Participants saw one of three types 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.

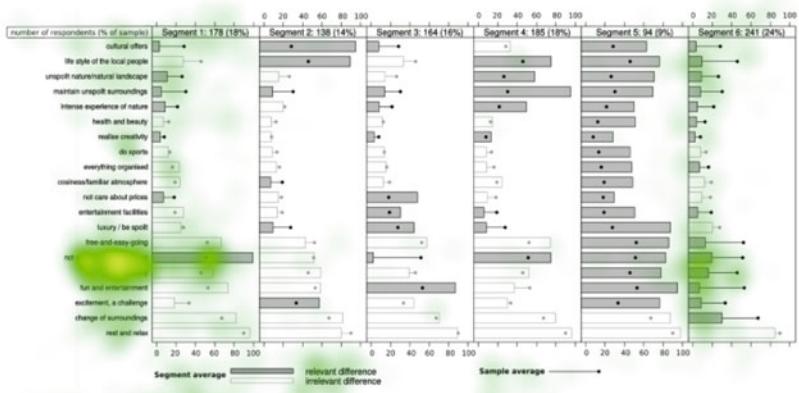
Segment separation plots are very simple if the number of segmentation variables is low, but become complex as the number of segmentation variables increases. But even in such complex situations, segment separation plots offer data analysts and users a quick overview of the data situation, and the segmentation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

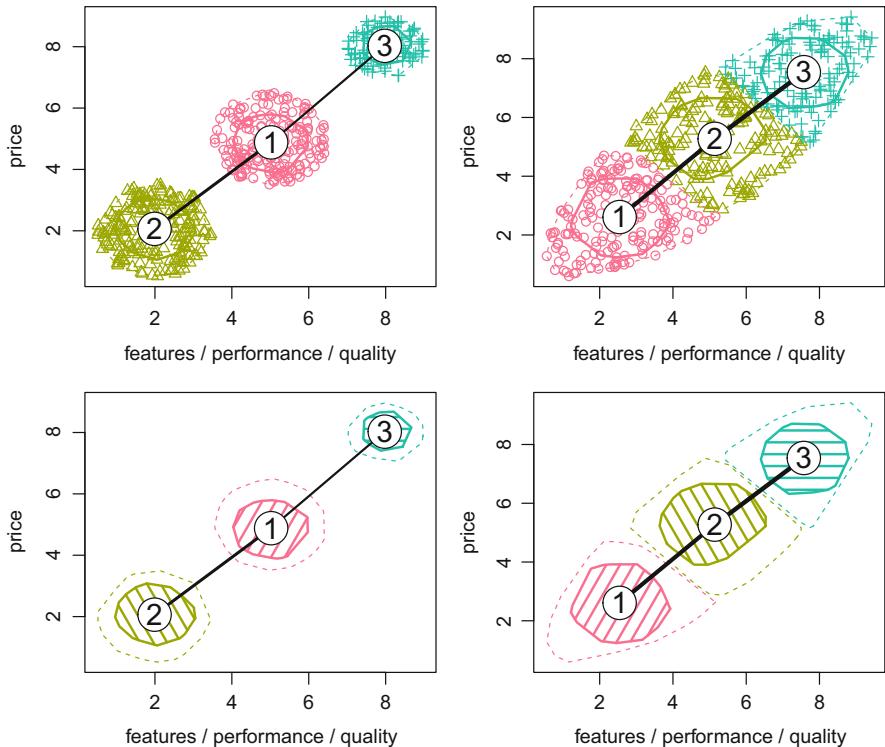


Fig. 8.4 Segment separation plot including observations (first row) and not including observations (second row) for two artificial data sets: three natural, well-separated clusters (left column); one elliptic cluster (right column)

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)
```

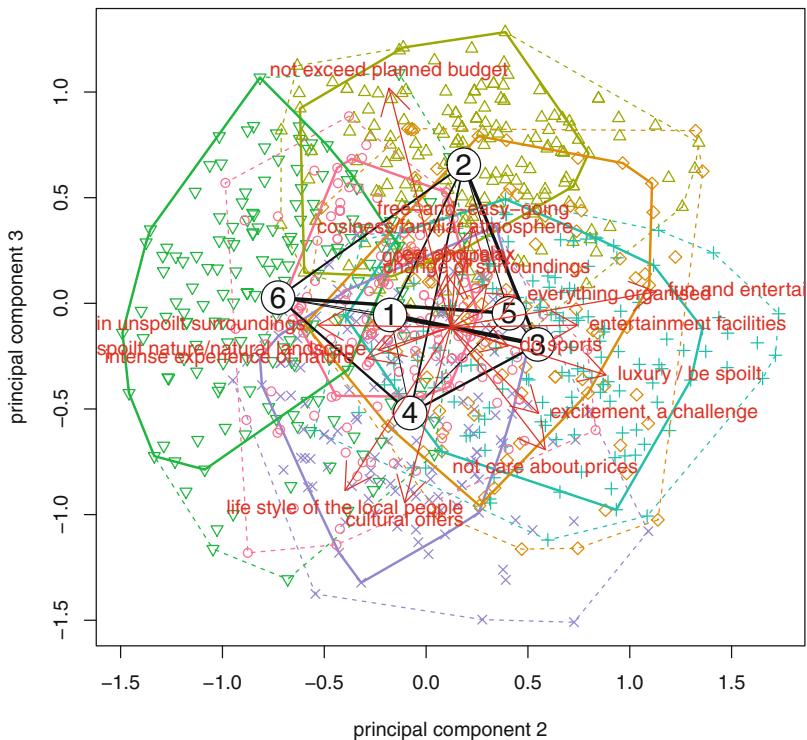


Fig. 8.5 Segment separation plot using principal components 2 and 3 for the Australian travel motives data set

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)

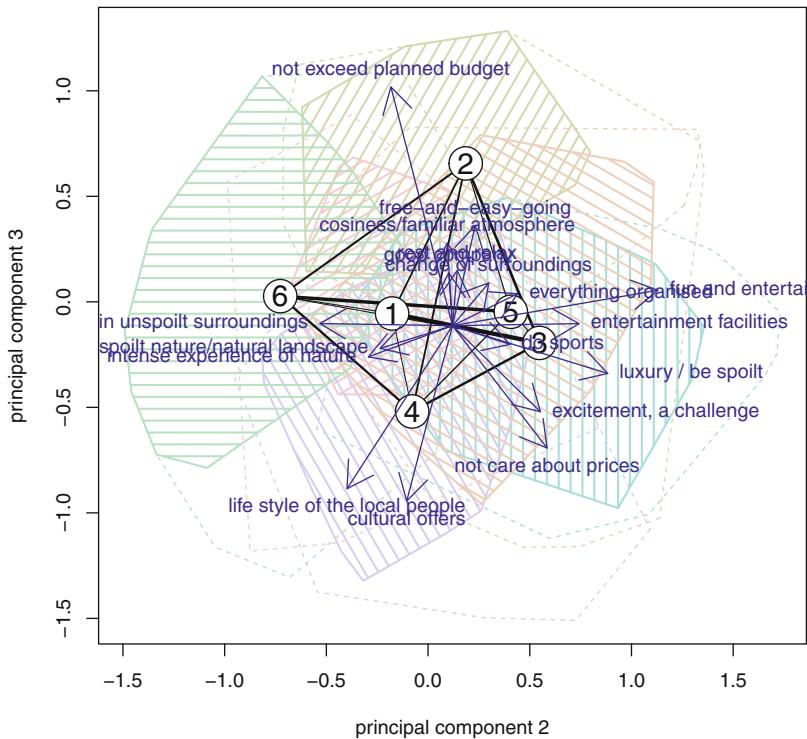


Fig. 8.6 Segment separation plot using principal components 2 and 3 for the Australian travel motives data set without observations

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.		<input type="checkbox"/>
Visualise segment profiles to learn about what makes each segment distinct.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Pass on the remaining segments to Step 7 for describing.		<input type="checkbox"/>

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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9

Step 7: Describing 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
```

Segment number	Male	Female	Gender
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 see 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

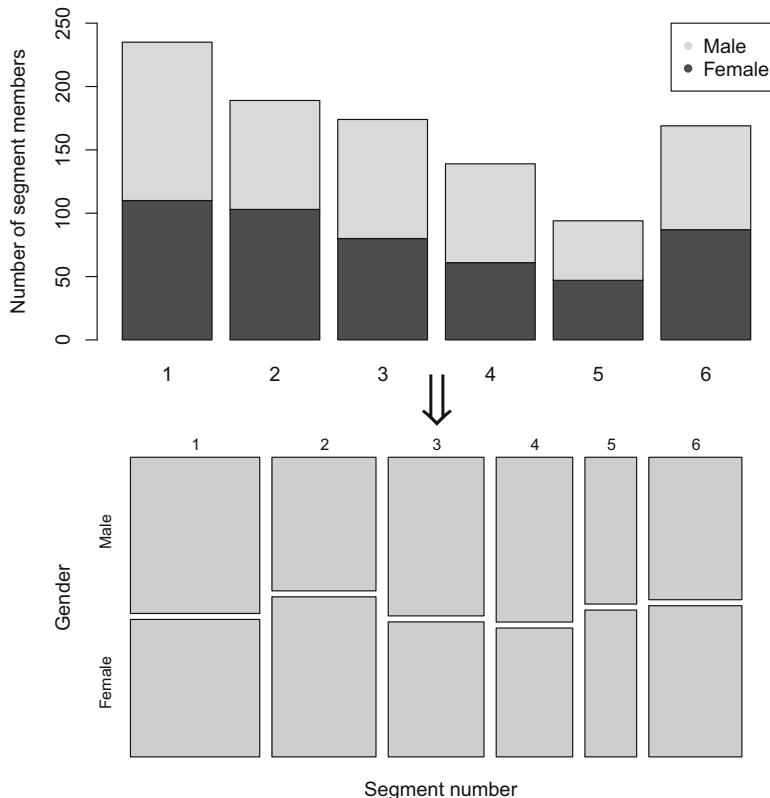


Fig. 9.1 Comparison of a stacked bar chart and a mosaic plot for the cross-tabulation of segment membership and gender for the Australian travel motives data set

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

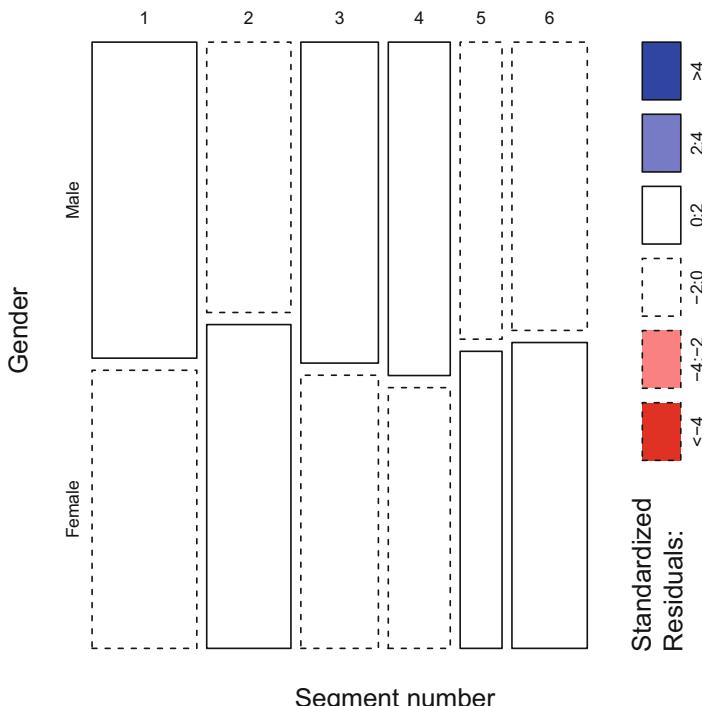


Fig. 9.2 Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set

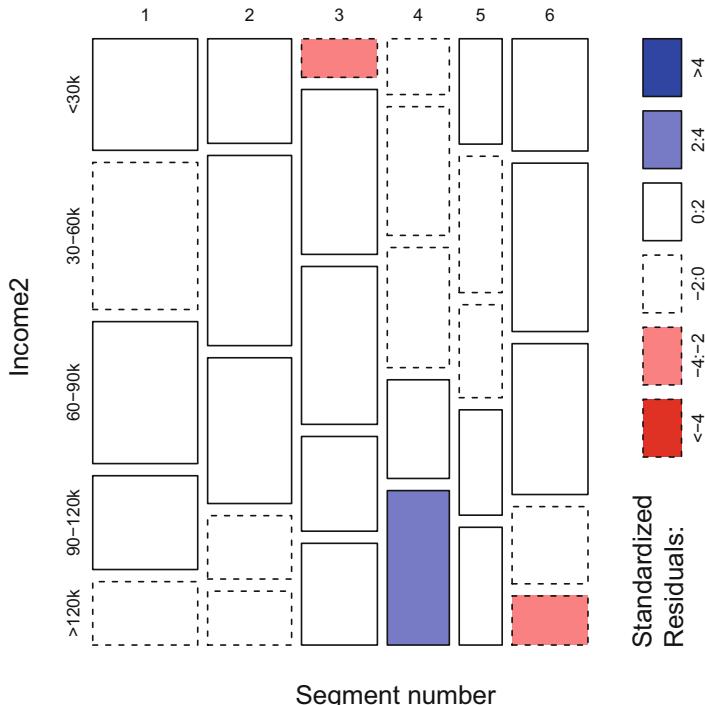


Fig. 9.3 Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set

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

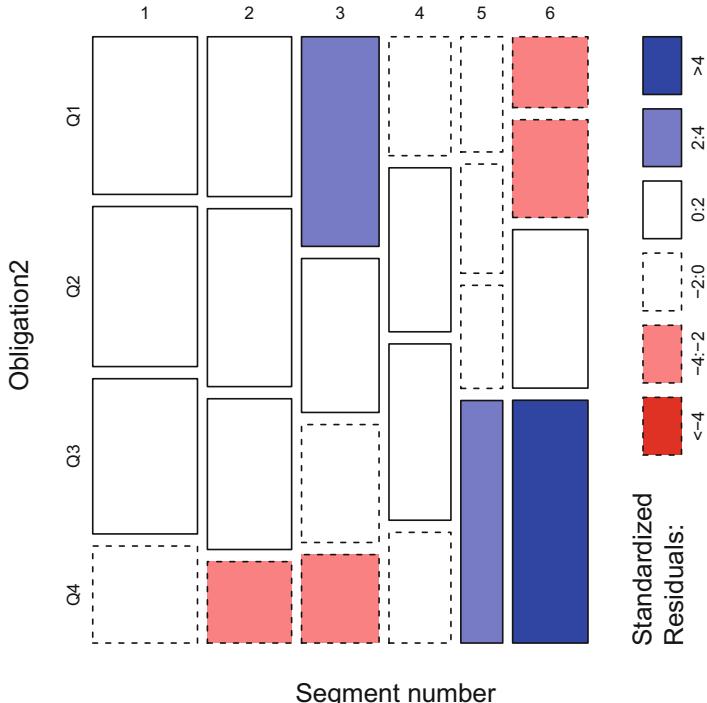


Fig. 9.4 Shaded mosaic plot for cross-tabulation of segment membership and moral obligation to protect the environment for the Australian travel motives data set

