Expert Systems With Applications 119 (2019) 456–475

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Marketing analytics: Methods, practice, implementation, and links to other fields

Stephen L. France[a,*], Sanjoy Ghose[b]

[a] College of Business, Mississippi State University, 114 McCool Hall, 40 Old Main, P.O. Box 5288, MS 39762, USA
[b] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, 3202 N. Maryland Ave., Milwaukee, WI 53201-0742, USA

ARTICLE INFO

Article history:
Received 28 January 2018
Revised 30 September 2018
Accepted 1 November 2018
Available online 2 November 2018

Keywords:
Analytics
Prediction
Marketing
Visualization
Segmentation
Data mining

ABSTRACT

Marketing analytics is a diverse field, with both academic researchers and practitioners coming from a range of backgrounds including marketing, expert systems, statistics, and operations research. This paper provides an integrative review at the boundary of these areas. The aim is to give researchers in the intelligent and expert systems community the opportunity to gain a broad view of the marketing analytics area and provide a starting point for future interdisciplinary collaboration. The topics of visualization, segmentation, and class prediction are featured. Links between the disciplines are emphasized. For each of these topics, a historical overview is given, starting with initial work in the 1960s and carrying through to the present day. Recent innovations for modern, large, and complex "big data" sets are described. Practical implementation advice is given, along with a directory of open source R routines for implementing marketing analytics techniques.

1. Introduction

It is estimated that the worldwide market in business intelligence and analytics will be worth $200 billion by 2020, up from $130 billion in 2016 (IDC, 2016). Much of this growth is driven by data (Chen, Chiang, & Storey, 2012). Large scale corporate databases, mobile-apps, web analytics data, social media, and sensor data, all contribute to what is commonly referred to as "information explosion". Many of the most interesting and practical applications of analytics are in the field of marketing. In fact, according to an IDG data analytics survey of information systems and analytics executives (IDG, 2016), the top objective (55% of respondents) for analytics implementations is to "improve customer relationships", a core aspect of marketing. In addition, the top three challenges for analytics are "finding correlations across multiple disparate data sources", "predicting customer behavior", and "predicting product or service sales". The second two objectives are direct marketing objectives, while the first objective encompasses a range of consumer analysis applications, including market basket analysis and customer segmentation. The implementation of techniques to solve these challenges is enabled by the availability of large amounts of marketing data. For example, the Oracle Cloud includes customer data gathered from a myriad of sources including web browsing behavior, online bookings, credit card bookings, scanner purchases, and media viewing habits. Overall, given the needs, challenges, and available data described above, there is increasing scope for work in marketing analytics, from the perspective of both practitioners and academic researchers.

2. Objectives

The purpose of this review is to provide a practical, implementation based overview of marketing analytics methodology. A major objective is to synthesize work from different academic areas. Contributions in marketing analytics come from a variety of fields including expert systems, marketing science, data mining, statistics, and operations research. Each discipline has its own literature, and quite often there is little knowledge of similar work in other fields, resulting in some reinventing of the wheel. Each field has it's own emphasis. Hand (1998) notes that while data mining covers many of the same areas as traditional statistics, there is an emphasis on large datasets, exploring data, pattern analysis, and dealing with poor quality data where the usual statistical assumptions do not hold, such as when there is non-stationary data, sample bias, statistical dependence, and contaminated data.

Analytics research in marketing science has particular niche areas of emphasis, including econometric analysis (Wansbeek & Wedel, 1998), Bayesian statistics (Rossi & Allenby, 2003), and psychometric methods (Carroll & Green, 1995; 1997). Operations re-

* Corresponding author.
E-mail addresses: sfrance@business.msstate.edu (S.L. France), sanjoy@uwm.edu (S. Ghose).

https://doi.org/10.1016/j.eswa.2018.11.002
0957-4174/

search has a particular focus on pricing (Bodea & Ferguson, 2014) and location optimization problems (ReVelle & Eiselt, 2005). However, marketing analytics is an area where it is impossible to impose strict disciplinary boundaries. For example, data mining and statistics have become much closer over the last few years, with more statistical rigor from data mining and machine learning researchers and more emphasis on computational implementations and larger datasets from statisticians. This is exemplified by statistical learning work (Hastie, Tibshirani, & Friedman, 2009) that develops data analytic techniques in a statistically rigorous manner. In the commercial arena, the hybrid of data mining and statistics is often referred to as data science.

Another major thread of this review is the increasing importance of expert and intelligent systems in marketing analytics applications. Expert systems can be thought of as computer systems that embody and improve some aspect of human behavior and decision making processes (Bramer, 1982) and utilize both domain knowledge and machine learning or artificial intelligence algorithms (Forsyth, 1984). As early as the 1970s, marketing academics realized that marketing models and methods needed to be implemented as part of integrated decision support systems for business (Little, 1979). Subsequently, marketing decision support systems have continued to be an important part of marketing research and have been utilized in a range of marketing areas including promotion planning (Silva-Risso, Bucklin, & Morrison, 1999), services marketing (Sisodia, 1992), and product design planning (Thieme, Song, & Calantone, 2000). The modern expert systems research field is somewhat eclectic and covers a range of application areas and a wide variety of methods from artificial intelligence and machine learning. However, there is a strong stream of classification research relevant to marketing, and while marketing systems make up the minority of expert systems implementations, a survey (Wagner, 2017) found that expert systems in marketing were more effective than those in any other domain. We hope that with this review, we will provide inter-disciplinary insight and cross-fertilization for marketing analytics researchers working in both expert systems and other arenas.

A major goal of this review is to provide a technical reference for practitioners and academics wishing to implement marketing analytic techniques. In each section, a table is given that summarizes some of the software implementations available for the described techniques. An emphasis is given towards R, as R is open source and is now the de-facto standard platform for statisticians, particularly in the field of data analysis (Ihaka & Gentleman, 1996). More importantly, a global 2015 survey of over 10,000 data science professionals (Rexer, 2016) shows that R is the most popular software package, with 76% of respondents using R, up from 23% in 2007. Given the breadth of the marketing analytics field, a review paper cannot cover the whole range of marketing analytics applications in detail. Thus, a review needs to be selective. In this review, topics were selected that i) are at the interface of multiple academic disciplines, ii) are commonly utilized in industry, and iii) are scalable to "big data" applications. Given the four Ps of marketing (Product, Promotion, Place, and Price), there is an emphasis towards product and promotion, as these topics are core to academic marketing and marketing analytics. Some pricing applications are given, but pure operations research revenue and inventory management applications are not. Likewise, some discussion of "place" is given in the context of visualization and geographic information systems research, but pure location optimization applications are not emphasized.

Fig. 1 gives an overview of the basic disciplines described previously, the methodological areas covered in the review, and the top ten business uses for marketing analytics as defined by the CMO 2016 survey (CMO, 2016) of marketing executives. Throughout the review, the relationships between these three concepts are emphasized. Three major topics have been chosen: Visualization, segmentation, and class prediction. These have been chosen because i) they are core to marketing and have been present in the major marketing journals from the 1960s onward, ii) have strong links with expert systems, statistics, data mining, and operations research, and iii) have recently experienced a "reawakening" in the academic marketing literature due to the explosion of interest in big data and business analytics. They are strongly linked, both methodologically and in business applications. This review does not attempt to give a comprehensive history of analytics in marketing science and some "purer" quantitative marketing science topics have been omitted. An excellent, detailed history of marketing science is given in Winer and Neslin (2014) and a historical discussion on the use of data in marketing is given in Wedel and Kannan (2016).

Other important areas, omited due to space considerations, such as social network analysis in marketing (Iacobucci, 1996), recommender systems (Adomavicius & Tuzhilin, 2005), and time series analysis (Dekimpe & Hanssens, 2000) may be included in a future review.

3. Visualization

3.1. Overview

Given the previously described problem of "information explosion", the ability to understand large, complex, and possibly unstructured data is an important one. Visualizing data in a parsimonious fashion can be used to help understand patterns in data and to forge new insights. John Tukey, who helped pioneer the use of data visualization for exploratory data analysis noted that "The greatest value of a picture is when it forces us to notice what we never expected to see" (Tukey, 1977). To meet increasing needs for visualization, a large number of tools have been developed. In industry, dedicated visualization packages such as Tableau and PowerBI provide an array of features to allow users to summarize and visualize data, while general providers of business intelligence and analytics software such as SAS, SAP, and IBM have incorporated additional visualization features into their offerings (Sallam et al., 2017).

3.2. Foundational methods

Visualization research has a long history in marketing. As early as the 1960s, techniques from psychometrics and applied statistics, such as multidimensional scaling (MDS) and factor analysis have been used to visualize marketing data. Many applications analyzed consumer preferences of products derived from survey data. For example, given a set products $i = 1 \ldots n$, a proximity matrix $\boldsymbol{\Delta} = (\delta_{ij})_{\{n \times n\}}$ can be defined, where δ_{ij} is the proximity between items i and j and is either elicited directly or derived using a proximity metric from preference data. MDS can use the proximity data to generate a p (usually 2) dimensional output data configuration $\mathbf{Y} = (y_{il})_{\{n \times p\}}$, which can be used to analyze how close products are to one another in the minds of consumers (Neidell, 1969). If product preferences are available $\mathbf{X} = (x_{il})_{\{n \times m\}}$, where x_{il} is the rating given by user i for product l, then "joint space" mappings (Green & Carmone, 1969) can be created that plot both consumers and products on the same map, so that $\mathbf{Y} = (y_{il})_{\{(n+m) \times p\}}$, where close proximity between consumers indicates similar preferences, and close proximity between a product and a consumer indicates a strong preference for the product from that consumer. Most multidimensional scaling mapping methods utilize either a decomposition approach or a distance fitting approach. For classical MDS, $\boldsymbol{\Delta}$ is converted to a double centered distance matrix $\mathbf{B} = \mathbf{D} - \bar{\mathbf{D}}_i - \bar{\mathbf{D}}_i' + \bar{\bar{\mathbf{D}}}_i$, where $\bar{\mathbf{D}}_i = \bar{\mathbf{x}}_i \mathbf{1}'$ is a matrix of row means and

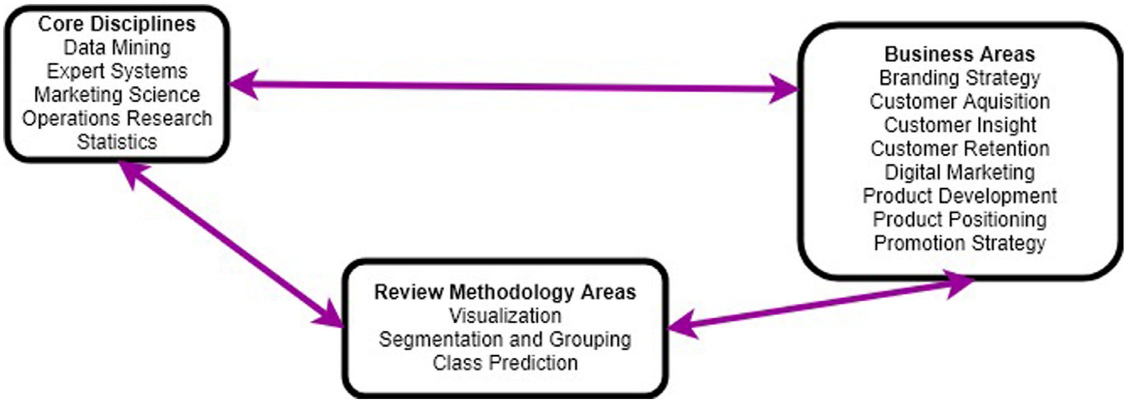

Fig. 1. Interplay of basic discipline, review methodology areas, and business areas.

$\bar{\bar{\mathbf{D}}}_i = \bar{\bar{\mathbf{x}}}_i \mathbf{1}\mathbf{1}'$ is a matrix of the overall mean. Then, an eigendecomposition $\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ is performed giving $\mathbf{Y} = \mathbf{Q}\mathbf{\Lambda}^{1/2}$. $\mathbf{\Lambda}$ is a diagonal matrix and each entry λ_{ii} contains the variance accounted for by the ith new component. Taking the first k columns of $\mathbf{Y}$ gives the derived k lower dimensional solution.

For distance based MDS, $\mathbf{\Delta}$ is transformed to $\hat{\mathbf{D}}$ using a non-metric or metric fitting function, where $\hat{d}_{ij} = F(\delta_{ij})$ and $\mathbf{Y}$ is found by optimizing a distance based minimization criterion. The basic Stress criterion (Kruskal, 1964) is given in Eq. (1). A range of different Stress like criteria and fitting functions are given in Borg and Groenen (2005).

$$Stress = \sqrt{\frac{\sum_i \sum_{j \neq i} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_i \sum_{j \neq i} d_{ij}^2}} \tag{1}$$

where d_{ij} is the Euclidean distance between points i and j in the derived solutions.

Many of these early studies note the trade-off between art and science when creating visualizations. For example, Green, Maheshwari, and Rao (1969) describe the process of rotating (Richman, 1986) the visualization and naming/interpreting the mapping "dimensions", either by business interpretation or by some measure of correlation with product attributes (Carroll & Chang, 1964). Extensions to the basic model exist for more complex datasets. For example, the INDSCAL model (Carroll & Chang, 1970) takes a proximity matrix $\mathbf{\Delta}_s$ for each subject $s = 1 \ldots r$ and creates a single base mapping configuration, but with dimensions weighted (stretched) differently for each individual subject, so that subject s on dimension l is given weight w_{sl}, as shown in (2).

$$\hat{d}_{ij}^s = \sum_{l=1}^{m} \sqrt{w_{sl}\left(x_{il} - x_{jl}\right)^2} \tag{2}$$

where $\hat{d}_{ij}^s$ is the input distance between items i and j for subject s, x_{il} is the output configuration value for item i in dimension l, and for subject s the scaled output values for item i in dimension l are defined as $y_{il}^s = w_{sl}^{\frac{1}{2}} x_{il}$.

When multiple data sources are present, canonical correlation can be used to combine brand mappings (Green & Carroll, 1988) into a single composite mapping. MDS techniques give point estimates, which may be unreliable when significant data variability is present. To help visualize the degree of reliability of the points on the MDS maps, several methods have been proposed to estimate confidence intervals for the MDS point estimates. Ramsay (1978) utilizes a maximum likelihood formulation for MDS to create Bayesian credible intervals for points, while Iacobucci, Grisaffe, and DeSarbo (2017) estimate mean, variance and correlation data from the weighted INDSCAL y_{il}^s values across subjects to create ellipsoidal confidence intervals.

3.3. Advanced methods

There has been a steady stream of product mapping work over the last 30–40 years. Typically, new methods have been developed to account for new and more complex datasets or more advanced statistical methodology. Multidimensional scaling based product mapping methods have been developed for binary choice data (DeSarbo & Hoffman, 1987), incomplete preference data (DeSarbo, Young, & Rangaswamy, 1997), pick any (i.e., choose k brands out of n) choice data (DeSarbo & Cho, 1989; Levine, 1979), asymmetric brand switching data (Desarbo & Manrai, 1992), and scanner data (Andrews & Manrai, 1999; Elrod, 1988). If brand data are nominal then joint space maps can be created using a related technique called correspondence analysis. For example, both yes/no purchase data (Hoffman & Franke, 1986) and nominal brand attribute information (Torres & Bijmolt, 2009) can be analyzed in this fashion (Murtagh, 2005).

Text data, used for opinion mining and sentiment analysis for brands (Pang & Lee, 2008), present a particular challenge, as they are complex and unstructured. Data can be parsed into a document by feature matrix $\mathbf{X} = (x_{il})_{\{n \times m\}}$, where x_{il} is the count in document i of feature l. Here, m, the number of so called "n-gram" features can be very large and can consist of characters, words, syntactic, and semantic features. To analyze and visualize the data, feature selection techniques, such as Abbasi, France, Zhang, and Chen (2011) can be used to reduce the number of features. Product attributes can be elicited from the data using text mining methods. Oelke et al. (2009) extract both attribute information and sentiment polarity information for attributes, then create grid and tree visualizations of product attribute sentiment. Lee and Bradlow (2011) derive product attributes from free-form reviews and then plot these attributes using correspondence analysis. Another approach is Latent Dirichlet Allocation (Blei, Ng, Jordan, & Lafferty, 2003), where each document is considered to consist of a mixture of latent topics and each topic is characterized as a probability distribution over a set of words. This method is utilized by Tirunillai and Tellis (2014), who derive topic mixtures for consumer satisfaction from online reviews, which they use to create MDS maps, and by Lee, Yang, Chen, Wang, and Sun (2016), who derive "service" and "user experience" dimensions for brands from user generated content and plot brands on perceptual maps and radar charts using these dimensions.

Recent work has applied mapping techniques to newly available internet and social media based data sources, such as online auction bidding data (France & Carroll, 2009), user generated content (Dwyer, 2012), and online reviews (Moon & Kamakura, 2017). A key component of this work is the creation of proximity values between products from complex and/or unstructured data. For example, France and Carroll (2009) utilize a temporal weighting scheme for bids in which bids by the same user for similar items are given a weight inversely proportional to the time between bids and Moon and Kamakura (2017) use text mining techniques to quantify different aspects of writing style that can then be compared using a distance metric.

Brand mapping techniques are not solely exploratory. They have particular use in product entry decisions. For example, Shugan (1987) shows how a plot of brand attributes normalized for product cost can be used to determine potential market share for new products and inform brand strategy, an approach that can be applied to scanner data (Hauser & Shugan, 1983). An alternative approach is to combine a product mapping procedure with the selection of product attributes from conjoint analysis in order to optimally position new products (Eliashberg & Manrai, 1992). In fact, Shocker and Srinivasan (1974), note that in order for perceptual product maps to be useful as part of an overall product development strategy, there needs to be some relationship between the perceptual maps and the underlying brand attributes.

3.4. Big data and analytics

While most of the original brand mapping methods were demonstrated on small scale data, the use of web and other large scale data has led to implementations for these data. For example, Ho, Chung, and Lau (2010) introduce a joint-space unfolding decomposition method, which is demonstrated by using ratings from 1000 movie reviewers for 1000 movies to create a joint-space preference map, with reviewers and movie stars plotted on the same map. Ringel and Skiera (2016), when mapping products using similarities derived from clickstream data, use the approach of creating product maps for sub-markets using an MDS-like criterion function, use a transformation to combine the sub-maps, and then add asymmetry to the data. With the growth in the number of available visualization techniques, methods of evaluating the quality of mappings/visualizations are required. As most models optimize some measure of goodness of fit to the original data, some "neutral" measure of quality is needed. A common method is to check rank order item neighborhood agreement between the source data and the derived mappings (Akkucuk, 2004; Akkucuk & Carroll, 2006; Chen, 2006; Chen & Buja, 2009; France & Carroll, 2007). Solution quality across different neighborhood sizes can be mapped elegantly using the idea of a co-ranking matrix (Lee & Verleysen, 2009; Lueks, Mokbel, Biehl, & Hammer, 2011).

Many innovative techniques have been developed to deal with the specifics of marketing data. For example, customer relationship management (CRM) (Payne & Frow, 2005) databases contain large amounts of consumer, contact, finance, and sales information. Yao, Sarlin, Eklund, and Back (2014) describe the use of a mapping technique to plot consumers of a general retailer onto a grid using similarities derived from CRM data and to create diagrams shaded by customer attributes related to customer demographics (e.g., age, gender), purchasing behavior (e.g., basket size, spending amount), product category (e.g., home, outdoors), and product mix.