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 incorporate elements of statistical hypothesis testing. For example, we can make the width of the

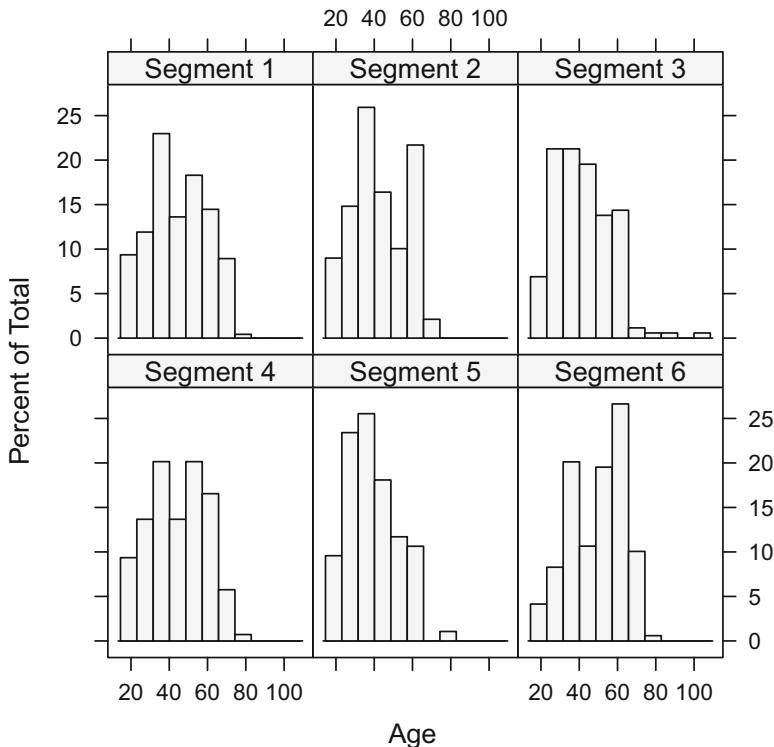


Fig. 9.5 Histograms of age by segment for the Australian travel motives data set

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

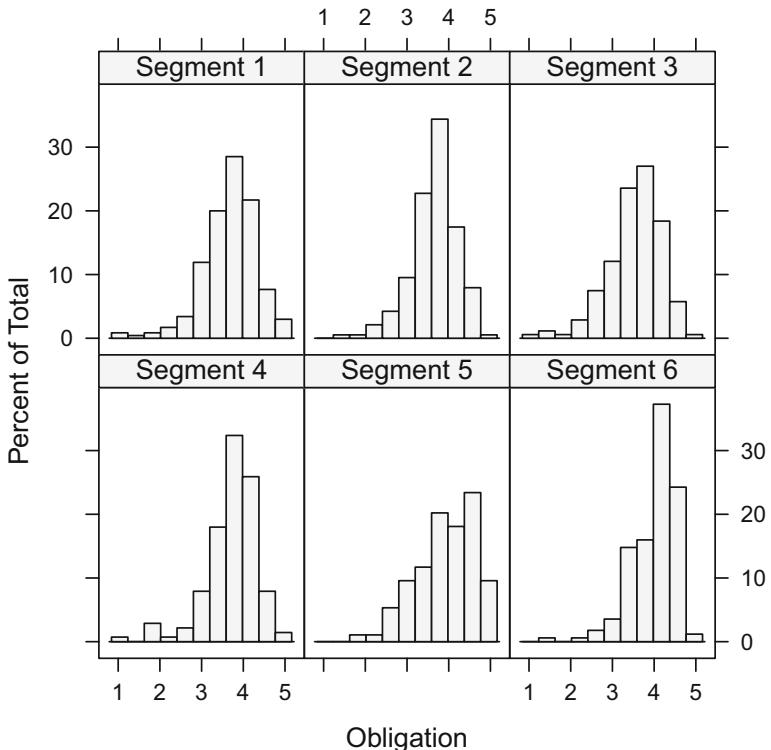


Fig. 9.6 Histograms of moral obligation to protect the environment by segment for the Australian travel motives data set

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

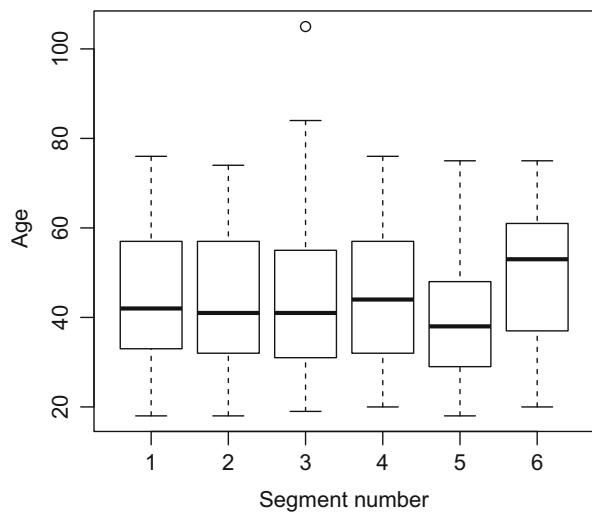
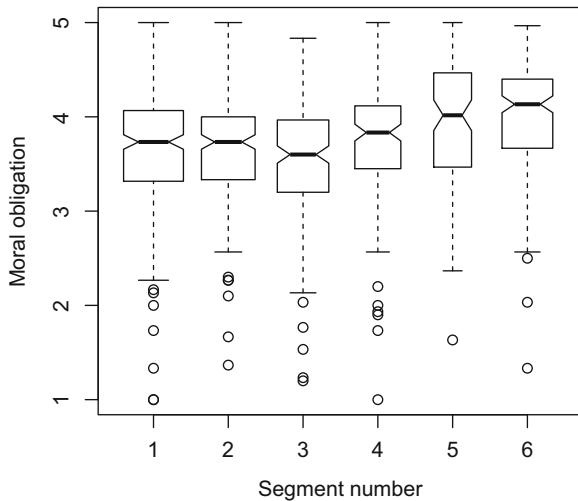


Fig. 9.8 Parallel box-and-whisker plot (with elements of statistical inference) of moral obligation to protect the environment by segment for the Australian travel motives data set



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.

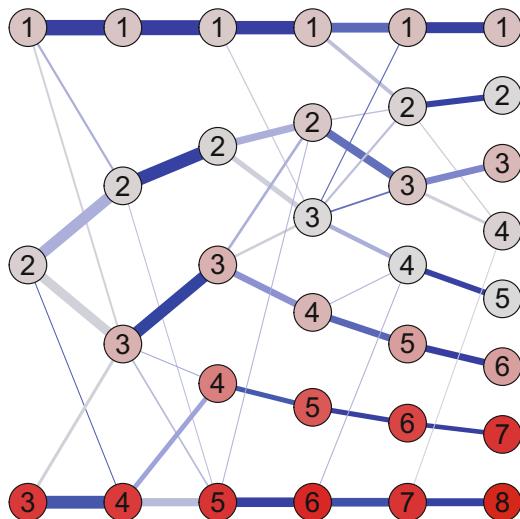
We can use a modified version of the segment level stability across solutions (SLS_A) plot to trace the value of a metric descriptor variable over a series of market segmentation solutions. The modification is that additional information contained in a metric descriptor variable is plotted using different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

```
R> slsplot(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

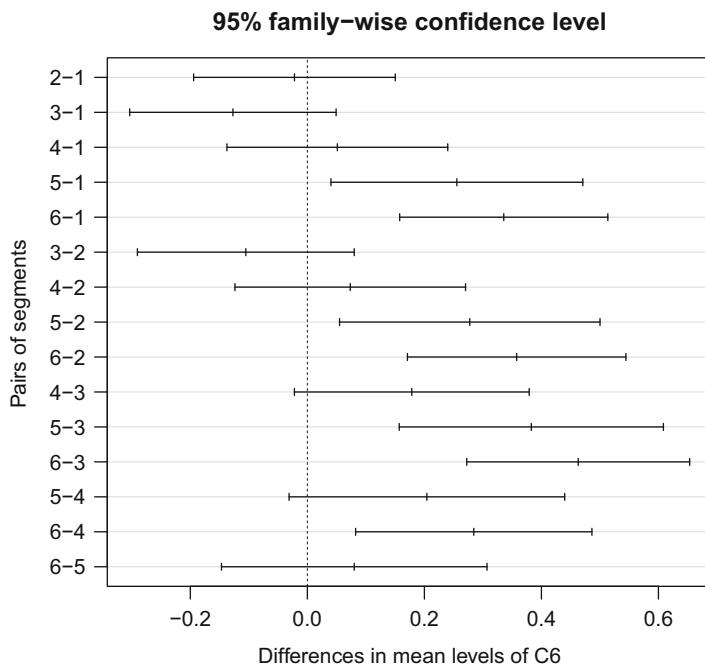


Fig. 9.10 Tukey's honest significant differences of moral obligation to behave environmentally friendly between the six segments for the Australian travel motives data set

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, \dots, x_p :

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

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

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 $f(x_1, \dots, x_p)$ and variance σ^2 . The relationship between the dependent variable y and the independent variables x_1, \dots, x_p is given by:

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

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, \dots, x_p is modelled by the linear function:

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

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:

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

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 \sim , and the two independent variables AGE and OBLIGATION2 on the right of \sim . 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, \dots, 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, \dots, 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

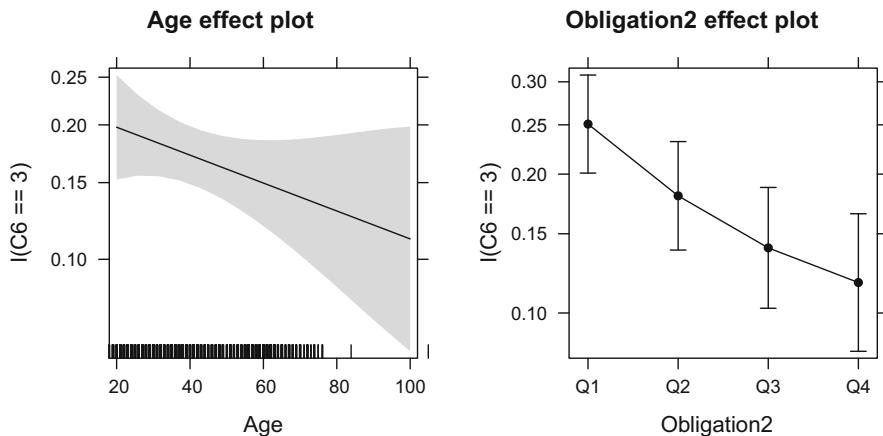


Fig. 9.11 Effect visualisation of age and moral obligation for predicting segment 3 using binary logistic regression for the Australian travel motives data set

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

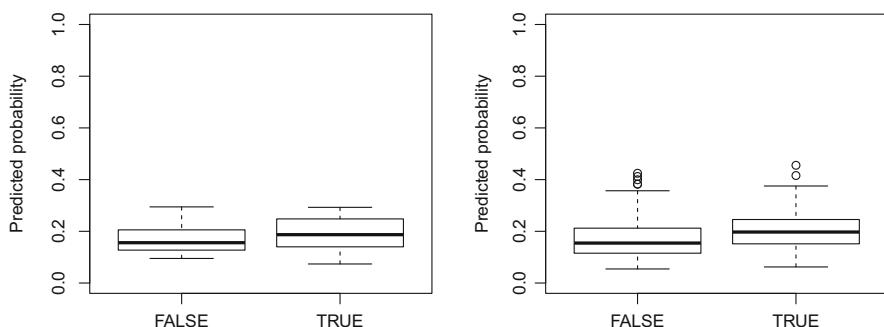


Fig. 9.12 Predicted probabilities of segment 3 membership for consumers not assigned to segment 3 (FALSE) and for consumers assigned to segment 3 (TRUE) for the Australian travel motives data set. The model containing age and moral obligation as independent variables is on the left; the model selected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

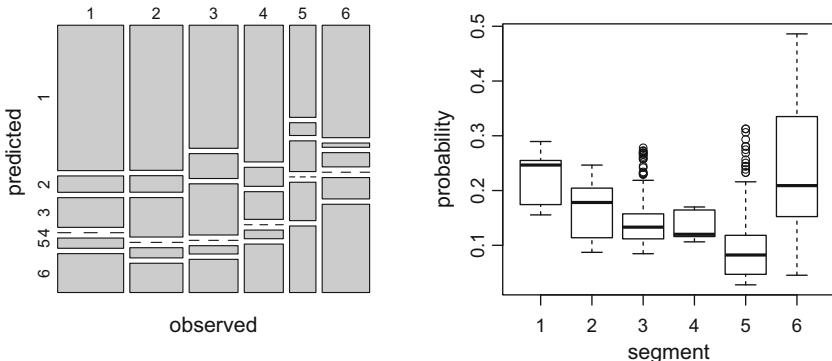


Fig. 9.13 Assessment of predictive performance of the multinomial logistic regression model including age and moral obligation as independent variables for the Australian travel motives data set. The mosaic plot of the cross-tabulation of observed and predicted segment memberships is on the left. The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

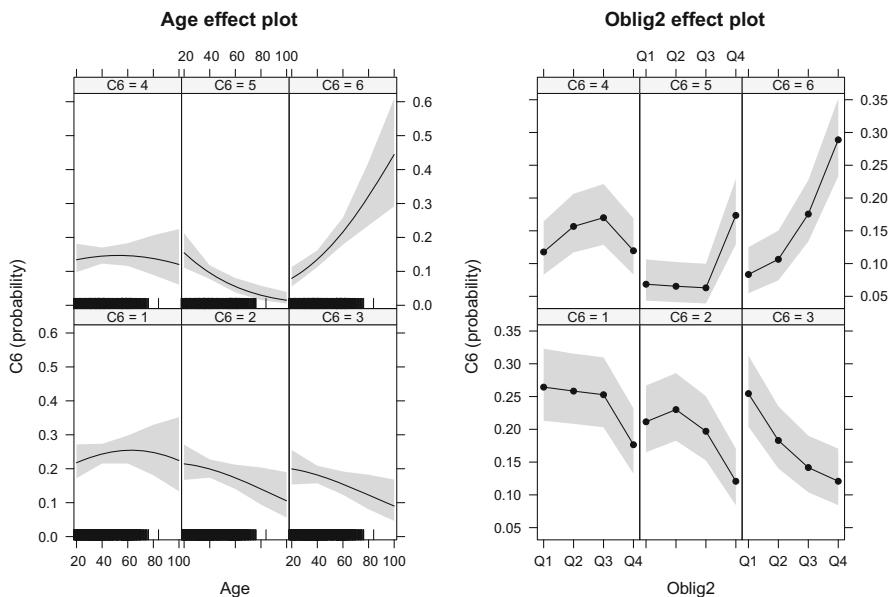


Fig. 9.14 Effect visualisation of age and moral obligation for predicting segment membership using multinomial logistic regression for the Australian travel motives data set

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 `cmtree()` 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* <

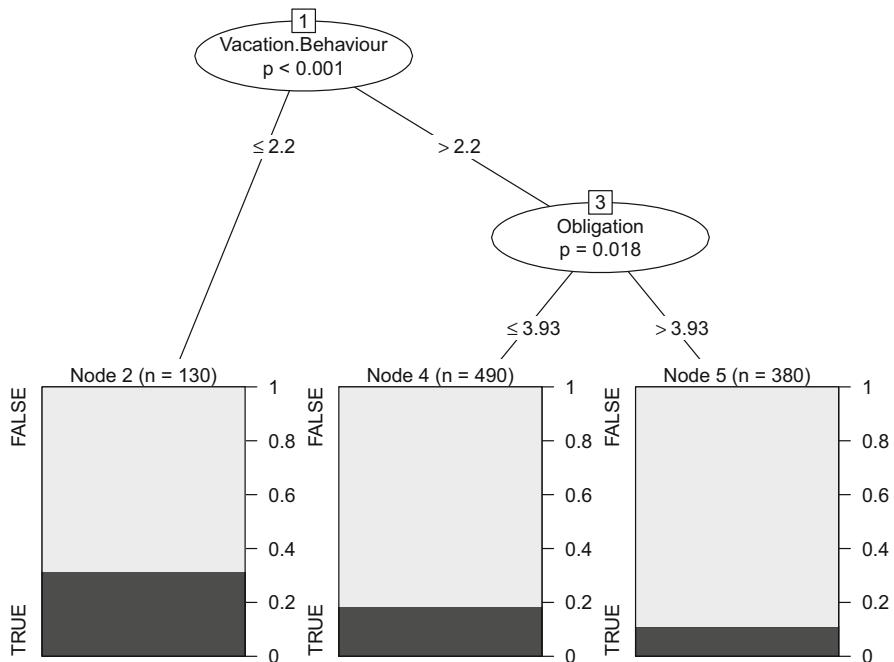


Fig. 9.15 Conditional inference tree using membership in segment 3 as dependent variable for the Australian travel motives data set

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
```

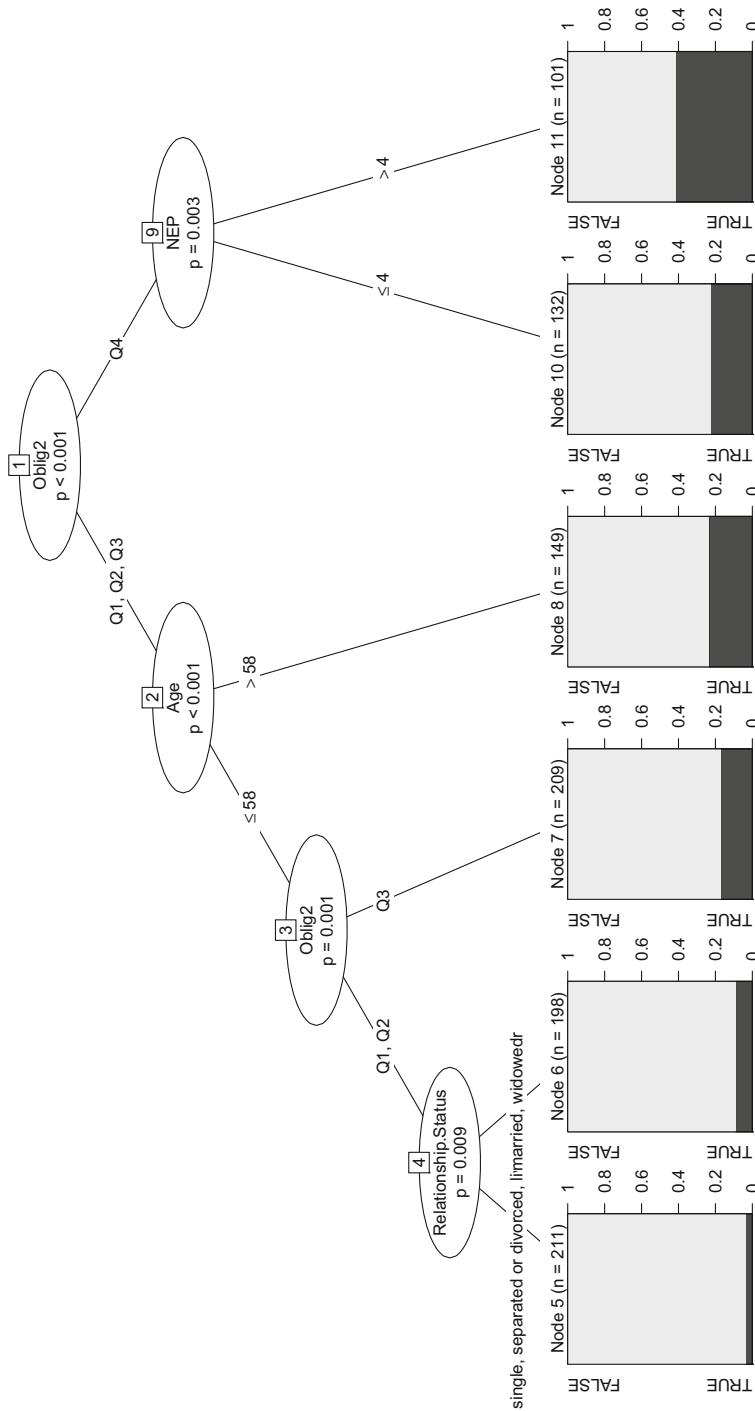


Fig. 9.16 Conditional inference tree using membership in segment 6 as dependent variable for the Australian travel motives data set

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.

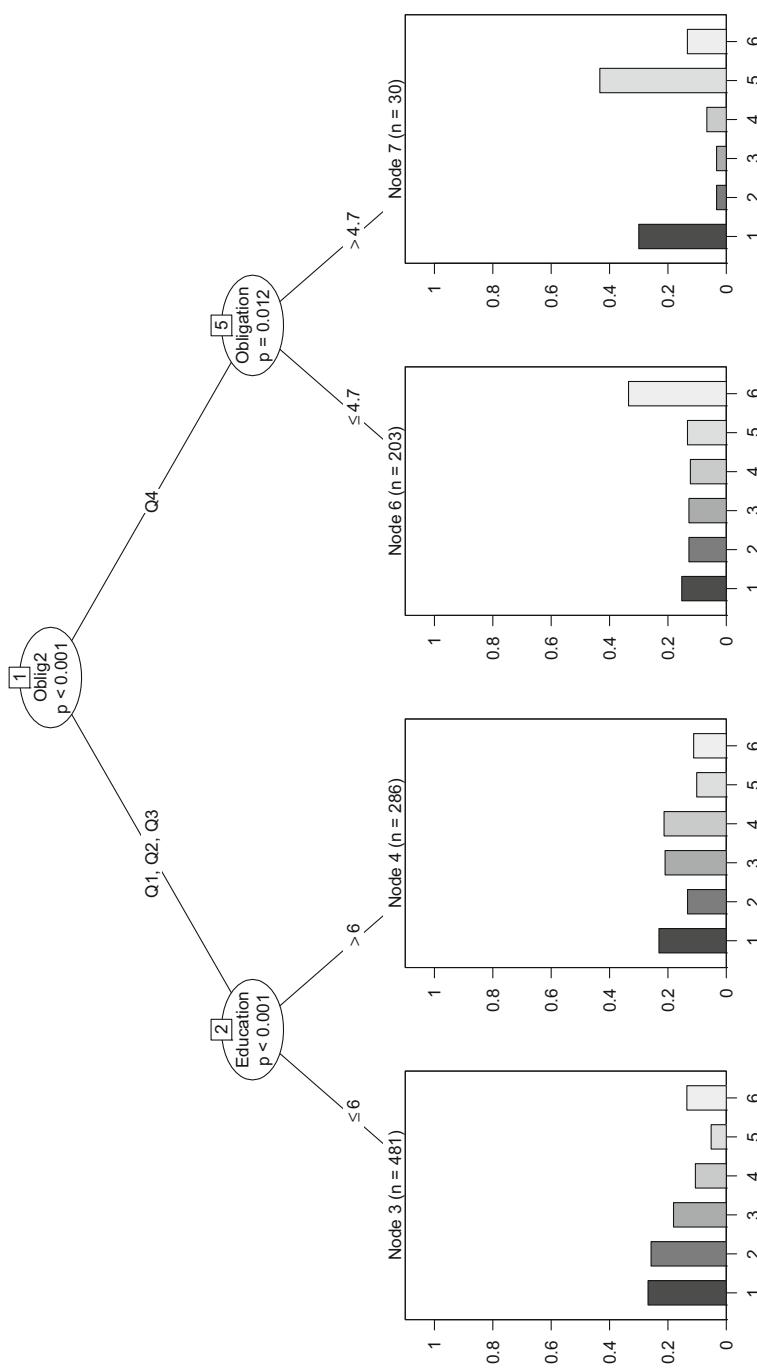


Fig. 9.17 Conditional inference tree using segment membership as dependent variable for the Australian travel motives data set

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.		<input type="checkbox"/>
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.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Test for statistical significance of descriptor variables.		<input type="checkbox"/>
If you used separate statistical tests for each descriptor variable, correct for multiple testing to avoid overestimating significance.		<input type="checkbox"/>
"Introduce" each market segment to the other team members to check how much you know about these market segments.		<input type="checkbox"/>
Ask if additional insight into some segments is required to develop a full picture of them.		<input type="checkbox"/>

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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 10

Step 8: Selecting the Target Segment(s)



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

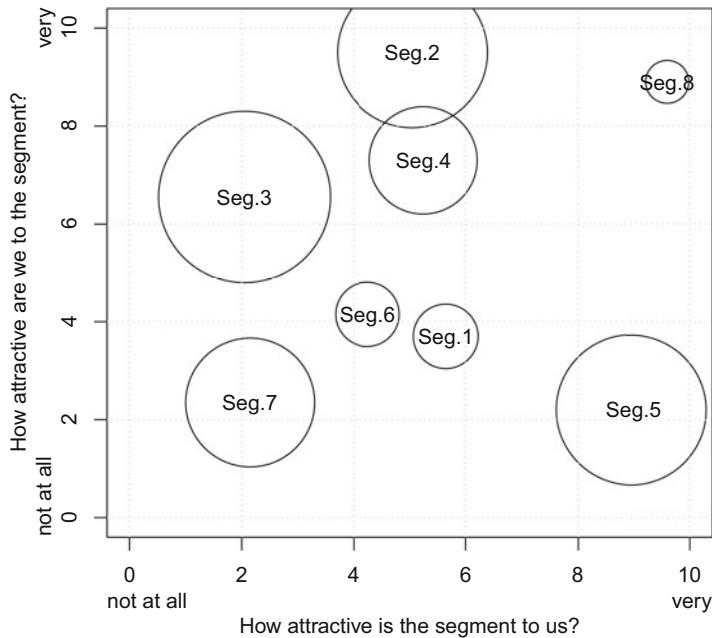


Fig. 10.1 Segment evaluation plot

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.		<input type="checkbox"/>
Determine which of the market segments profiled in Step 6 and described in Step 7 are being considered as potential target markets.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Discuss and agree on values for each market segment for each segment attractiveness criterion.		<input type="checkbox"/>
Discuss and agree on values for each relative organisational competitiveness criterion for each of the market segments.		<input type="checkbox"/>
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.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Plot the values onto a segment evaluation plot.		<input type="checkbox"/>
Make a preliminary selection.		<input type="checkbox"/>
If you intend to target more than one segment: make sure that the selected target segments are compatible with one another.		<input type="checkbox"/>
Present the selected segments to the advisory committee for discussion and (if required) reconsideration.		<input type="checkbox"/>