Beyond mapping, there are many other uses for visualization in marketing analytics. Many of these are extensions to some of the original exploratory data analysis techniques described by Tukey or are adapted from the information visualization literature. For example, in the parallel coordinates approach (Inselberg & Dimsdale, 1987), high dimensional data are plotted onto a two dimensional map by placing the different data dimensions spaced at equidistant intervals along the x-axis and plotting a line for each item, where the item values are plotted for each dimension and then joined. Brodbeck and Girardin (2003) combine an overall tree structure with a parallel coordinates system in order to visualize a customer survey. Klemz and Dunne (2000) use this technique on a longitudinal scanner dataset to examine the interplay between price and market share for coffee brands by plotting both the market share and price points for the brands over time. In fact, the visualization of data over time can lead to important marketing insights. For example, Shmueli and Jank (2005), examine the price curves for Ebay auctions over time and model commonalities across auctions. France, Vaghefi, and Zhao (2016) visualize and model growth curves for viral video views. Keim, Hao, Dayal, and Hsu (2002) describe a pixel-based extension to the bar chart, where for each chart, each brand is plotted in a separate bar and for each customer, the qualitative attributes are coded with different colored pixels. The use of this technique is demonstrated on sales records for over 400,000 customers.

Most of the applications reviewed thus far give fixed visualizations. As described close to the beginning of the chapter, one of the primary motivations behind visualization is to help users "explore", find patterns in, and gain insight from data. Thus, to be useful to managers, visualization environments must be incorporated into business systems and allow users to explore data interactively. In fact, in an overview of modern visualization, Heer and Shneiderman (2012) note that visualization environments for large, complex datasets need to guide users throughout the visualization process and provide tools to sort, select, navigate, annotate, coordinate, and share data. A good example of an interactive system is OpinionSeer (Wu et al., 2010), which takes online reviews for hotels and uses opinion mining and subjective logic to create word cloud visualizations paired with perceptual maps that relate opinions to underlying customer categories such as gender, trip type, and age range.

3.5. Geographic and spatial visualization

One area of visualization that is particularly pertinent to marketing is geographic visualization and mapping. Here, the word "mapping" generally refers to the overlay of business information onto a cartographic map of physical location. While the word "mapping" in marketing usually refers to perceptual or brand mapping, the concept of applying geographic analysis to marketing problems in not new. In fact, some early work on multidimensional scaling compared physical maps of location with derived perceptual maps of location, for cities (Kruskal & Wish, 1978) and for supermarkets (Olshavsky, MacKay, & Sentell, 1975). In recent years, there has been increasing interest in spatial models in marketing, particularly with regards to econometric models that incorporate spatial or distance effects (Bradlow et al., 2005). For example, Bell and Song (2007) build a proportional hazard model on data from an online grocer, which shows that adoption is greatly increased when consumers in neighboring zip codes have also adopted.

Many practical uses of geographic visualization come under the banner of geographic information systems, which are systems that allow users to visualize, explore, and annotate visual data. A key concept of GIS systems is that of layers. Data are built up across different layers, with raster layers forming images and vector layers defining features (land boundaries, roads, stores, etc.). GIS systems have been used extensively in retail analytics, having been used to analyze retail location analysis problems, analyze store performance, and plan shopping malls (Hernández, 2007). The retail location problem is an interesting one, as it is at the intersection of several fields. Original work on gravity models of attraction for stores was given by the Huff gravity model (Huff, 1964), which has been used extensively in GIS/remote sensing data applications (Huff, 2003). The Huff model for the attractiveness of a location is summarized in Eq. (3).

$$p_{ij} = \frac{U_j d_{ij}^{\lambda}}{\sum_{k=1}^{n} U_j d_{ik}^{\lambda}} \tag{3}$$

Here, p_{ij} is the probability that user i uses location j, d_{ij} is the distance between user i and location j, λ is a distance decay parameter, and U_j is the utility for location j. In retail location models, U_j is often operationalized as the size of the store. In the GIS arena and in traditional retail applications, population density data are used to estimate the model. For facility location, several locations can be analyzed with the respect to the model. In a retail context, locations can be selected to maximize the market share taken from competitors and to minimize self-cannibalization. The model is a special case of the Luce choice model (Haines, Simon, & Alexis, 1972). A large number of extensions have been developed for the Huff gravity model, include those that incorporate store image (Stanley & Sewall, 1976), elasticities of demand (Griffith, 1982), and social media location data (Lovelace, Birkin, Cross, & Clarke, 2016). Douard, Heitz, and Cliquet (2015) note that many consumers shop outside of their residence area and describe a model that captures customer flow across geographic areas.

There has been a tradition of building “location analysis” (ReVelle & Eiselt, 2005) optimization models for facility location problems in operations research, many which are based on the original Huff gravity model. Practical location analysis work can combine visual analysis and retail knowledge with operations research methodology. For example, Hernández and Bennison (2000), in a survey of UK retailers, find that while GIS and quantitative decision support systems are increasingly applied to retail location decision problems, the final choice of location is often manual and local retail knowledge, dubbed “retail nose”, is important. Roig-Tierno, Baviera-Puig, Buitrago-Vera, and Mas-Verdu (2013) provide a case study of using GIS for location decisions. They utilize GIS to analyze customer density and retail competition, using kernel density estimation to identify possible retail sites and combine this analytic work with an overall decision analysis methodology.

3.6. Software

R packages for visualization are listed in Table 1. The packages listed are chosen to present a range of basic methods such as PCA and MDS, along with more complex methods designed for larger datasets, mixed measurement types, and textual data. In some cases, particularly with methods developed in marketing/OR, software for cited material is not available, so similar methods have been chosen that can best implement the described material. GIS methods are not included, as a large number of interlinked packages are required, but there are several excellent books and tutorials for implementing GIS analyses in R (Bivand, Pebesma, & Gómez-Rubio, 2013; Brunsdon & Comber, 2015). An example is given in Fig. 2, where a joint space plot was created using “smacof” for a sample of data on youth preferences[1] for different movie genres. One can almost see two distinct clusters of music genres, with the different individuals numbered 1–100 positioned relative to the genres. This configuration has some face validity. For example, rock music, punk, alternative, rock ‘n roll and metal, are all types of guitar driven music performed by bands and are clustered together at the right of the diagram. Pop, dance, hip-hop, reggae, and techno music tend to be driven by electronic sounds and beats and are clustered together at the left of the diagram.

[1] Data can be found at https://www.kaggle.com/miroslavsabo/young-people-survey

3.7. Conclusions

In summary, much has changed since the initial marketing visualization and mapping research in the 1960s and 1970s. Computers have become more powerful, datasets have become more complex, and methodology for analyzing data has become more advanced. However, there are certain research commonalities, which have remained over time. Visualization will always remain a combination of art and science. While quantitative techniques can guide interpretation, there is still a need for managerial insight in order to make business decisions from visualizations. There is a degree of art in how visualizations are pieced together to form a narrative, a process called visualization storytelling (Kosara & Mackinlay, 2013). This trade-off between art and science is common to all areas of marketing. As early as the 1960s, Taylor (1965) noted that while marketing academics developed scientific theory, the implementation of this theory would require “art” based on practitioner experience. Subsequently, the trade-off and tension between art and science in marketing has been a constant topic of interest over the last 30–40 years (Brown, 1996). In this respect, visualization is very typical of marketing practice.

4. Segmentation and grouping

4.1. Overview

Segmentation is a core activity for marketers. In academia, the idea of segmentation arose in the 1950s. Smith (1956) defined market segmentation as the process of “viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing product preferences” and notes the interplay between market segmentation and product differentiation strategies. Initial attempts at segmentation utilized both demographic and psychographic (Yankelovich, 1964) variables.

From an analytics perspective, segmentation refers to the process of grouping or splitting items using a range of segmentation criteria or bases. While the term, “market segmentation” implies the segmentation of consumers or products in a market, any meaningful business entity can be segmented, including countries (Wind & Douglas, 1972), sales territories (Zoltners & Sinha, 2005), and employees (Waite, 2007). Segmentation is now an established part of the marketing management literature. Kotler and Keller (2015) summarize work over prior decades and note that segmentation can be carried out on a wide range of demographic, psychographic, geographic, and behavioral variables and that in order to be managerially useful, segments need to be substantial, measurable, accessible, and actionable, and differentiable.

4.2. Foundational methods

In many ways, the development of segmentation methods is closely linked to the development of the visualization methods described in the previous section, with authors publishing in both areas. Green, Frank, and Robinson (1967) give an application of clustering market territories and describe how the method of cluster analysis can be used to create market partitions. As with MDS, cluster analysis requires input distances or dissimilarities between items. Frank and Green (1968) describe a range of metrics including the Euclidean metric for continuous data and the Tanimoto or Jaccard metric for categorical data, provide additional examples of segmentation for audience profile analysis, customer brand loyalty, experimental gaming, and inter-brand competition, and show how clusters can be overlaid onto MDS solutions.

Two major clustering methods were explored in this early work. The first is hierarchical clustering (Johnson, 1967), where a tree structure is derived by either starting at a single cluster and re-

Table 1
Visualization packages in R.

CRAN	Reference	Description
base	None	cmdscale() implements classical multidimensional scaling.
base	None	prcomp() and princomp() implement principal components analysis.
ca	Nenadic and Greenacre (2007)	Implements simple and multiple correspondence analysis, along with tools for plotting solutions.
dimRed	Kraemer (2017)	Interface for dimensionality reduction techniques, including PCA, MDS, ICA (independent component analysis), and techniques for nonlinear data.
FactoMineR	Lê, Josse, and Husson (2008)	General data mining library containing principal components analysis, correspondence analysis, and multiple correspondence analysis. Includes measures for mixed measurement types.
irlba	Baglama (2016)	PCA for large, sparse datasets of the type found in word count data and online review data.
kohonen	Wehrens and Buydens (2007)	Methods to implement and visualize SOM (self organizing maps).
MASS	Venables and Ripley (2003)	Contains ca() for correspondence analysis, isoMDS() for MDS, and parcoord() for parallel coordinates plot.
PCAmixdata	Chavent, Kuentz-Simonet, Labenne, and Saracco (2014)	Methods for mixed data including mixed PCA.
smacof	de Leeuw and Mair (2009)	Implements distance MDS, INDSCAL, and joint space (unfolding methods).
SpatialPosition	Commenges and Giraud (2017)	Implements the Huff model and associated distance based models.
syuzhet	Jockers (2017)	Connects to a range of sentiment analysis parsers, so text can be scored in terms of sentiment and emotional content.
tidytext	Silge and Robinson (2016)	Contains framework to parse free text into item × feature representation needed for most visualization techniques.
topicmodels	Hornik and Grün (2011)	Fits topic models using latent dirichlet allocation.
vegan	Dixon and Palmer (2003)	Includes functions for MDS, rotating and interpreting MDS solutions, and cannonical correlation analysis.