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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 11

Step 9: Customising the Marketing Mix



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 (NAs) 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)
```

	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:

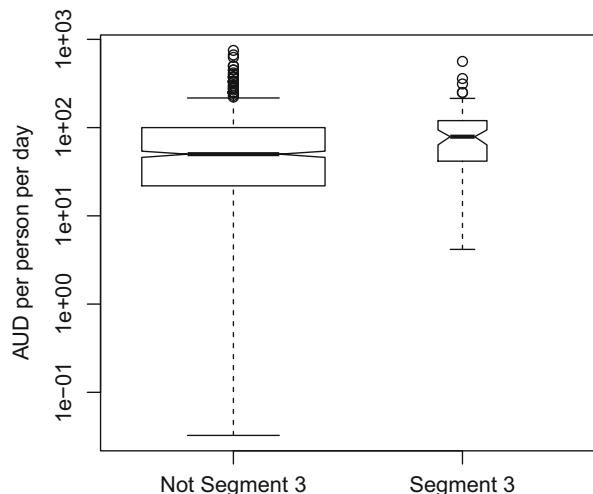
```
R> data("ausActivDesc", package = "MSA")
R> boxplot(spendpppd ~ cl12.3, data = ausActivDesc,
+   notch = TRUE, varwidth = TRUE, log = "Y",
+   ylab = "AUD per person per day")
```

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

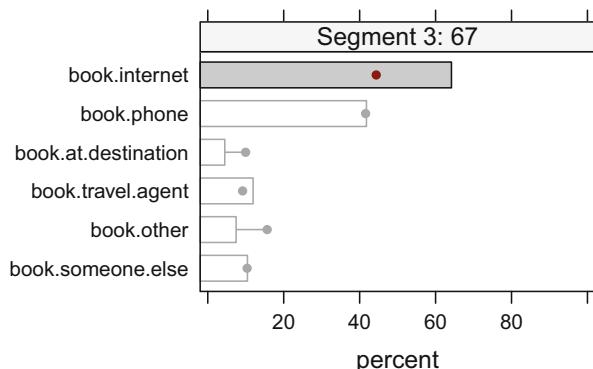


Figure 11.3 indicates that members of segment 3 differ from other tourists in terms of how they booked their hotel on their last domestic vacation: they book their hotel online much more frequently than the average tourist. This information has clear implications for the place dimension of the marketing 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 = "")
```

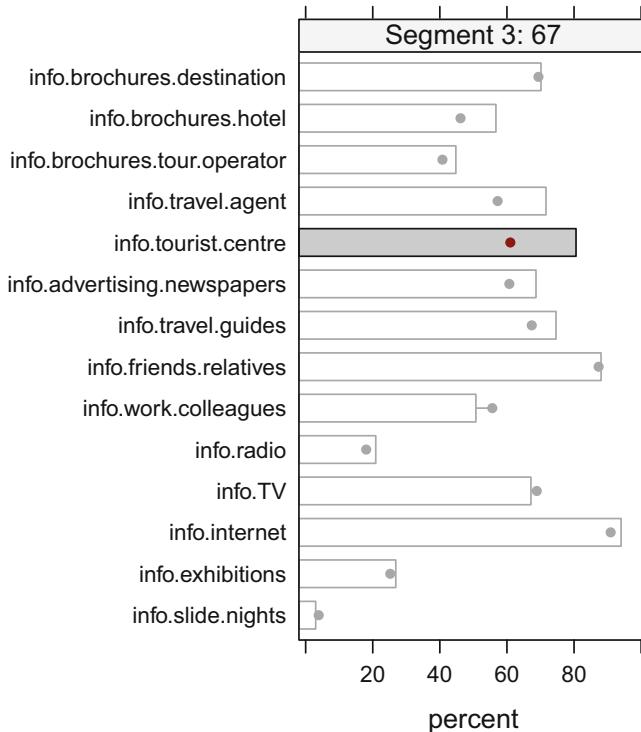


Fig. 11.4 Information sources used by segment 3 and by the average tourist.

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.

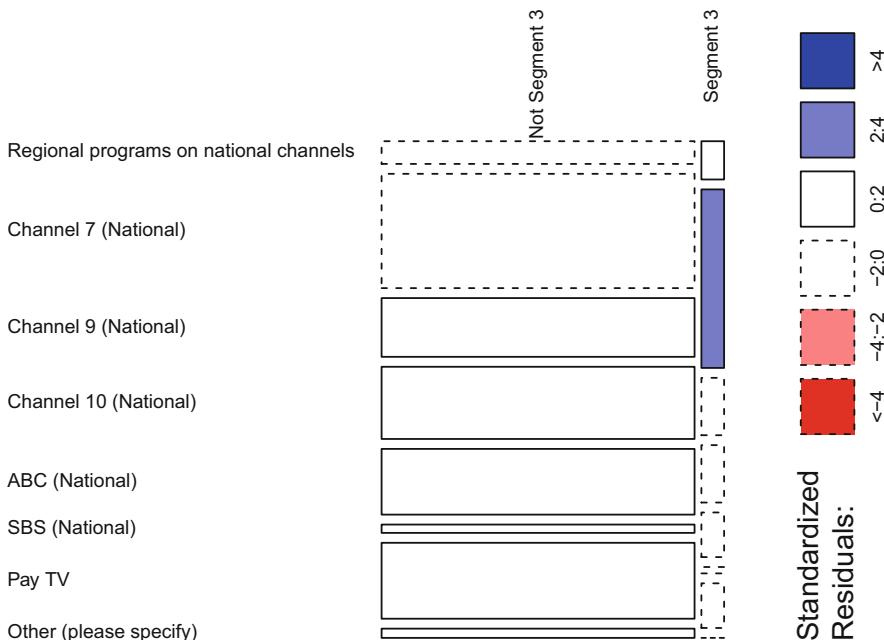


Fig. 11.5 TV station most frequently watched by segment 3 and all other tourists

11.6 Step 9 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		<input type="checkbox"/>
Study the profile and the detailed description of the target segment again carefully.		<input type="checkbox"/>
Determine how the product-related aspects need to be designed or modified to best cater for this target segment.		<input type="checkbox"/>
Determine how the price-related aspects need to be designed or modified to best cater for this target segment.		<input type="checkbox"/>
Determine how the place-related aspects need to be designed or modified to best cater for this target segment.		<input type="checkbox"/>
Determine how the promotion-related aspects need to be designed or modified to best cater for this target segment.		<input type="checkbox"/>
Review the marketing mix in its entirety.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Present an outline of the proposed marketing mix to the advisory committee for discussion and (if required) modification.		<input type="checkbox"/>

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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 12

Step 10: Evaluation 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 automated 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(v = 0, 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:

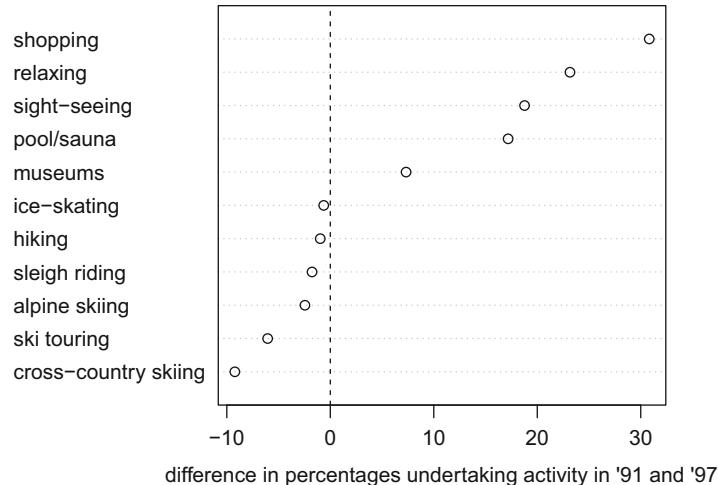


Fig. 12.1 Difference in the percentage of tourists engaging in 11 winter vacation activities during their vacation in Austria in 1991/92 and 1997/98

```
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:

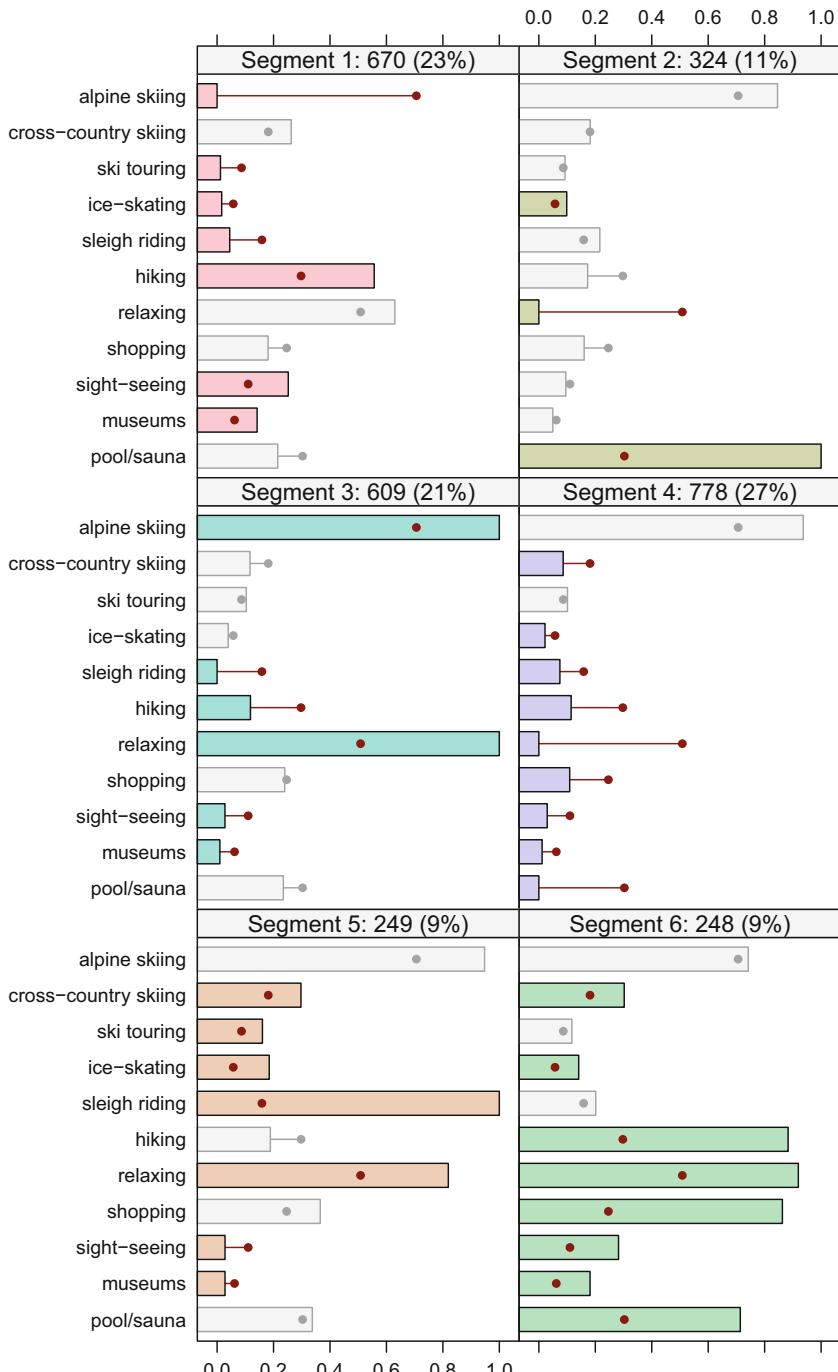


Fig. 12.2 Segment profile plot for the six-segment solution of winter vacation activities in 1991/92

```
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

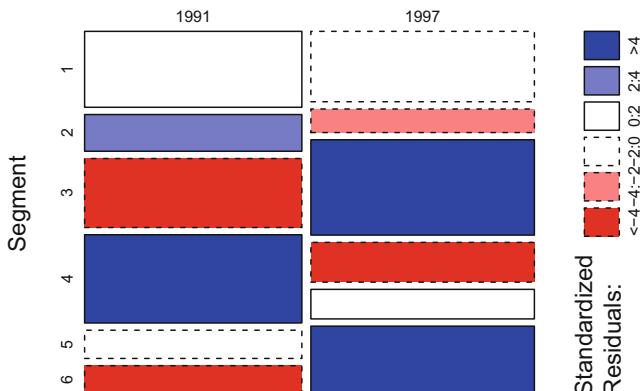


Fig. 12.3 Mosaic plot comparing segment sizes in 1991/92 and 1997/98 based on the segmentation solution for winter activities in 1991/92

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.