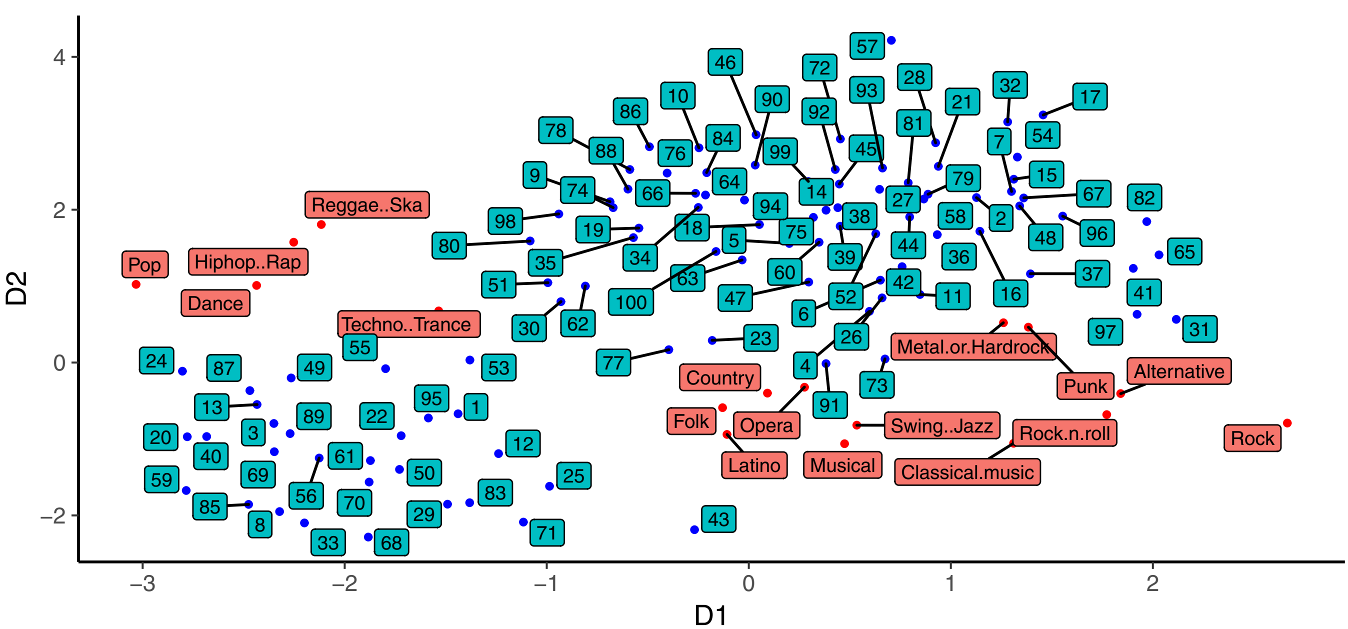

Fig. 2. Joint Space “smacof” unfolding solution for youth music preferences.

peatedly splitting the cluster(s) until every item is in its own cluster (divisive clustering) or by doing the opposite (agglomerative clustering) and taking each individual item and combining until there is a single cluster. For agglomerative clustering, given a distance matrix between items of $\mathbf{D} = (d_{ij})_{\{n\times n\}}$, at each stage of the algorithm the two clusters with the lowest value of d_{ij} are combined. If two items a and b are combined then the distance between the newly combined item and a new item c can be $min\{d_{ac}, d_{bc}\}$ (single linkage), $max\{d_{ac}, d_{bc}\}$ (complete linkage), or $\frac{d_{ac}+d_{bc}}{2}$ (average linkage). There are a multitude of additional schemes, including Ward's method (Ward, 1963) to minimize variance, and centroid clustering, where cluster means are explicitly calculated. Each combination scheme has its positives and negatives. For example, single linkage clustering produces long drawn out clusters and is susceptible to outliers (Milligan, 1980) and complete-linkage clustering produces more spherical clusters. Various Monte Carlo studies have been run to compare the performance of hierarchical cluster analysis algorithms. It is generally believed that Ward's method gives the most consistent performance (Morey, Blashfield, & Skinner, 1983), but some studies have shown that average linkage and centroid clustering algorithms can give better cluster recovery, particularly in cases where there are outliers (Milligan, 1981).

A second type of clustering is partitioning clustering. Here the items are directly clustered into some predefined k number of clusters. Consider a data matrix $\mathbf{X} = (x_{il})_{\{n\times m\}}$, where each row contains the observations for a single item to be clustered. The most common method of doing this is by minimizing the within cluster sum of squares criterion (SSW) given in (4).

$$SSW = \sum_{i=1}^{k}\sum_{\mathbf{x}_j\in C_i}\|\mathbf{x}_j-\bar{\mathbf{x}}\|^2 \qquad (4)$$

Here, the total Euclidean distance between each row vector $\mathbf{x}_j$ and its centroid $\bar{\mathbf{x}}$ (average value of items in its cluster) is minimized.

A common method of minimizing this criterion is through the k-means algorithm (MacQueen, 1967), where items are randomly assigned to initial seed clusters, the cluster centroids are calculated, and then at each iteration of the algorithm, each item is assigned to its nearest centroid and the centroids are updated.

The authors of this early work noted several challenges. These include i) how to decide on the number of clusters, ii) how to operationalize distance or similarity measures, particularly in cases when different dimensions/attributes have different scales, iii) how to handle data where dimensions are correlated, and iv) how to properly define the boundaries of the clusters. Subsequent papers test the use of a large number of metrics including city-block, variance adjusted (Mahalanobis), and ordinal metrics (Green & Rao, 1969). Morrison (1967) develops a metric that accounts both for variation and for weighting dimensions by managerial intuition. Punj and Stewart (1983) note a separate challenge, that of cluster validation. It is possible that cluster solutions are unstable and change with minor changes in the dataset. Thus metrics to ensure the stability of the cluster analysis with respect to data variation are needed. A range of metrics have been developed to test cluster validity (Milligan, 1996); the most common of which is the Hubert–Arabie adjusted Rand index for comparing partitions (Hubert & Arabie, 1985).

4.3. Managerial intuition

While initial marketing analytics work on cluster analysis concentrated on the development of algorithms, managerial intuition was not ignored. Cluster analysis can be used to split customers into homogeneous subsets based on their characteristics, but does not provide guidance on how to utilize these characteristics as part of an overall marketing strategy. Several methods were developed to address these challenges. Green, Carroll, and Carmone (1977) note that there are two typical approaches for segmentation strategy. The first is “a-priori” segmentation, where there is some cluster defining descriptor, such as a favorite brand or brand category. The second is post-hoc segmentation, where analytic segmentation is carried out on a range of demographic, behavioral, or psychographic characteristics. The results of the segmentation are then analyzed with respect to the original segmentation bases (e.g., mean income, brand awareness, etc.) with an eye towards managerial action. However, Green et al. (1977) argue that post-hoc segmentation, while widely used, does not necessarily predict response to future products or services. Thus, they give a factorial design based segmentation scheme, which simultaneously analyzes consumer categories and sets of desirable product feature categories using a multi-way linear model. Mahajan and Jain (1978) describe how the two-stage process of calculating segments and then assigning resources to segments can be inefficient and build a resource based mathematical programming model to counter this issue.

Wind (1978) summarizes early research in segmentation and notes that any segmentation study should i) start with a managerial problem definition and go through research design, data collection, data analysis, and interpretation stages. The authors also note that the choice of segmentation bases will depend on the application. For example, appropriate segmentation bases for product positioning studies include product usage, preference, and benefits, for pricing decisions include price sensitivity and deal proneness, and for advertising decisions include benefits sought, media usage, and psychographic variables. Young, Ott, and Feigin (1978) describe practical limits to what is possible with segmentation and give situations where segmentation is not appropriate. These situations include when a market is too small to be profitable, when a few users dominate the market and most marketing effort is targeted to these users, and when a single brand dominates the market and is purchased by all segments of the population.

Lifestyle segmentation (Lazer, 1963) was developed as existing forms of segmentation were found to be insufficient. Demographic segmentation is too broad-scope and does not contain enough information about behavior, segmentation on psychological attributes is not reliable, and specific brand usage segmentation can be too narrow. Plummer (1974) notes that lifestyle information contained in the AIO (activities, interests, and opinions) framework (Wells & Tigert, 1971) can be used to supplement demographic segmentation and gives insight into consumer behavior that cannot be accounted for by either broad-scope demographic segmentation or usage segmentation on specific products. Wells (1975) describes limits to lifestyle segmentation in that if a segmentation solution is too abstract then it is managerially useless, but conversely, if it is too specific, then it is too close to actual behavioral data to give any additional insight. Here, the author gives an example of applying lifestyle segmentation to profiles of heavy users vs. non-users of shotgun ammunition. They note that the biggest differentiator between these categories is an enjoyment of hunting, which is closely tied to actual hunting behavior.

Dhalla and Mahatoo (1976) posit that any one type of market segmentation cannot be a silver bullet. All types of segmentation have drawbacks. A purely demographic segmentation is limited in predicting marketing response. Psychographic tests adapted to marketing are designed to test underlying psycho-social traits, and have limited ability to predict purchases of specific products. Behavioral data, such as brand loyalty, can be used to group consumers using purchase patterns, but cannot distinguish between a consumer who buys a product because it is the only product available and a consumer who has high utility for the product. Overall, Dhalla and Mahatoo (1976) conclude that segmentation is only useful when it covers all aspects of consumer and buyer behavior and if segments respond differently to a firm’s marketing efforts. This idea is empirically tested by Woodside and Motes (1981), who carry out a large scale survey of tourism preferences, finding that different segments had different sensitivity elasticities to different advertising strategies.

4.4. Beyond partitioning clustering

The partitioning clustering methods described thus far assume that all consumers or brands belong to one and only one cluster. When it comes to real life interpretation of marketing data, this is not necessarily a realistic assumption. Consider a segmentation of movies based on customer preferences, where segments have been found for “romance”, “action”, “horror”, “sci-fi”, “teen”, and “comedy” movies. A movie may belong in multiple segments; for example, the popular teen movie Twilight could feasibly belong in the “romance”, “sci-fi”, and “teen” genres. From a modeling standpoint, define a k cluster solution $\mathbf{P} = (p_{ij})_{\{n\times k\}}$, where $p_{ij} \in \{0,\ 1\}$ is a cluster assignment from item i to cluster j. If $p_{ij} = 1$ and $\sum_{j=1}^{k} p_{ij} = 1, \forall i = 1 \ldots n$, then this is partitioning clustering. An alternative way of conceptualizing clustering is overlapping clustering. Here, each item can be assigned to more than one cluster and $\sum_{j=1}^{k} p_{ij} \geq 1, \forall i = 1 \ldots n$. Overlapping clustering was introduced as a method for product positioning by Arabie, Carroll, DeSarbo, and Wind (1981), noting that in benefit segmentation (Haley, 1968), a product may belong to multiple segments. For example, chewing gum may be used both as a candy substitute and for dental health. The authors implement the ADCLUS (ADditive CLUStering) model (Shepard & Arabie, 1979), in which a similarity matrix $\mathbf{S} = (s_{ij})_{\{n\times n\}}$ is decomposed as $\mathbf{S} = \mathbf{PWP}'$, where $\mathbf{P}$ is the aforementioned overlapping assignment matrix and $\mathbf{W} = (w_{ij})_{\{k\times k\}}$ is a weighting matrix for the different clusters. The model is used to analyze different usage scenarios for different breakfast food products.

A number of extensions to the ADCLUS model have been implemented. The INDCLUS model (Carroll & Arabie, 1983) allows a similarity matrix for each user or data source and can be used in instances where each user compares multiple products, for example an online review scenario. It can also be used when data are split by qualitative attributes, such as region or demographic group (Chaturvedi & Carroll, 2001). INDCLUS can be thought of as a discrete variant of INDSCAL. In the INDCLUS model, for each user $i = 1 \ldots r$, $\mathbf{S}_i = \mathbf{P}\mathbf{W}_i\mathbf{P}'$, with each user assigned a weight for each cluster. DeSarbo (1982) generalizes INDCLUS for asymmetric data, overlapping or non-overlapping clustering, and a range of weighting options and demonstrates the utility of the model on both brand switching data and celebrity/brand congruence data. Chaturvedi and Carroll (2006), drawing on psychological justification for the use of hybrid models (Carroll, 1976), introduce the CLUSCALE model, which contains both continuous and discrete dimensions and is a hybrid of both INDCLUS and INDSCAL models. This model is demonstrated using a segmentation of the car market.

Yet another clustering conceptualization is that of fuzzy clustering. Here, $p_{ij} = 1$ as per partitioning clustering, but the values of p_{ij} are membership probabilities with $p_{ij} \in [0, 1]$. Fuzzy clustering is useful when dealing with items that are positioned towards the "edge" of clusters. For example, consider a customer segmentation where each segment S is targeted with a promotion for a specific brand. However, the customers at the edge of each cluster may have a lower probability of utilizing the promotion and are thus less profitable. A marketer could save money by only targeting consumers where p_{ij} is above a certain threshold. The most common fuzzy clustering algorithm is a fuzzy variant of the McQueen k-means algorithm entitled c-means clustering (Bezdek, Ehrlich, & Full, 1984). Casabayó, Agell, and Sánchez-Hernández (2015) describe a procedure that fuzzifies existing partitioning clusters, to allow extra insight and context to be gained on consumers who are weaker members of segments. Hruschka (1986) describes a fuzzy version of the ADCLUS model that can be used to segment both customers and brands. A related model that is somewhat intermediate to k-means clustering and additive clustering is k-centroids clustering (Chaturvedi, Carroll, Green, & Rotondo, 1997). Here, a consumer $\times$ features matrix $\mathbf{X} = (x_{il})_{\{n\times m\}}$ is decomposed into a binary cluster indicator matrix $\mathbf{P} = (p_{ij})_{\{n\times k\}}$ and a matrix of cluster centroids $\mathbf{W} = (d_{jl})_{\{k\times m\}}$, so that $\mathbf{X} = \mathbf{PW}$. The advantage of this model is a direct correspondence between items (or customers), segments, and features. To demonstrate the use of the model, a conjoint analysis (Green & Srinivasan, 1978; 1990) experiment was performed for 600 users and 9 product features, resulting in a 600×9 matrix of attribute utilities, which were then input into the segmentation procedure. Cross-validation and comparison to exogenous variables was used to validate the solutions.

4.5. Model based and econometric approaches to segmentation

Most of the methods described thus far do not make any parametric or distributional assumptions. This gives advantages in terms of flexibility, but limits some of the output from the procedures in terms of parameter significance and model selection criteria. Mixture model based clustering (Banfield & Raftery, 1993; Fraley & Raftery, 2002; McNicholas, 2016) approaches that make distributional assumptions have been increasingly utilized for marketing segmentation purposes, particularly over the last few decades. In mixture model based clustering, the observed data are considered to come from a "mixture" of different distributions. Given an observation vector $\mathbf{x}_i$ for each item to be clustered and k clusters, the basic mixture model based clustering formulation is given in (5).

$$f(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \tau_k f_k\left(\mathbf{x}_i|\boldsymbol{\theta}_k\right) \tag{5}$$

Here, $f_k\left(\mathbf{x}_i|\boldsymbol{\theta}_k\right)$ is a probability function with parameters $\boldsymbol{\theta}_k$ and τ_k is some prior probability of membership of cluster k. Most commonly, $\sum_{k=1}^{K} \tau_k f_k\left(\mathbf{x}_i|\boldsymbol{\theta}_k\right)$ is implemented as a Gaussian mixture across k, with covariance matrix $\boldsymbol{\Sigma}_k$. Different model assumptions specify different covariance matrices. The mixture model returns probabilities, which can be scaled as per the membership parameters in fuzzy clustering. Let z_{ik} be the assignment of item i to cluster k. Let CI_i be the chosen cluster index for the ith item. To get a partitioning solution, the maximum value of $p(CI_i = k|\mathbf{x}_i, \boldsymbol{\Theta})$ is selected. For a model based overlapping clustering solution, the cluster assignment $z_{ik} = 1$ if $p(CI_i = k|\mathbf{x}_i, \boldsymbol{\Theta}) > \lambda$, where λ is a threshold parameter (Banerjee, Krumpelman, Basu, & Mooney, 2005).

In marketing, model based clustering often comes under the banner of latent class analysis, which can be used to analyze brand switching data (Jain, Bass, & Chen, 1990). In a marketing segmentation context, Magidson and Vermunt (2002) note that latent class clustering has several advantages over k-means clustering, which include the fact that membership probabilities are generated, which can be used to estimate error, likelihood based diagnostic statistics such as the AIC and BIC (Kuha, 2004; Vrieze, 2012) can easily be used to estimate the number of clusters, a wide range of measurement types are allowed, and exogenous information can be modeled, allowing for simultaneous clustering and descriptive analysis of the clusters. However, k-means clustering has a range of methods for choosing the number of clusters (Steinley, 2006), including the Caliski and Harabasz (1974) ratio of between to within cluster sum of squares and the gap statistic (Tibshirani, Walther, & Hastie, 2001), in which the gap between the sum of squares criterion for the clustering solution and the average sum of squares criterion for clustering solutions from data generated uniformly from the range of the original data is minimized across k. An experimental evaluation of several of these methods is given by Chiang and Mirkin (2010). In fact, on a set of experimental data, Steinley and Brusco (2011) found that combining k-means clustering with the Caliski-Harabasz method gave better recovery of the number of clusters than model based clustering using AIC and BIC. In addition, variants of k-means have been developed that can account for categorical data (Huang, 1998) and external information. To summarize, both k-means based and mixture model based styles of clustering have their adherents and advantages/disadvantages. The technique to be used will depend on the specific dataset and researcher preference.

In a very separate tradition, market share and latent class based methods have been developed by marketing researchers to help explain segmentation behavior and analyze specific market segmentation scenarios for customers and products. Much of this segmentation work builds on the concepts of structured markets and submarkets, where brands within submarkets compete with one another. Urban, Johnson, and Hauser (1984) define the concept of submarkets and develop a set of statistical tools to analyze the existence of submarkets based on categorical features, for example {diesel, gas} cars, or {mild, medium, dark} roasted coffee. A modern extension to identify and visualize brand submarkets is given by France and Ghose (2016).

Grover and Srinivasan (1987) build an explicit latent class model to model brand loyal segments and brand switching segments from brand switching data. Consider a brand switching matrix $\mathbf{S} = (s_{ij})_{\{n\times n\}}$ that records cross purchases over two purchase occasions, where s_{ij} is the number of households who purchase brand i on the first occasion and brand j on the second occasion.