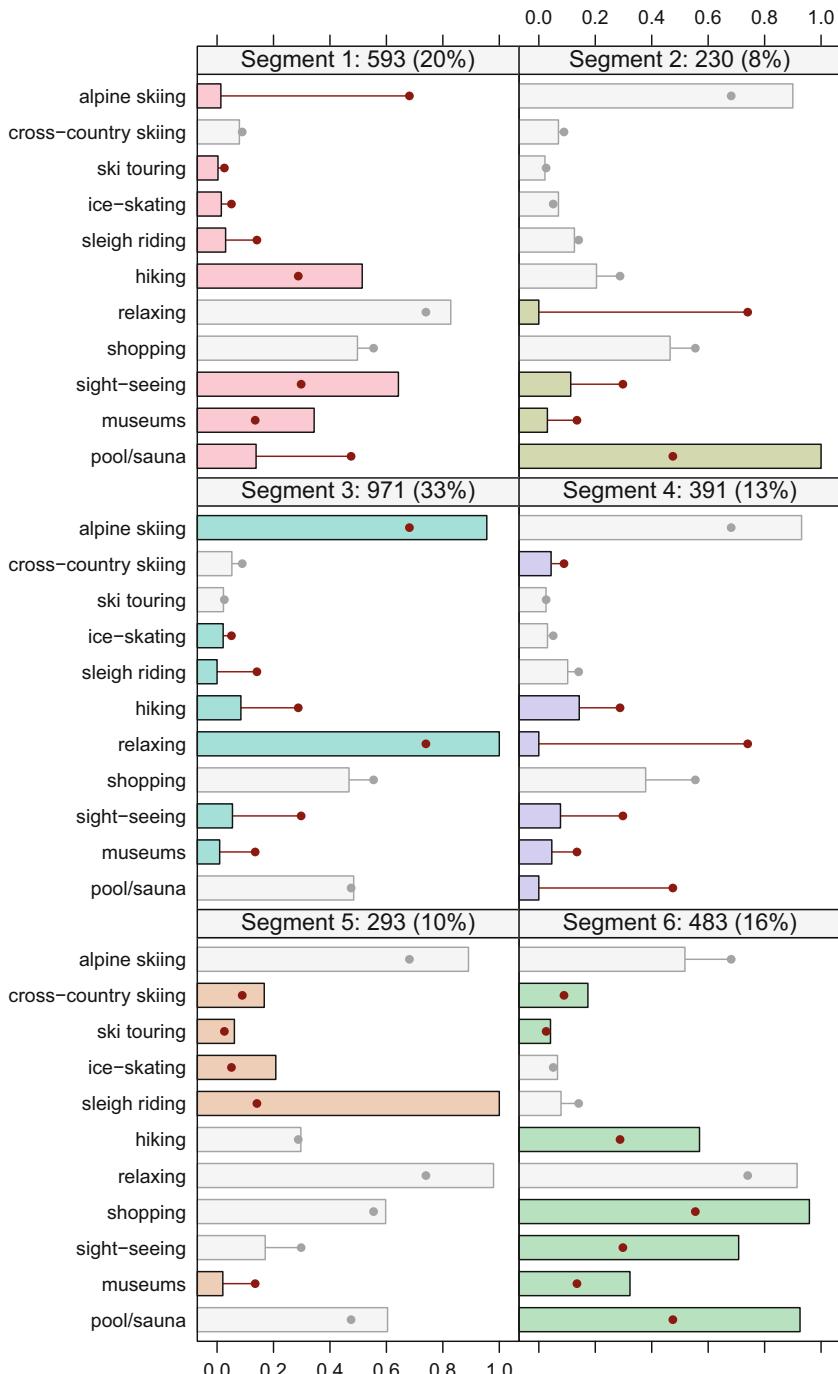


Fig. 12.4 Segment profile plot for the six-segment solution of winter vacation activities in 1997/98

12.5 Step 10 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		<input type="checkbox"/>
Determine which indicators of short-term and long-term success will be used to evaluate the market segmentation strategy.		<input type="checkbox"/>
Operationalise how segmentation success indicators will be measured and how frequently.		<input type="checkbox"/>
Determine who will be responsible for collecting data on these indicators.		<input type="checkbox"/>
Determine how often the segmentation team will re-convene to review the indicators.		<input type="checkbox"/>
Determine which indicators will be used to capture market dynamics.		<input type="checkbox"/>
Remind yourself of the baseline <i>global stability</i> to ensure that the source of instability is attributed to the correct cause.		<input type="checkbox"/>
Remind yourself of the baseline <i>segment level stability</i> to ensure that the source of instability is attributed to the correct cause.		<input type="checkbox"/>
Operationalise how market dynamics indicators will be measured and how frequently.		<input type="checkbox"/>
Determine who will be responsible for collecting data on market dynamics.		<input type="checkbox"/>
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.		<input type="checkbox"/>
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.		<input type="checkbox"/>
Run the indicators, measures of indicators, reviewing intervals and the draft adaptation checklist past the advisory committee for approval or (if necessary) modification.		<input type="checkbox"/>

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, Ntoutsis 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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



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, FATENING, 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   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 segmentation 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 0.2 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:

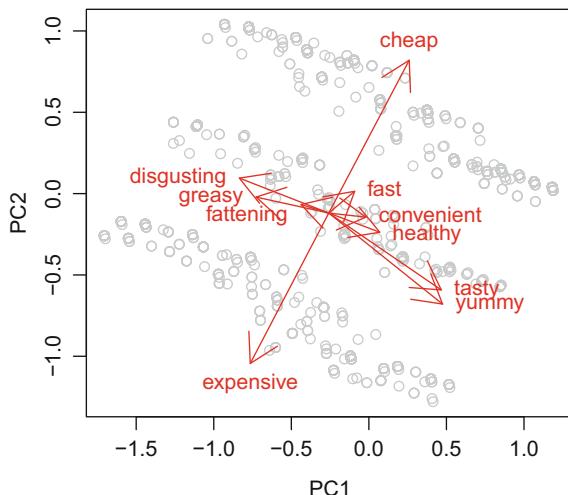
```

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.

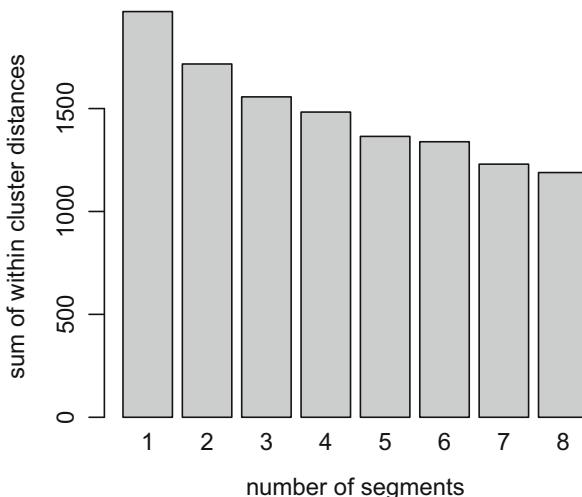


Fig. A.2 Scree plot for the fast food data set

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.

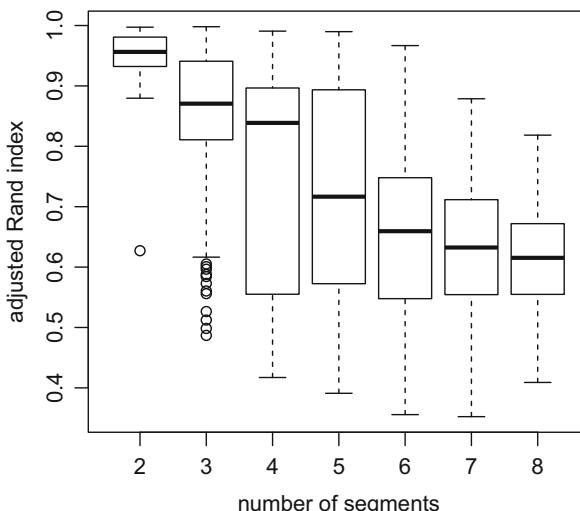


Fig. A.3 Global stability of k -means segmentation solutions for the fast food data set

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 `s1saplot` (`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

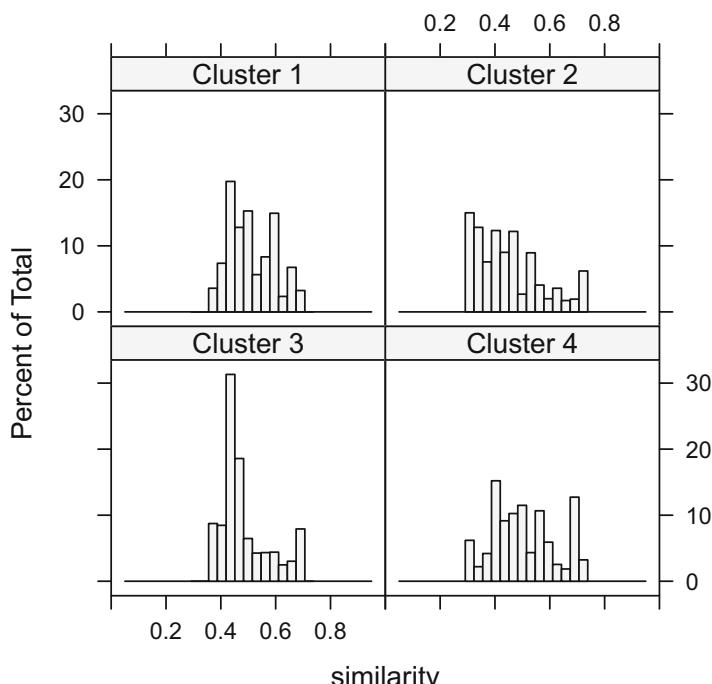


Fig. A.4 Gorge plot of the four-segment k -means solution for the fast food data set

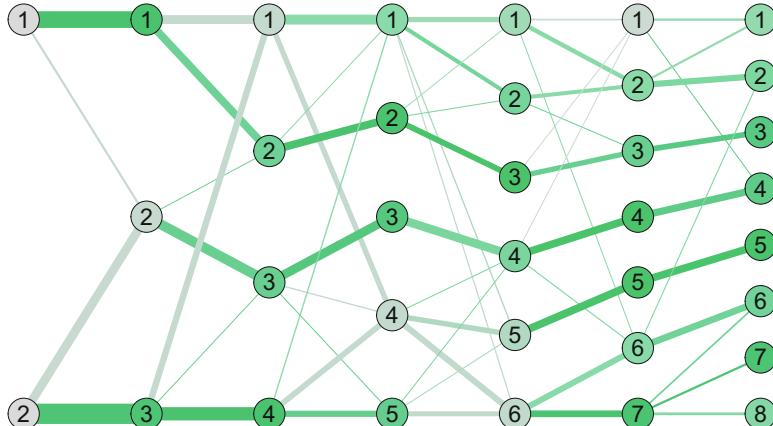


Fig. A.5 Segment level stability across solutions (SLS_A) plot from two to eight segments for the fast food data set

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).

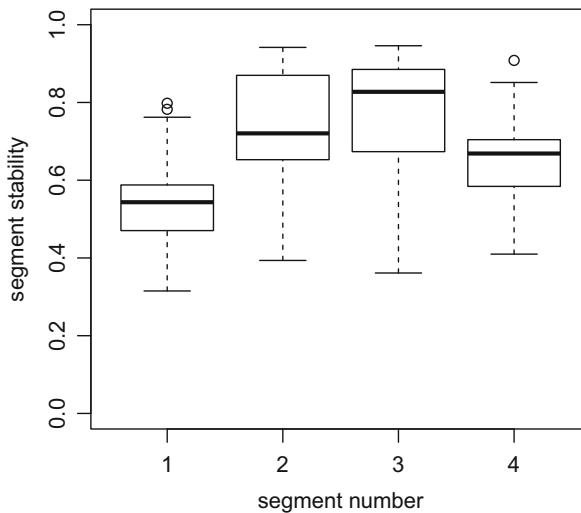


Fig. A.6 Segment level stability within solutions (SLS_W) plot for the fast food data set

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

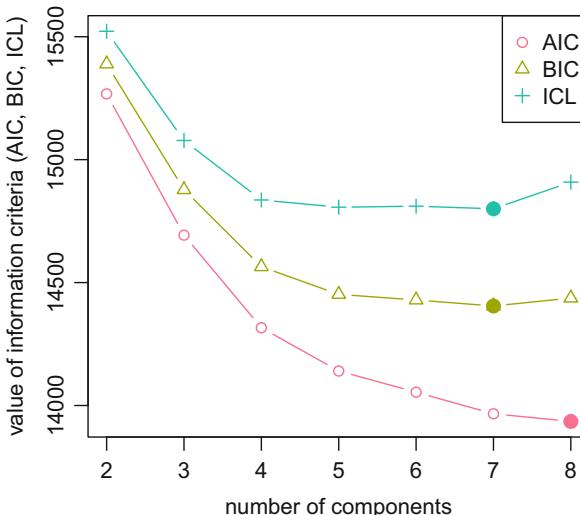


Fig. A.7 Information criteria for the mixture models of binary distributions with 2 to 8 components (segments) for the fast food data set

```

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

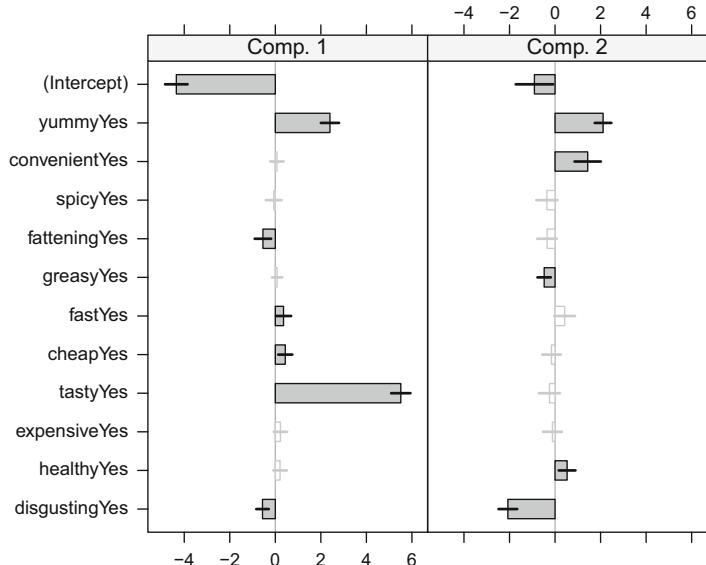


Fig. A.8 Regression coefficients of the two-segment mixture of linear regression models for the fast food data set

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.

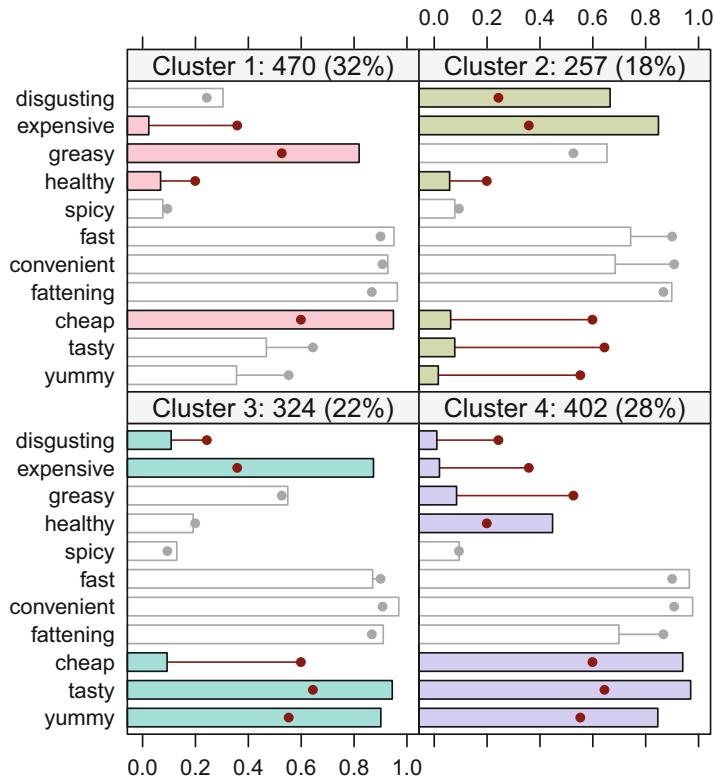


Fig. A.9 Segment profile plot for the four-segment solution for the fast food data set

```
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)
```

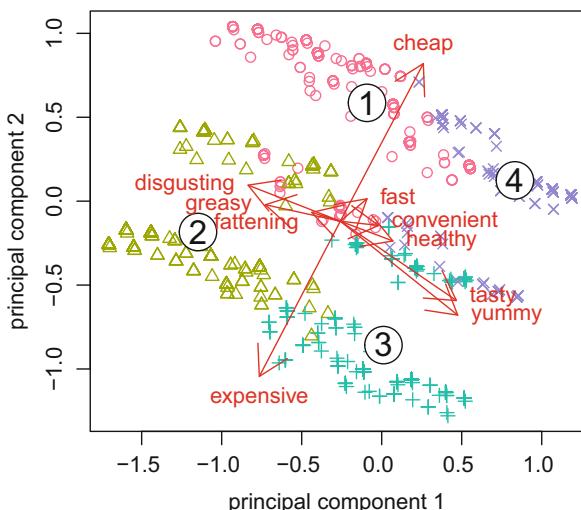


Fig. A.10 Segment separation plot using principal components 1 and 2 for the fast food data set

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

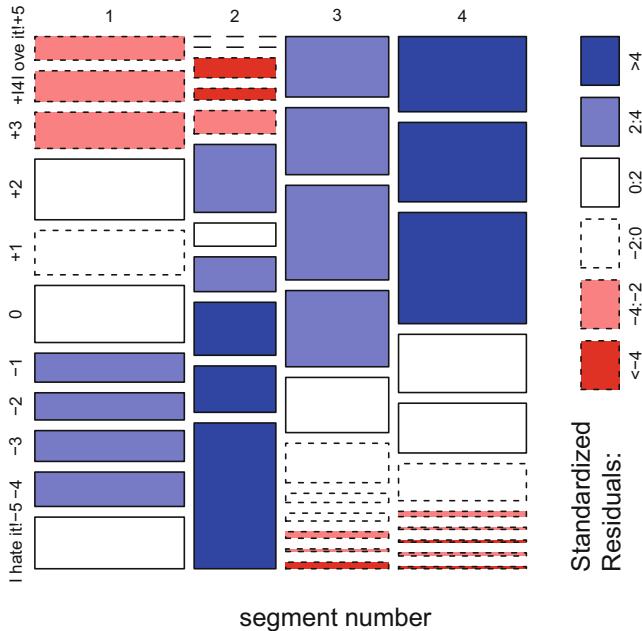


Fig. A.11 Shaded mosaic plot for cross-tabulation of segment membership and I LIKE IT for the fast food data set

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

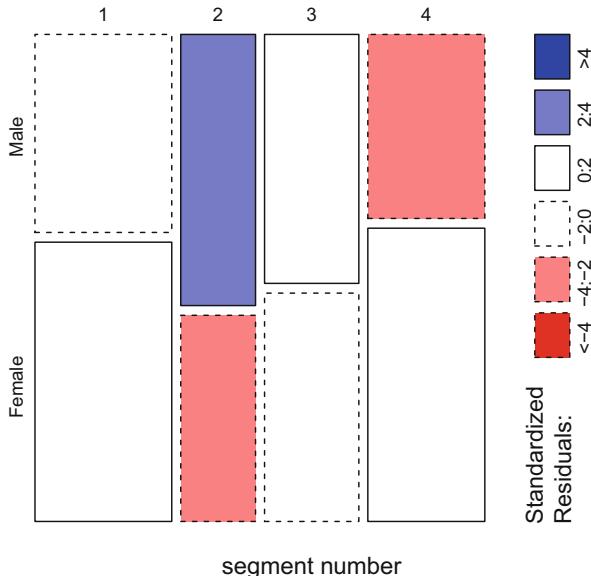


Fig. A.12 Shaded mosaic plot for cross-tabulation of segment membership and gender for the fast food data set

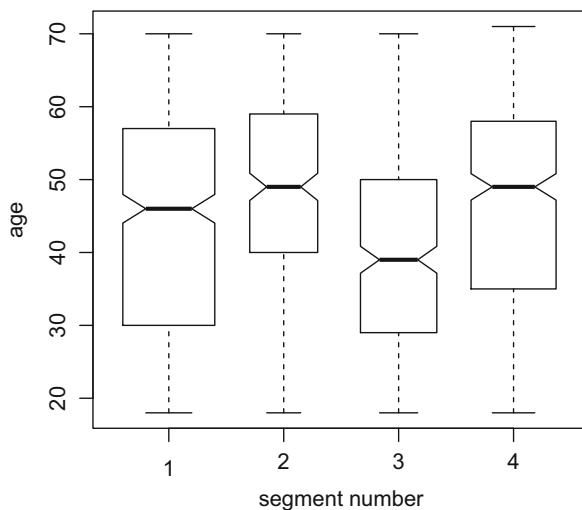


Fig. A.13 Parallel box-and-whisker plot of age by segment for the fast food data set

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.

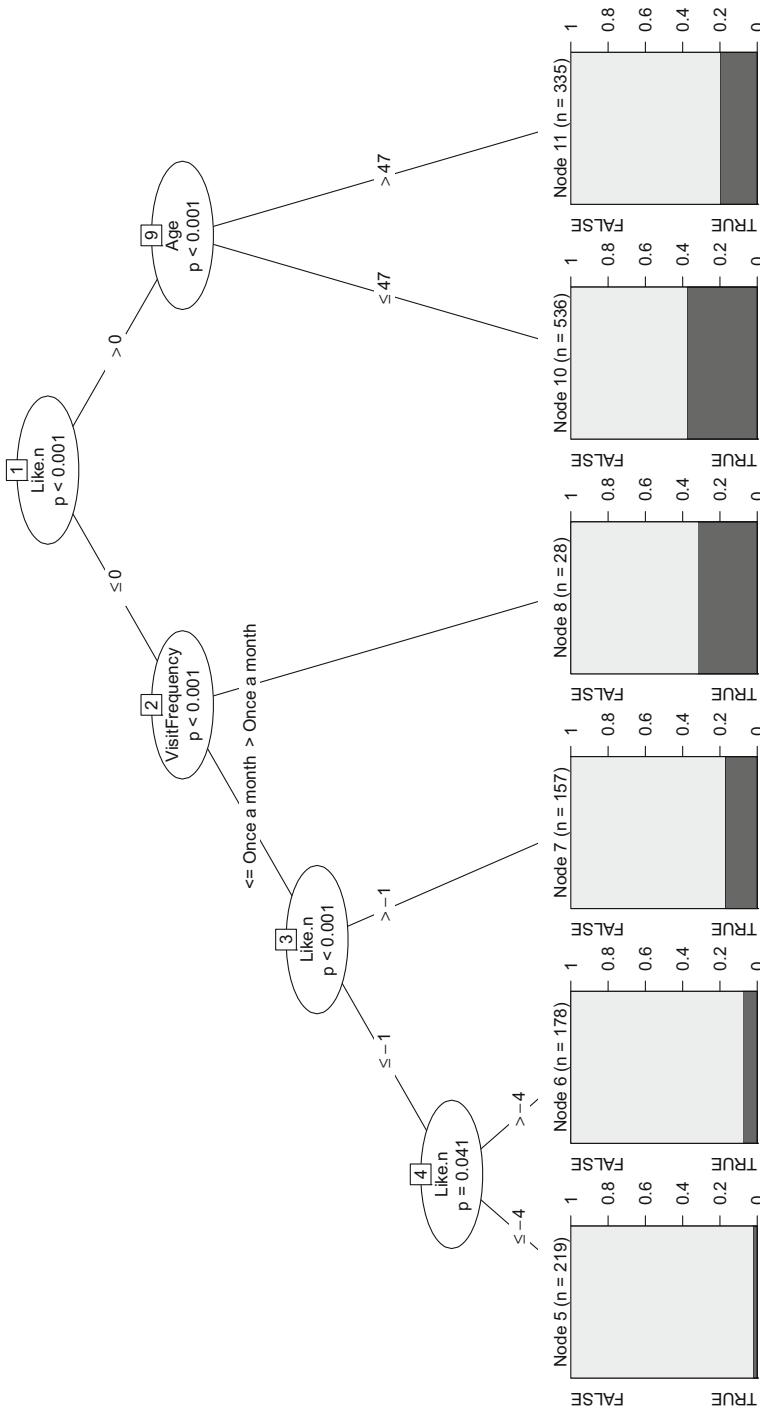


Fig. A.14 Conditional inference tree using segment 3 membership as dependent variable for the fast food data set