The resulting segmentation model is given in (6).

$$s_{ij} = \sum_{h=1}^{n+k} \beta_h q_{ih} q_{jh} \tag{6}$$

Here, there are n brand loyal segments and k switching segments, β_h is the size of segment h, and q_{ih} is the probability that a consumer in segment h purchases brand i across multiple time periods. For brand loyal segments, $q_{ih} = 1$ if h is the brand loyal for segment i and $q_{ih} = 0$ otherwise. The model can easily be fit using a maximum likelihood fitting procedure. There have been many extensions to the core choice-based model to deal with various business scenarios. These include analyzing segmentation structure over time (Grover & Srinivasan, 1989), incorporating promotion effects (Grover & Srinivasan, 1992), incorporating price elasticity variables (Kamakura & Russell, 1989), and using multinomial logit models to analyze price sensitivity across purchase incidence and brand choice segments (Bucklin & Gupta, 1992).

The models described above have found most practical use in analyzing scanner data from supermarkets. Russell and Kamakura (1994) note that the "micro" household data used to track latent class models allow for more detailed analysis than previously used "macro" store level market share data and propose using the segmentation results from latent class analysis to help estimate macro level brand share and momentum parameters. In addition, Montgomery (1997) shows using a hierarchical Bayesian model that customized store level pricing and promotion strategies derived from micro-level scanner data can improve gross profit margins from 4% to 10%.

4.6. Clusterwise regression

As described previously, to be useful, segmentation solutions need to be actionable and some analysis must be made of the resulting segmentation solutions with respect to consumer behavior. One way of examining this is by using a technique called clusterwise regression, described by DeSarbo and Cron (1988), who expand initial statistical work (Späth, 1979) to create a maximum likelihood model that simultaneously clusters a set of independent variables in a regression, while fitting optimal regression equations relating the independent variables to a dependent variable. The formulation, given in (7), is similar to the mixture model clustering formulation (5), but the mixture is defined across the continuous variable y, with each segment $k = 1 \ldots K$ having a regression equation defined by a set of parameters $\boldsymbol{\beta}_k$ and a homoskedastic variance σ_k.

$$y_i = \sum_{k=1}^{K} \tau_k f_{ik}\left(y_i | \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k\right) \tag{7}$$

An example is given with trade show data, where the problem was to examine which factors managers considered important for the success of a trade-show visit. Here the dependent variable was the overall rating for a trade show visit, while the independent variables were ratings for success in certain sub-areas, such as sales, new clients, new product launches, corporate image, morale, and information gathering. Two distinct segments were found, with one segment prioritizing sales and the other, a more general marketing segment, having more balanced priorities. This approach has been applied to more general customer segmentation problems. For example, DeSarbo and Edwards (1996) cluster consumers with compulsive behavior shopping problems into two clusters with different drivers for compulsive behavior.

Wedel and Steenkamp (1989) extend clusterwise regression to allow for the previously described fuzzy clustering paradigm. A product benefit segmentation example is given in Wedel and Steenkamp (1991). Here, based on an MDS configuration derived from a consumer survey, twelve brands of margarine are clustered using ratings on exclusiveness, vegetable content, multiple purpose flexibility, and packaging, with regression used to compare these data with actual product attributes. As most products could be used for most purposes, a fuzzy clustering solution is more appropriate in this scenario than a partitioning clustering solution. Brusco, Cradit, and Tashchian (2003) note that within a segmentation context there are often cases where a single dependent variable is not indicative of behavior; for example, when analyzing car repurchase behavior, consumers with a high degree of customer satisfaction are still liable to switch, so to analyze behavior a second variable measuring propensity to brand switch is required. Thus, they develop a multi-criterion programming approach in which clustering is optimized over multiple dependent variables. Brusco, Cradit, Steinley, and Fox (2008) note that while clusterwise regression provides a flexible framework for actionable market segmentation, care must be taken, as it is prone to overfitting, which can be mitigated using a procedure in which the model results are compared to those gained from fitting a model with randomly generated dependent variables.

4.7. Modern, large scale segmentation approaches

As per visualization, over the last 10–20 years, segmentation methods have been developed to both account for increasing large scale, complex, online and corporate data. Many of the basic clustering/segmentation techniques described previously have been adapted to deal with this reality. This has generally been achieved by either by improving the efficiency of the implementation algorithms or by parallelization, i.e., splitting up computation into multiple threads and running simultaneously. For hierarchical clustering, Koga, Ishibashi, and Watanabe (2007) develop a method for agglomerative clustering that approximates the process of finding the nearest neighbor using hashing for choosing items to be merged, reducing the complexity of the algorithm. In addition, there have been several papers on how to best parallelize agglomerative clustering algorithms (Dahlhaus, 2000; Li, 1990; Olson, 1995). Similar methods exist for k-means clustering. One method is to reformulate the k-means data structure in a tree problem and then develop distance based criteria to prune the tree (Alsabti, Ranka, & Singh, 1998; Kanungo et al., 2002), thus reducing the number of possible distance comparisons. In a similar fashion, Elkan (2003) utilizes the triangle equality to discard distance comparisons that are not possible. As per agglomerative clustering algorithms, there have been several algorithms for parallelizing computation (Stoffel & Belkoniene, 1999; Zhang, Xiong, Mao, & Ou, 2006). A method for combining parallel k-means clustering with the general MapReduce framework for distributed computing is given by Zhao, Ma, and He (2009). Model based clustering approaches have been typically constrained by the performance of the EM algorithm (McLachlan & Krishnan, 2007) used to maximize the likelihood functions. However, several approaches have been used to speed up estimation. In a basic sampling approach, a random sample of the data is used to calculate the clusters and then an additional "E" (expectation) step is used to classify the remaining items. This approach can be improved by building multiple models from the initial sample and then running through several steps of the EM algorithm to fit the whole dataset to these models (Wehrens, Buydens, Fraley, & Raftery, 2004) or by looking to create new clusters for observations in the full dataset that are fit badly by the sample clusters (Fraley, Raftery, & Wehrens, 2005). In addition, as per other clustering approaches, parallel methods have been developed (Kriegel, Kroger, Pryakhin, & Schubert, 2005; McNicholas, Murphy, McDaid, & Frost, 2010).

Recent segmentation research has applied a broad range of algorithms to modern datasets. Of particular interest are neural network algorithms. These algorithms have in-built parallelization and can be applied to a range of segmentation and classification scenarios. Neural networks take a set of problem inputs and use a hidden layer to transform the inputs to a set of output variables. Neural networks are applied to a simple product segmentation problem in Hruschka and Natter (1999). The hidden layer equation (8), uses a multinomial logit formulation similar to (6).

$$s_{il} = \frac{exp\left(\sum_{j=1}^{m} \alpha_{jl} x_{ij}\right)}{\sum_{h=1}^{k} exp\left(\sum_{j=1}^{m} \alpha_{jh} x_{ij}\right)}, \quad \hat{y}_{ij} = \frac{1}{exp\left(-\sum_{h=1}^{k} \beta_{hj} s_{ih}\right)} \tag{8}$$

Here s_{il} is the segment probability for item i in segment l, α_{jl} is the weight for feature j in segment l, and x_{ij} is the value of feature j for item i. The fitted output $\hat{y}_{ij}$ for item i in feature j is calculated across all values of the weighted segment probabilities $\beta_{hj} s_{ih}$ and the neural network error function minimizes the error between each $\hat{y}_{ij}$ and its segment average. A neural network is applied to real world retail segmentation exercises in Boone and Roehm (2002). The authors utilize an error function that forces probabilities of cluster membership towards 0 or 1 and as per k-means, minimizes the distances to the cluster centroids (Kamgar-Parsi, Gualtieri, Devaney, & Kamgar-Parsi, 1990). The authors found that this method had less reliance on the starting cluster centroid configuration and lower overall error than both k-means and model based clustering. Neural networks have been used for a range of segmentation applications including online shopping behavior (Vellido, Lisboa, & Meehan, 1999), web-log sequence mining (Park, Suresh, & Jeong, 2008), tourism visitor segmentation (Brida, Disegna, & Osti, 2012), and visit/usage segmentation at a dental clinic (Wei, Lin, Weng, & Wu, 2012). Tuma, Decker, and Scholz (2011), in a survey of segmentation with cluster analysis, note that there are now a range of different neural network types that have been applied to cluster analysis, including topology representing networks (TRNs), Self-organizing maps (SOMs), and Hopfield-Kagmar (HK) neural networks. Each method has advantages and disadvantages.

A range of other previously described techniques have been adapted for modern datasets. Brito, Soares, Almeida, Monte, and Byvoet (2015) segment a fashion database of 7000 consumers using k-medoids clustering (a variant of k-means with an L_1 city-block distance) and subset mining, which is designed to find interesting relations for items/customers who have an specific distribution of a target variable. In this case, it was used to elicit fashion preferences for overweight consumers. Griva, Bardaki, Pramatari, and Papakiriakopoulos (2018) segment customer visits rather than customers. This type of analysis allows products and product classes to be matched to different types of purchase occasions, including breakfast, light meal, extended visits around food, and detergents and hygiene. Arunachalam and Kumar (2018) give a modern analysis of benefit segmentation, testing a range of different distance metrics and clustering methods. The authors find that SOM and fuzzy clustering give strong clustering solutions and that the Gower distance, where data are range scaled within dimensions and then added up as per the Manhattan distance, and the generalized distance metric, which is a measure of generalized correlation, give good results when segmenting on ordinal Likert scale data.

In the marketing literature, over the last 20 years there has been an influential work at the intersection of marketing and accounting/finance on measuring customer lifetime value (CLV) (Berger & Nasr, 1998) and optimizing marketing processes to maximize this value (Venkatesan & Kumar, 2004). Chen, Fu, and Zhu (2008) develop a segmentation methodology, which involves calculating CLV and segmenting on CLV along with some measure of customer loyalty. The methodology is demonstrated on a database of frequent flier information for a Chinese airline.

4.8. Software

R packages for classification and grouping are listed in Table 2. A range of both classical and mixture model/latent class based procedures are included. Packages that have features for visualization and cluster validation are included. Procedures designed for large datasets and datasets with mixed measurement types are also included.