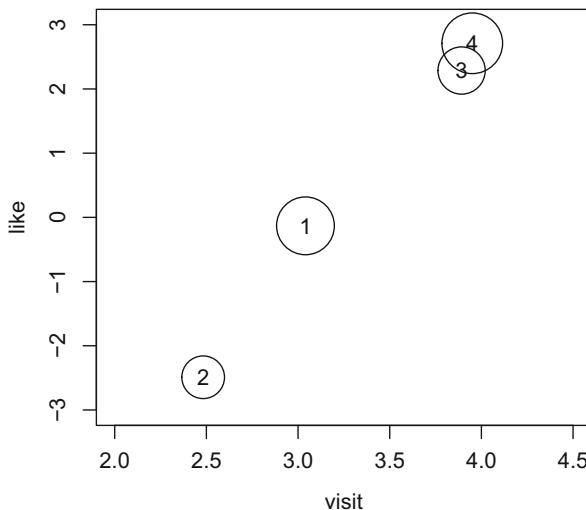


Fig. A.15 Example of a simple segment evaluation plot for the fast food data set

```
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,
+       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:

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

<code>clusterhulls()</code> :	Plot data with cluster hulls.
<code>decisionMatrix()</code> :	Segment evaluation plot.
<code>twoStep()</code> :	Infer segment membership for two-step clustering.
<code>vsbd()</code> :	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:

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

<code>flexmix()</code> :	Finite mixtures of distributions and regression models.
<code>stepFlexmix()</code> :	Repeated random initialisations of <code>flexmix()</code> .

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 L ^A T _E X (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/zoo/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.

Cluster: The term cluster is used in distance-based segment extraction methods to describe groups of consumers or market segments.

Clustering: Clustering aims at grouping consumers in a way that consumers in the same segment (called a cluster) are more similar to each other than those in other segments (clusters). Clustering is also referred to as *unsupervised learning* in machine learning. Statistical clustering algorithms can be used to extract market segments.

Component: The term components is used in model-based segment extraction methods to refer to groups of consumers or market segments.

Constructed market segments: Groups of consumers (market segments) artificially created from unstructured data. They do not re-occur across repeated calculations.

Data cleaning: Irrespective of the nature of empirical data, it is necessary to check if it contains any errors and correct those before extracting segments. Typical errors in survey data include missing values or systematic biases.

Data-driven segmentation: See a posteriori market segmentation.

Data structure analysis: Exploratory analysis of the segmentation variables used to extract market segments. Stability-based data structure analysis provides insights into whether market segments are naturally existing (permitting natural segmentation to be conducted), can be extracted in a stable way (requiring reproducible market segmentation to be conducted), or need to be artificially created (requiring constructive market segmentation to be conducted). Stability-based data structure analysis also offers guidance on the number of market segments to extract.

Dendrogram: A dendrogram visualises the solution of hierarchical clustering, and depicts how observations are merged step-by-step in the sequence of nested partitions. The height represents the distance between the two sets of observations being merged. The dendrogram has been proposed as a visual aid to select a suitable number of clusters. However, in data without natural clusters the identification of a suitable number of segments might be difficult and ambiguous.

Descriptor variable: Descriptor variables are *not* used to extract segments. Rather, they are used after segment extraction to develop a detailed description of market segments. Detailed descriptions are essential to enable an organisation to select one or more target segments, and develop a marketing mix that is customised specifically to one or more target segments.

Exploratory data analysis: Irrespective of the algorithm used to extract market segments, one single correct segmentation solution does not exist. Rather, many different segmentation solutions can result. Randomly choosing one of them is risky because the chosen solution may not be very good. The best way to avoid choosing a bad solution, is to invest time into exploratory data analysis. Exploratory data analysis provides glimpses of the data structure from different perspectives,

thus guiding the data analyst towards a managerially useful market segmentation solution. A range of tools is available to explore data, including tables and graphical visualisations.

Factor-cluster analysis: Factor-cluster analysis is sometimes used in an attempt to reduce the number of segmentation variables in empirical data sets. It consists of two steps: first the original segmentation variables are factor analysed based on principal components analysis. Principal components with eigenvalues equal or larger than one are then selected and suitably rotated to obtain the factor scores. Factor scores are then used as segmentation variables for segment extraction. Because only a small number of factors are used, a substantial amount of information contained in the original consumer data might be lost. Factor-cluster analysis is therefore not recommended, and has been empirically proven to not outperform segment extraction using the original variables. If the number of original segmentation variables is too high, a range of other options are available to the data analyst to select a subset of variables, including using algorithms which simultaneously extract segments and select variables, such as biclustering or the variable selection procedure for clustering binary data (VSBD).

Geographic segmentation: Geographic segmentation is the result of using geographic information as segmentation variable(s). Examples include postcodes, country of origin (frequently used in tourism market segmentation) or travel patterns recorded using GPS tracking.

Global stability: Global stability is a measure of the replicability of an overall market segmentation solution across repeated calculations. Very high levels of global stability point to the existence of natural market segments. Very low levels of global stability point to the need for constructive market segmentation. Global stability is visualised using a global stability boxplot.

Hierarchical clustering: Distance-based method for the extraction of market segments. Hierarchical methods either start with the complete data set and split it up until each consumer represents their own market segment; or they start with each consumer being a market segment and merge the most similar consumers step-by-step until all consumers are united in one large segment. Hierarchical methods provide nested partitions as output which are visualised in a so-called *dendrogram*. The dendrogram can guide the selection of number of market segments to extract in cases where data sets are well structured.

***k*-means clustering:** *k*-means clustering is the most commonly used distance-based partitioning clustering algorithm. Using random consumers from the data sets as starting points, the standard *k*-means clustering algorithm iteratively assigns all consumers to the cluster centres (centroids, segment representatives), and adjusts the location of the cluster centres until cluster centres do not change anymore. Standard *k*-means clustering uses the squared Euclidean distance. Generalisations using other distances are also referred to as *k*-centroid clustering.

Knock-out criteria: Criteria a market segment must comply with to qualify as a target segment, including homogeneity (similarity of members to one another), distinctness (difference of members of one segment to members of another segment), sufficient size to be commercially viable, match with organisational strengths, identifiability (recognisability of segments members), and reachability.

Marker variable: Marker variables are subsets of segmentation variables that discriminate particularly well between market segments. They serve as key characteristics in the profiling of market segments.

Market segment: A group of similar consumers. A market segment contains a subset of consumers who are similar to one another with respect to the segmentation criterion, for example, a characteristic that is relevant to the purchase of a certain product. Optimally, members of different market segments are very different from one another.

Market segmentation analysis: The process of grouping consumers into naturally existing or artificially created segments of consumers who share similar product needs.

Masking variable: Masking variables – also referred to as noisy variables – are segmentation variables that do not help the segmentation algorithm to extract market segments. Rather, they blur the true structure of the data. By not contributing any information relevant to the segmentation analysis, masking variables increase the number of segmentation variables and, in so doing, make the segment extraction task unnecessarily difficult.

Mosaic plot: The mosaic plot visualises the joint distribution of categorical (nominal or ordinal) variables based on their cross-tabulation. The mosaic plot allows to compare the distribution of a nominal or ordinal variable across market segments. A shaded mosaic plot colours the cells according to the standardised residuals obtained from comparing the observed cell size with the expected cell size if the variables are not associated, and thus allows easy identification of differences in the distributions across market segments. It is a useful tool for the description of market segments using nominal descriptor variables (such as gender, country of origin, preferred brand), or ordinal variables (such as age groups, the agreement with a range of statements).

Natural segmentation: The concept of natural segmentation can be used when natural market segments exist in the data. Such natural market segments are distinct and well-separated. Being able to extract them repeatedly across multiple independent calculations is an indicator of their existence. Natural segmentation is the textbook case of market segmentation, but natural segments rarely occur in consumer data.

Natural market segments: Groups of similar consumers existing naturally in the market. Such market segments rarely exist in consumer data. High stability of segmentation solutions when repeated is an indicator of the existence of natural market segments.

Noisy variable: See masking variable.

Partitioning clustering: Distance-based method for the extraction of market segments. Partitioning methods aim at finding the optimal partition with respect to some criterion and thus require the number of market segments to be specified in advance.

Post hoc market segmentation: See a posteriori market segmentation.

Principal components analysis: Principal components analysis (PCA) finds principal components in data sets containing multiple variables. These principal components differ from the original variables in two ways: they are uncorrelated and they are ordered by information contained (the first principal component contains the most information about the data). As long as the full set of principal components is retained, the components offer a different angle of looking at the data. If, however, only a small number of principal components are used as segmentation variables – which typically occurs when data analysts are faced with too many original variables as segmentation variables – a substantial amount of information collected from consumers is typically lost. It is therefore preferable to use the original variables for segment extraction. If the number of segmentation variables is too high, principal components (or expert assessment) can guide the selection of a subset of available variables to be used for segment extraction, or algorithms like biclustering can be used.

Psychographic segmentation: Psychographic segmentation is the result of using psychological traits of consumers or their beliefs or values as segmentation criterion. Examples include travel motives, benefits sought when purchasing a product, personality traits, and risk aversion.

Rand index: The Rand index measures how similar two market segmentation solutions are. It takes values between 1 and 0, where 1 indicates that the two segmentation solutions are identical.

Recursive partitioning: Recursive partitioning can be used as a regression or classification algorithm; it generates a decision tree also referred to as classification or regression tree. The algorithm aims at identifying homogeneous subsamples with respect to the outcome variable by stepwise splitting of the sample into subsamples based on the independent variables. The trees obtained using recursive partitioning are easy to interpret and allow for convenient visualisation. The disadvantage of recursive partitioning is that the trees are unstable and their predictive performance is often outperformed by other regression or classification algorithms.

Regression: Regression is the statistical problem of learning a prediction algorithm where the predicted variable is a metric variable. Regression is also referred to as *supervised learning* in machine learning. Linear regression or recursive partitioning algorithms are examples for regression algorithms. Regression is used as segment-specific model in model-based clustering using a mixture of regression models.

Reproducible segmentation: The concept of reproducible market segmentation is used when natural, distinct, and well-separated market segments do not exist, yet the segmentation variables underlying the analysis are not entirely unstructured. The existing (unknown) structure of the data can be harvested to extract relatively stable segments. Stable segments are segments which re-emerge in similar form across repeated calculations. In reproducible market segmentation, it is essential to conduct a thorough data structure analysis to gain as much insight as possible about the data before extracting segments. Reproducible market segmentation is the most common case when extracting segments from consumer data (Ernst and Dolnicar 2018).

Sample size: The number of people whose information is contained in the data set which forms the basis of the market segmentation analysis. Sample size requirements for market segmentation analysis increase with the number of segmentation variables used. As a rule of thumb, the sample size should be at least 100 times the number of variables (Dolnicar et al. 2016).

Segment attractiveness criteria: Once market segments have been extracted, they have to be assessed in terms of their attractiveness as target markets for an organisation. Segment attractiveness criteria have to be selected and weighted by the users of the market segmentation solution (the managers considering to pursue a market segmentation strategy). Optimally, this occurs before data is collected. After segments have been extracted from the data, segment attractiveness criteria are used to develop a segment evaluation plot that assists users in choosing one or a small number of target segments.

Segment evaluation: After market segments have been extracted from consumer data, profiled and described, users – typically managers in the organisation considering to adopt a segmentation strategy – have to select one or a small number of market segments for targeting. To do this, market segments have to be evaluated. This is achieved by agreeing on desirable segment characteristics, assigning weights to them, and using the summated values to create a segment evaluation plot. The plot guides the discussion of users as they select one or a small number of target segments.

Segment evaluation plot: The segment evaluation plot visualises the decision matrix assisting users of market segmentation solutions (managers) to compare market segments before selecting one or a small number of target segments. The segment evaluation plot depicts the attractiveness of each segment to the organisation on one axis, and the attractiveness of the organisation's product or service to each of the segments on the other axis. The values for both of these axes result from both the segment extraction stage as well as managers' evaluation of which segment attractiveness criteria matter most to them. The bubble size of the segment evaluation plot can be used to visualise another key feature of each segment, such as an indicator of their profitability.

Segmentation criterion: This is a general term for the nature of the segmentation variables chosen; it describes the construct used as the basis for grouping consumers. Travel motives or expenditure patterns, for example, are segmentation criteria.

Segmentation variable: Segmentation variables are used to extract segments. Market segments can be based on one single segmentation variable (such as age or gender) or on many segmentation variables (such as a set of travel motives, or patterns of expenditure for a range of different products). The approach using one single (or a small number of) segmentation variable(s) inducing a segmentation solution which is known in advance is referred to as commonsense segmentation. The approach using many segmentation variables where segments need to be extracted is referred to as data-driven segmentation.

Segment level stability across solutions (SLS_A): Segment level stability across solutions (SLS_A) indicates how stable one market segment is across repeated calculations of market segmentation solutions containing different numbers of segments. It can best be understood as the stubbornness with which a market segment reappears across repeated calculations with different numbers of segments. Segment level stability across solutions (SLS_A) is visualised using a segment level stability across solutions (SLS_A) plot.

Segment level stability within solutions (SLS_W): Segment level stability within solutions (SLS_W) indicates how stable a market segment is across repeated calculations of market segmentation solutions containing the same number of segments. Very high levels of segment level stability within solutions (SLS_W) for a market segment point to this market segment being a natural market segment. Very low levels for a market segment indicate that this segment is likely to be artificially constructed. Segment level stability within solutions (SLS_W) is visualised using a segment level stability within solutions (SLS_W) plot.

Segment profile plot: The segment profile plot is a refined bar chart visualising a market segmentation solution. The segment profile plot requires less cognitive effort to process than a table containing the same information. As a consequence, the segment profile plot makes it easier for users of market segmentation solutions (managers) to gain insight into the key characteristics of market segments. Segment profile plots portray market segments using segmentation variables only.

Segment separation plot: The segment separation plot allows to assess a segmentation solution. The plot consists of a projection of the data into two dimensions (using, for example, principal components analysis); colouring the data points according to segment memberships; and indicating segment shapes using cluster hulls. The plot is overlayed with a neighbourhood graph indicating the segment representatives (cluster centres) as nodes, and their similarity through the inclusion of edges and adapting edge widths. For simplicity, data points can be omitted.

Socio-demographic segmentation: Socio-demographic segmentation is the result of using socio-demographic information about consumers as segmentation variable(s). Examples include age, gender, income, and education level.

Stability analysis: Stability analysis provides insight into how reproducible market segmentation analyses are. Stability can be assessed at the overall level for the entire market segmentation solution (global stability), or at the segment level (segment level stability within solutions (SLS_W)), segment level stability across solutions (SLS_A)). Stability information points to the most appropriate market segmentation concept (natural segmentation, reproducible segmentation or constructive segmentation); assists in choosing the number of segments to extract; and identifies stable segments.

Target segment: The target segment is the market segment that has been selected by an organisation for targeting.

Validity: See data structure analysis.

References

- ACM (1999) ACM honors Dr. John M. Chambers of Bell Labs with the 1998 ACM software system award for creating “S system” software. ACM press release on 1999-03-23
- Brusco M (2004) Clustering binary data in the presence of masking variables. *Psychol Methods* 9(4):510–523
- Cliff K (2009) A formative index of segment attractiveness: optimising segment selection for tourism destinations. Ph.D. thesis, School of Management and Marketing, Faculty of Commerce, University of Wollongong
- Dahl DB (2016) xtable: export tables to L^AT_EX or HTML. <https://CRAN.R-project.org/package=xtable>, R package version 1.8–2
- de Leeuw J, Mair P (2007) An introduction to the special volume on “Psychometrics in R”. *J Stat Softw* 20(1):1–5
- Dolnicar S, Grün B (2008) Challenging “factor-cluster segmentation”. *J Travel Res* 47(1):63–71
- Dolnicar S, Grün B (2014) Including ‘don’t know’ answer options in brand image surveys improves data quality. *Int J Mark Res* 56(1):33–50
- 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 (2008a) An investigation of tourists’ patterns of obligation to protect the environment. *J Travel Res* 46:381–391
- Dolnicar S, Leisch F (2008b) Selective marketing for environmentally sustainable tourism. *Tour Manag* 29(4):672–680
- Dolnicar S, Leisch F (2010) Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Mark Lett* 21:83–101
- Dolnicar S, Leisch F (2012) One legacy of Mazanec: binary questions are a simple, stable and valid measure of evaluative beliefs. *Int J Cult Tour Hosp Res* 6(4):316–325, special issue in honour of the contributions of Josef Mazanec to tourism research
- Dolnicar S, Leisch F (2014) Using graphical statistics to better understand market segmentation solutions. *Int J Mark Res* 56(2):97–120
- Dolnicar S, Leisch F (2017) Using segment level stability to select target segments in data-driven market segmentation studies. *Mark Lett* 28(3):423–436

- Dolnicar S, Kaiser S, Lazarevski K, Leisch F (2012) Biclustering – overcoming data dimensionality problems in market segmentation. *J Travel Res* 51(1):41–49
- Dolnicar S, Grün B, Leisch F (2016) Increasing sample size compensates for data problems in segmentation studies. *J Bus Res* 69:992–999
- Ernst D, Dolnicar S (2018) How to avoid random market segmentation solutions. *Journal of Travel Research* 57(10): 69–82
- Fox J (2017) Using the R Commander: a point-and-click interface for R. Chapman & Hall/CRC Press, Boca Raton
- Grün B, Dolnicar S (2016) Response-style corrected market segmentation for ordinal data. *Mark Lett* 27:729–741
- Grün B, Leisch F (2007) Fitting finite mixtures of generalized linear regressions in R. *Comput Stat Data Anal* 51(11):5247–5252
- Grün B, Leisch F (2008b) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J Classif* 25(2):225–247
- Hajibaba H, Dolnicar S (2017) Helping when disaster hits. In: Dolnicar S (ed) Peer-to-peer accommodation networks: pushing the boundaries, chap 21. Goodfellow Publishers, Oxford, pp 235–243
- Hajibaba H, Karlsson L, Dolnicar S (2017) Residents open their homes to tourists when disaster strikes. *J Travel Res* 56(8):1065–1078
- Hennig C (2000) Identifiability of models for clusterwise linear regression. *J Classif* 17:273–296
- IBM Corporation (2016) IBM SPSS statistics 24. IBM Corporation, Armonk. <http://www.ibm.com/software/analytics/spss/>
- Leisch F (2004) FlexMix: a general framework for finite mixture models and latent class regression in R. *J Stat Softw* 11(8):1–18
- Leisch F (2006) A toolbox for k-centroids cluster analysis. *Comput Stat Data Anal* 51(2):526–544
- Leisch F (2010) Neighborhood graphs, stripes and shadow plots for cluster visualization. *Stat Comput* 20(4):457–469
- Leisch F, Dimitriadou E (2012) mlbench: machine learning benchmark problems. <https://CRAN.R-project.org/package=mlbench>, R package version 2.1-1
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2017) nlme: linear and nonlinear mixed effects models. <https://CRAN.R-project.org/package=nlme>, R package version 3.1-129
- Sarkar D (2008) Lattice: multivariate data visualization with R. Springer, New York
- Turner R (2017) deldir: Delaunay triangulation and Dirichlet (Voronoi) tessellation. <https://CRAN.R-project.org/package=deldir>, R package version 0.1-14
- Wood SN (2006) Generalized additive models: an introduction with R. Chapman and Hall/CRC, Boca Raton

Index

Symbols

- χ^2 -test, 210, 211, 263
- k -centroid clustering, 90
- k -means, 76, 90, 92–94, 96, 98–101, 110, 166, 274
- k -medians, 92
- t -test, 212, 213
- z -test, 138, 221
- 4Ps, 246, 247

A

- a posteriori segmentation, 15
- a priori segmentation, 15
- absolute distance, *see* Manhattan distance
- acquiescence bias, 47
- adjusted Rand index, 49–51, 154, 158, 159, 163, 164
- agglomerative hierarchical clustering, 83–85, 90
- agreement scale, 66
- AIC, *see* Akaike information criterion
- Akaike information criterion, 132, 222, 280
- analysis of variance, 210–212
- ANOVA, *see* analysis of variance
- answer option, 46
- approaches to market segmentation, 13
- artificial segment, 17, 121, 172, 173, 176
- asymmetric binary distance, 80, 81
- auto-encoder, 105
- average linkage, 84, 85

B

- bagged clustering, 110, 112, 114, 115
- Ball-Hall index, 155
- bar chart, 89, 100, 114, 134, 234
- Bayesian information criterion, 124, 132, 280
- behaviour, 41, 52
- behavioural segmentation, 44
- bias, 47, 52, 145
- BIC, *see* Bayesian information criterion
- bicluster membership plot, 146, 147
- biclustering, 143–145
- big data, 14, 186
- Bimax, 146
- binary data, 46, 142
- binary logistic regression, 217, 220
- Bonferroni correction, 213
- bootstrap, 110, 112, 113, 162, 163, 166, 167, 276
- Boston matrix, 238
- box-and-whisker plot, *see* boxplot
- boxplot, 62–64, 115, 163, 168, 206–209, 223, 226, 276, 289, 290
- boxplots, 276

C

- categorical variable, 59, 201
- centaur, 257
- centroid, 90, 92, 110
- choice experiment, 52
- classification, 215

classification plot, 126–128
 classification tree, 228–234, 291
 cluster index, 154
 co-clustering, 143
 commonsense segmentation, 15, 39, 183
 commonsense/commonsense segmentation, 16
 competitive advantage, 7
 complete linkage, 84–88
 concomitant variable, 142
 conditioning plot, 206
 conjoint analysis, 52
 constructive segmentation, 18, 162, 163, 165, 183
 convenience-group segmentation, *see* commonsense segmentation
 correlation, 50, 68
 covariance matrix, 68, 120, 122, 124, 125
 crisp segmentation, 106
 cross-tabulation, 202–205, 289, 290
 curse of dimensionality, 48

D

data, 39, 110
 data cleaning, 59
 data collection, 270, 288
 data exploration, 57, 271
 data quality, 41, 50
 data structure analysis, 18, 20, 153, 170, 275
 data visualisation, 64, 65, 71, 186, 272, 274
 data-driven market segmentation, 15, 39, 41, 44, 75, 183, 184, 186
 data-driven/data-driven segmentation, 16
 decision matrix, 238
 defining characteristics, 187
 dendrogram, 85–87, 89, 99, 110–112
 describing segments, 39, 199
 descriptor variable, 39, 142, 199, 200, 210, 215, 289
 dichotomous data, 46
 dimensionality, 45, 71
 directional policy matrix, 238
 dissimilarity, 78
 distance, 77–79, 81, 84–86, 89, 92, 98–100, 154, 187
 distance measure, 46, 47, 79
 distance-based method, 77, 78, 116
 distribution, 250
 divisive hierarchical clustering, 83, 84
 DLF IIST, *see* doubly level free answer format
 with individually inferred thresholds
 doubly level free answer format with
 individually inferred thresholds, 47
 dynamic latent change models, 142

E

elbow, 98, 99, 155, 275
 empirical data, 39, 41
 ensemble clustering, 112
 entropy, 118, 174
 Euclidean distance, 80–82, 90, 92
 experimental data, 52
 exploratory data analysis, 57
 external cluster index, 154, 157
 eye tracking, 191

F

factor analysis, 152, 153
 factor-cluster analysis, 151–153, 272
 finer segmentation, 8
 finite mixture model, 116
 finite mixture of binary distributions, 127, 279
 finite mixture of distributions, 119, 120, 127, 279
 finite mixture of normal distributions, 120
 finite mixture of regressions, 133, 282, 285
 five number summary, 62
 fuzzy segmentation, 106

G

Gaussian distribution, *see* normal distribution
 General Electric / McKinsey matrix, 238
 generalised linear model, 216, 217, 224
 geographic segmentation, 42
 global optimum, 90, 102, 144, 281
 global stability, 161–166, 168, 276–278
 global stability boxplot, 164, 165, 276
 gorge plot, 159, 277
 graphical statistics, 186, 200
 graphics, 186
 grouping, 75

H

hard competitive learning, 101, 103
 hierarchical clustering, 83–85, 89, 110–112, 144, 187, 188, 285
 histogram, 61, 62, 206–208
 Holm's method, 213
 hybrid approach, 106
 hybrid consumer, 257
 hyper-segmentation, 8

I

ICL, *see* integrated complete likelihood
 information criterion, 118, 124, 132, 280

initialisation, 90, 96, 101
integrated completed likelihood, 118, 132, 280
internal cluster index, 154–157
internal data, 51
interpretation, 183, 184, 200
irrelevant item, 50

J

Jaccard index, 154, 158, 159, 167

K

Kaiser criterion, 152
knock-out criteria, 237, 238, 270, 291
Kohonen map, *see* self-organising map

L

label switching, 157
latent class analysis, 119, 127, 279
latent class regression, 282
layers of market segmentation analysis, 12
LCA, *see* latent class analysis
learning vector quantisation, 101
linear regression, 215, 216
linkage method, 84
local optimum, 101, 144, 281

M

machine learning, 93, 215
Mahalanobis distance, 126
Manhattan distance, 80, 82, 85, 87, 92
marker variable, 187, 188, 190, 286
market attractiveness-business strength matrix, 238
market dominance, 7
market research, 4
market segmentation, 6, 11
marketing mix, 39, 152, 200, 246, 294
marketing planning, 3
masking variable, 45, 148
McDonald four-box directional policy matrix, 238
measurement level, 66, 119
metric variable, 46, 59, 68, 142, 200, 206, 289
micro marketing, 8
model-based method, 77, 116, 119, 120, 127, 133, 279, 282
monitoring, 295
mosaic plot, 200–206, 210, 211, 226, 288–290
multi-stage segmentation, 16

multinomial logistic regression, 224, 227
multiple testing, 213, 214
mutation, 13

N

natural segmentation, 18
natural segments, 108, 121, 162–165, 168, 170, 172, 173, 176, 183, 192, 194
neural gas, 102, 103, 170, 171, 185, 193
neural network, 105
niche market, 110
niche segment, 7, 110, 145
noisy variable, 45, 46, 51, 143
nominal variable, 46, 200, 210
normal distribution, 120
number of respondents, 146
number of segments, 96, 98, 113, 163, 164, 275, 276, 278
number of variables, 145, 146

O

order of variables, 187
ordinal data, 47
ordinal scale, 66
ordinal variable, 200, 210
organisational constraints, 13
outlier, 68
overlap, 190, 193

P

partitioning clustering, 89, 110, 112
PCA, *see* principal components analysis
perceptual map, 272, 273
place, 4, 250, 251, 294
positioning, 4
post hoc segmentation, 15
pre-processing, 57, 65, 142, 145
price, 4, 247, 249, 294
principal components analysis, 68, 71, 145, 193, 272, 274, 288
product, 4, 247, 294
profiling, 183, 184, 186
promotion, 4, 251, 294
psychographic segmentation, 44
purchase data, 51

Q

questionnaire, 45

R

Rand index, 154, 158, 159
 randomness, 162
 recursive partitioning, 228
 redundant item, 46
 reproducibility, 163
 reproducible segmentation, 18, 162, 163, 165, 183
 resampling, 161
 respondent fatigue, 45
 response bias, 50
 response option, 46
 response style, 47, 50
 return on investment, 8

S

sample size, 48–51, 146
 sampling error, 50
 scale, 46, 57, 66, 82, 119
 scale development, 46
 scree plot, 98–100, 155, 275
 segment attractiveness criteria, 237, 239, 270, 291
 segment evaluation plot, 34, 35, 238–241, 291, 293, 294
 segment evolution, 13, 14, 258
 segment extraction, 75
 segment hopping, 256, 257
 segment level stability, 166, 277, 278
 segment level stability across solutions, 172, 209, 210, 277, 278
 segment level stability within solutions, 167, 169–171, 278, 279
 segment mutation, 14
 segment neighbourhood graph, 102
 segment profile, 152
 segment profile plot, 89, 133, 134, 150, 152, 187, 189–191, 261, 262, 264, 265, 285, 286
 segment revolution, 13
 segment separation plot, 190, 192–195, 287
 segmentation criterion, 6
 segmentation strategy, 12
 segmentation variable, 6, 15, 39, 40, 45, 151, 152, 199, 210, 237, 259
 segmentation-targeting-positioning approach, 245
 self-organising map, 103, 104

similarity, 77, 78, 144
 single linkage, 76, 84–86
 slider scale, 47
 socio-demographic segmentation, 43
 SOM, *see* self-organising map
 stability, 20, 275
 stability analysis, 153, 258
 stacked bar chart, 201, 202, 230
 standardisation, 67, 82
 STP approach, *see* segmentation-targeting-positioning approach
 strategic marketing plan, 3
 supervised learning, 93, 215
 survey, 41, 45–47, 50, 66, 75, 86
 symmetric binary distance, 81

T

tactical marketing plan, 3
 targeting, 4
 team building, 8
 topology representing network, 102, 103
 transformation, 68, 145
 tree-based method, 228, 292
 TRN, *see* topology representing network
 Tukey, 62, 213, 214
 two-mode clustering, 143
 two-step clustering, 107, 108

U

uncertain, 121
 uncertainty plot, 121, 126
 unsupervised learning, 93

V

validation, 133, 153
 variable reduction, 151
 variable selection, 142, 215
 vector quantisation, 107
 visual analogue scale, 47
 visualisation, 88, 94, 98, 102, 146, 186, 200, 206, 287

W

Ward clustering, 85, 188