An example is given in Fig. 3, which illustrates the results from cluster analysis performed with the previously utilized youth preferences dataset. The first subplot utilizes the “cluster” package to give the gap statistic for $k = 1..20$ k-medoids clustering solutions for partitioning the survey participants. Using the algorithm described by Tibshirani et al. (2001), a 12 cluster solution is suggested. The second subplot uses the “dendextend” package and the hclust() function to compare single and average linkage clustering solutions for the different music genres. There are some differences between the solutions, but also some commonalities. As with the previous visualization solution, there is a degree of face validity, particularly for the average linkage solutions. The sets of genres of {pop, dance, hip-hop}, {punk, metal, rock}, and {folk, country} are clustered together in groups towards the bottom of the tree, indicating close correspondence.

4.9. Conclusions

Segmentation, as a method, has been utilized by both marketing academics and practitioners. Segmentation methods remain important for marketing practitioners. In fact, in a recent survey designed to determine the influence of major marketing science innovations on marketing practice (Roberts, Kayande, & Stremersch, 2014), segmentation was the highest rated innovation by both academics and practitioners. However, despite this overall impact, in a critical review of market segmentation in the Harvard Business Review, Yankelovich and Meer (2006) described a survey of 200 senior executives, where 59% reported carrying out a major segmentation exercise within the last two years, but only 14% reported that any real business value resulted from the exercise. The authors note that psychographic segmentation has become widely established over the last 50 years and has been instrumental in some foundational advertising campaigns, including the Pepsi “Generation” campaign, which melded together groups of consumers who identified with youth culture. However, the authors note that psychographic segmentation still has limited usefulness when it comes to predicting specific brand behavior. This ties in with some of the issues previously raised by Young et al. (1978) and the trade-offs between general behavioral and brand specific attributes. Furthermore, Barron and Hollingshead (2002) note that despite technical advances in segmentation in the 1980s and 1990s, segmenting towards needs or feelings rarely finds actionable segments of customers who can be targeted in the real business world. To counter this and other issues, the authors recommend a multidisciplinary team to brainstorm segmentation criteria and to paying attention to a broad range of factors including the purchase/user environment, a customer’s desired experience, and product beliefs/associations, in order to be able to derive segments related to purchase and usage behavior.

Several recent advances hold promise in the segmentation arena. The first is the use of neuroscience, which provides a set of segmentation bases based on unconscious cognitive responses to stimuli (Venkatraman, Clithero, Fitzsimons, & Huettel, 2012), can provide brand-specific psychological responses from consumers and when combined with traditional segmentation bases, can provide great insight into brand behavior. Another advance is the rise of micro-segmentation. A New York Times article (Rosen, 2012) describes how the BlueKai platform (now owned by Oracle), which

Table 2
Segmentation and grouping packages in R.

CRAN	Reference	Description
base	None	hclust() implements hierarchical clustering with a range of linkage methods.
base	None	kmeans() implements basic k-means clustering procedure.
CluMix	Hummel, Edelmann, and Kopp-Schneider (2017)	Cluster analysis and visualization for mixed (continuous and categorical) data.
cluster	Kaufman and Rousseeuw (1990)	General clustering package. Contains clara() for large scale k-means clustering and pam() for partitioning on medoids.
dendextend	Galili (2015)	A package for visualizing and comparing dendograms from hierarchical clustering.
fastclust	Müllner (2013)	Fast, scalable hierarchical clustering algorithms for large datasets.
fclust	Giordani, Ferraro, and Giordani (2015)	Implements fuzzy clustering algorithms.
flexclust	Leisch (2006)	K-centroids clustering, with a choice of distance metrics and various advanced methods such as neural clustering.
fpc	Hennig (2015)	Contains clusterwise regression, a range of fixed point clustering methods, and clustering validation routines.
MCLUST	Fraley and Raftery (2003)	Gaussian model based clustering models and algorithms.
poLCA	Linzer and Lewis (2011)	Polytomous variable latent class analysis, including latent class (clusterwise) regression.
skmeans	Karypis (2002)	Contains an interface to the CLUTO vmeans function which provides a range of criteria for partitioning clustering from distance matrices. Particularly useful for text/document clustering.

contains a wide variety of behavioral and financial data from a range of sources, including browsing behavior, credit card records, and e-commerce purchases, can be used to group consumers into small micro-segments based on behavior and demographics. Examples given include “Hawaii-vacation-seeking price-sensitive Democrat” and “baseball-loving safety-net senior oenophile”. These segments contain quite specific product needs and can be algorithmically targeted by flexible e-commerce engines.

5. Class prediction

5.1. Discriminant analysis and related techniques

Marketing prediction encompasses a plethora of models for a range of responses, including purchase behavior, review ratings, customer loyalty, customer lifetime value, sales, profit, and brand visibility. To keep length manageable we will concentrate on methods of prediction where predictions are made for either a specific class label and for which the scope of the prediction is at a high level of “granularity”, i.e., predictions for an individual consumer or product. For example, for a customer renewal prediction application, let $y_i \in \{0, 1\}$, where 0 indicates that a customer “churned” or did not renew a contract and $y_i = 1$ indicates the converse. Class prediction techniques can be applied to multi-class classification problems where there are more than two categories and also instances where there are more than two labels to be predicted, the so-called multi-label classification problem (Tsoumakas, Katakis, & Vlahavas, 2010). Though class prediction requires discrete data, class prediction methods can be applied to discretized continuous data; for example, Ballings and den Poel (2015) operationalize change in service utilization as a dichotomous increase/decrease variable.

In most instances, models will be built and tested using data for which y_i is known and then applied on data for which y_i is not known. From a data mining terminology standpoint, this is known as “supervised learning”, as opposed to the unsupervised learning techniques of dimensionality reduction and cluster analysis, where there is no known dependent variable that can be used to guide the output. Typically, models will be tested with some type of holdout or cross validation scheme (Arlot & Celisse, 2010). Here, a proportion of the dataset is denoted as the training data and is used to predict y_i on the remaining holdout or test data. This can be repeated iteratively by repeatedly sampling test data without replacement until every observation is included in a holdout dataset. If p items are held out at each iteration then the procedure is known as “hold-p-out cross validation”. If a proportion of $1/n$ of the data is held out at each iteration then the procedure is known as “n-fold cross validation”. The predictions for the test

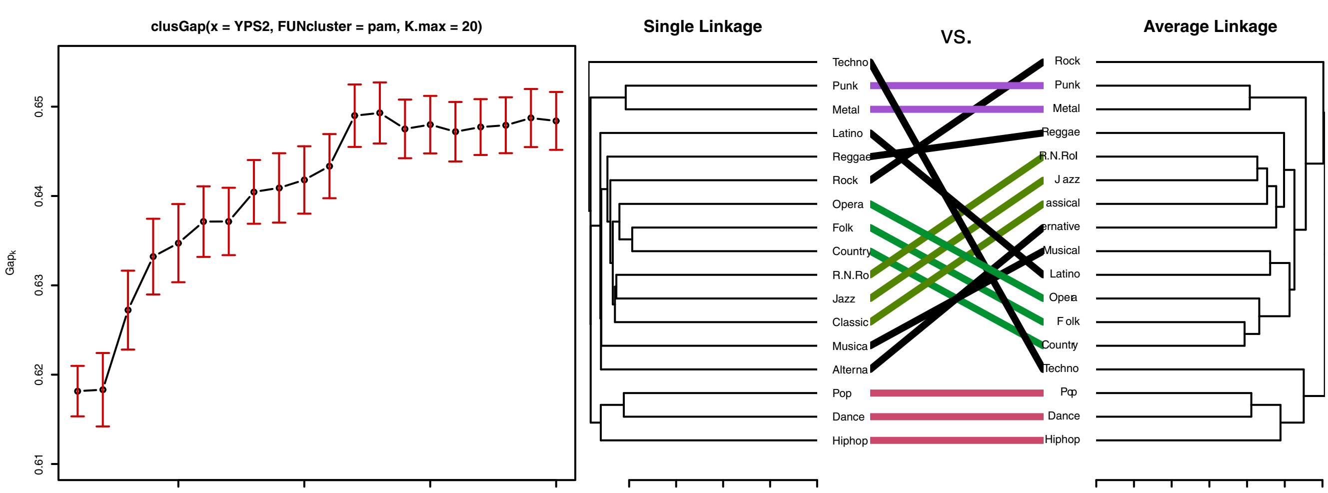

Fig. 3. Analysis of youth music, movie, and activity preferences.

data $\hat{y}_i$ will then be compared against the actual data y_i, either using a cross classification table (for class label predication), and/or some measure of error, for example the mean squared error (MSE), where $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$.

Most early class prediction papers in marketing utilized discriminant analysis (Fisher, 1936). In discriminant analysis, each item has a set of features $\mathbf{x}_i$ and a class label y_i. The data are split into groups using the class label y_i. In the binary case, a discriminant function line is drawn between the groups to maximize the variance between groups relative to the variance within groups. This "discriminant function" is defined as a weighted combination of the features of $\mathbf{x}_i$, so if $\mathbf{w}'\mathbf{x}_i > c$ the item i is classified into one group and if $\mathbf{w}'\mathbf{x}_i \leq c$, the item is classified into the other group. An initial application of discriminant analysis in marketing was to predict brand switching behavior (Farley, 1964) (switch/not switch) using predictors of quantity purchased, income, and family size. Frank, Massy, and Morrison (1965) describe applications for multiple discriminant analysis, where there may be more than two classes, e.g., one wishes to predict consumer choice between three brands A,B,C. They note that prediction performance on the sample may lead to overall bias and recommend either using holdout validation or adjusting the prediction success using the success predicted by chance. Further examples of the use of discriminant analysis include the prediction of consumer innovators (Robertson & Kennedy, 1968), new product purchasers (Pessemier, Burger, & Tigert, 1967), choice of retail facilities (Bucklin, 1967), business to business source loyalty (Wind, 1970) and private label brands (Burger & Schott, 1972). Issues surrounding the use of discriminant analysis are very similar to those found in market segmentation, including the selection of analysis variables and interpretation of results. Morrison (1969) notes that the probability of group membership calculated by discriminant analysis is a product of its likelihood ratio and its prior odds, so for a two label problem where $y_i \in \{0, 1\}$, the ratio of probabilities for group membership is given in (9).

$$\frac{p(y_i = 1|\mathbf{x}_i)}{p(y_i = 0|\mathbf{x}_i)} = \frac{p(\mathbf{x}_i|y_i = 1)}{p(\mathbf{x}_i|y_i = 0)} \times \frac{p(y_i = 1)}{p(y_i = 0)} \tag{9}$$

Here, the first term is the posterior ratio of group membership probabilities, the second term is the likelihood ratio, and the third term is the ratio of prior probabilities, which are operationalized as the proportion of items in each class. The log of likelihood ratio is the discriminant function $\mathbf{w}'\mathbf{x}_i$, which links discriminant analysis to logistic regression. Morrison (1969) further notes that the ratio of membership probabilities can be further altered by the economic cost of misclassifying items. For uneven classes, the ratio of prior probabilities can dominate the likelihood ratio and thus lead to all the items predicted to be in the larger class. A possible solution is to assign the items closest to the decision boundary to the smaller class, thus damping the effect of the uneven class memberships. The weight coefficients of the discriminant function can be interpreted in a similar way to regression. If the variables are not standardized then the coefficients give the absolute effect of one unit of each of the elements of $\mathbf{x}_i$ on the discriminant function.

5.2. Trends in data mining and statistical learning

Over the subsequent decades, a whole host of algorithms have been developed for class prediction. Many of these give better performance on certain problems than linear discriminant analysis. However, at their core, class prediction algorithms have two main purposes. The first is to "predict" the class labels and the second is to explain how the item (customer etc.) features inform the prediction result. Consider a brand loyalty example. Marketers may wish to pay extra attention to those customers who are predicted to churn. In the context of planning a marketing campaign, the actual prediction probabilities give more information than the class labels and may help with resource allocation. In addition, parameter estimates can be used to help relate $\mathbf{x}_i$ to the predicted y_i and give a broader context to the prediction decision.

The method of support vector machines (SVMs) (Cui & Curry, 2005) has found wide applicability in marketing prediction problems and can give performance increases over the logit/probit models that are often used in marketing prediction applications (Gensch & Recker, 1979). The intuition for SVMs is very similar to discriminant analysis. Here, $y_i \in \{-1, 1\}$ and a boundary line (or hyperplane) between classes is defined as $\mathbf{w}'\mathbf{x}_i - c = 0$. The value of $\mathbf{w}$ is then optimized using (10). The inner maximization equation tries to keep separation between classes, by defining a margin of size 1 and penalizing items that fall within the margin. The value of λ provides a trade off as to how strongly the margin is enforced.

$$\min\left[\frac{1}{n}\sum_{i=1}^{n} max\left(1 - y_i\left(\mathbf{w}'\mathbf{x}_i - c\right), 0\right)\right] + \lambda\|\mathbf{w}\|^2 \tag{10}$$

Both discriminant analysis[2] and SVMs utilize distances and the dot products in the distances $\langle \mathbf{x}, \mathbf{x}\rangle$ can be transformed by a "kernel" to give a nonlinear boundary between decision groups. For discriminant analysis and SVM, this "kernel trick" (Schölkopf, 2001), gives rise to non-linear decision boundaries and by selecting the kernel type (Gaussian, polynomial, linear, etc.) and tuning parameters, strong performance can be gained from a wide range of datasets.

A wide range of other methods have been applied to class prediction problems in marketing. These include probabilistic tree methods, such as CAR (Steinberg & Colla, 2009) and random forests (Breiman, 2001), neural networks (Zhang, 2000), nearest neighbor methods (Cover & Hart, 1967), stochastic gradient boosting (Zhang, 2004), and naïve Bayes (Lewis, 1998). For parsimony, full descriptions are left out in this review, but there are many excellent resources available, for example, Hastie et al. (2009), who discuss classification techniques from a statistical learning standpoint and Duda, Hart, and Stork (2002), who follow a purer machine learning/pattern recognition approach.

There are several major trends in class prediction. The first is the rise in the use of so called "ensemble" methods, where results from multiple algorithms are combined to give better results. This approach gets around the problem that for class prediction applications, relative algorithm performance can be dataset specific. In fact, several of the aforementioned techniques, random forests and stochastic gradient boosting, are predicated on combining multiple weak classifiers into a strong classifier. Another development has been for online prediction competitions on platforms such as https://www.Kaggle.com , where datasets from industry/science are uploaded for a specific prediction task, with competitors building models on training data and then having their solutions validated on a hidden test dataset. In this realm of practical prediction analysis, the best solutions often are "ensemble solutions". Common methods for creating ensemble solutions include the statistical pooling techniques of "boosting" and "bagging" (Bauer & Kohavi, 1999; Hoch, 2015; Puurula, Read, & Bifet, 2014). In addition, neural networks can be used to find the best set of weights for combining solutions (Kim, Kim, & Lee, 2003).

A second trend is the use of feature reduction and extraction techniques. Modern datasets can be extremely complex with hundreds of thousands of features or dimensions, many of which are mainly noise and don't improve prediction accuracy. Feature selec-

[2] Discriminant analysis, also has a quadratic variant, which occurs when the group variances are assumed to be unequal.

tion methods can improve the signal to noise ratio and reduce the size of the dataset, which improves scalability. The previously described dimensionality reduction technique of PCA, along with related methods, can be used as a precursor to or combined with class prediction techniques such as SVM (Ji & Ye, 2009). An alternate approach is to select the most useful features. This can be done using a wide range of criteria including high correlations with the class labels, low correlations with other predictor variables, predictor performance for single variable classifiers, and information theoretic measures (Guyon & Elisseeff, 2003). There are many bespoke feature selection algorithms for specific types of data. For example, in an interesting application, Moon and Russell (2008) take binary brand purchase data, utilize the "pick-any" joint space mapping technique previously referenced (Levine, 1979) and use the output dimensions from this technique to predict brand purchases.

5.3. Churn prediction

Recent marketing academic interest in class prediction techniques has been spurred by the availability of data in CRM systems (Ngai, Xiu, & Chau, 2009). Neslin, Gupta, Kamakura, Lu, and Mason (2006) describe the results of a tournament, where both academics and practitioners used a training dataset to build models for churn prediction, which were then evaluated on a test dataset. Several success metrics were reported, which all rely on items being ordered in order of the probability of churn, i.e. $p(y_i = 1)$. The decile lift gives the ratio of the proportion of churners in the top 10% of predicted churners to the proportion in the full dataset. The GINI coefficient, adapted from the GINI measure of economic inequality, gives a measure of how unequally distributed the actual churners are on the ordered list. The authors note that logit based models performed well on the dataset and outperformed discriminant and tree based techniques. The data mining aspects of churn management are linked to marketing work in CLV or customer lifetime value (Venkatesan & Kumar, 2004), to produce a profitability metric for a churn management campaign, where potential customers are offered an incentive to remain, which is given in (11).

$$Profit = N\alpha[(\gamma CLV + c_{IN}(1-\gamma))\beta_0\lambda - c_{CO} - c_{IN}] - c_{FX} \qquad (11)$$

Here, β_0 is the proportion of churners, λ is the "lift", i.e., the proportion of churners for the targeted customers divided by the proportion of churners across all customers, γ is the success of the incentive, i.e., the proportion of targeted consumers who remain loyal, c_{CO} is the cost of contacting consumers, c_{IN} is the cost of the incentive, and c_{FX} are fixed promotion costs. An extension is given by Verbraken, Verbeke, and Baesens (2013), who model profit using a parameterized beta distribution and from this create an EPMC (expected profit maximization criterion) as an evaluation metric for churn prediction, which can be used for feature selection (Maldonado, lvaro Flores, Verbraken, Baesens, & Weber, 2015).

Many of the described churn prediction algorithms can be applied to other scenarios, for example customer targeting prediction (Coussement, Harrigan, & Benoit, 2015) or yes/no recommendation prediction. With churn data, there can be a strong class imbalance problem, with only a few churners and many non-churners. Predictive accuracy can be improved either by under sampling non-churners (Burez & den Poel, 2009) or by oversampling churners (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Douzas & Bacao, 2017). As described previously, the usefulness of class prediction algorithms is predicated on both prediction accuracy and the ability to interpret model parameters. Bock and den Poel (2012) describe a framework based on an ensemble of additive logit models, and build in a set of feature importance scores to help interpret/select features. They implement graphs built using splines that show the probability function of churning and associated confidence intervals alongside a histogram of churn class distribution.

Corporate data utilized for churn prediction models have been annotated with derived sentiment data from emails (Coussement & den Poel, 2009) and from information available from company websites (D'Haen, den Poel, & Thorleuchter, 2013). Given increasingly strict data protection laws, particularly in Europe, it is sometimes necessary to delete past customer data, rendering it unamenable to analysis. To get around this problem, Holtrop, Wieringa, Gijsenberg, and Verhoef (2017) describe a method that only needs model parameters to estimate prediction models and uses Kalman filters to update the parameters as new data come in. Another issue is that training and implementation datasets may have differing distributions due to rapid changes in the business environment due to biased sampling. To deal with the first situation, Ballings and den Poel (2012) define and optimize a time window for which customer events are included in the prediction model. Xiao, Xiao, Huang, Liu, and Wang (2015) start with the assumption that training and test datasets have different distributions and employ a transfer learning (Pan & Yang, 2010) algorithm, which uses a neural network to match features between training and test datasets and combines this process with an ensemble classifier. To be used in practice for large scale consumer datasets, churn prediction must be scalable. Huang et al. (2015) describe the deployment of a large scale churn prediction system for a Chinese mobile operator with millions of active customers and multiple large databases, containing 2.3 terabytes of information. To achieve scalability, a tiered architecture was used with a data layer containing the databases, a "big data" layer consisting of data integration, modeling, and mining components, and an applications layer containing the business process functions, including churn prediction. The system is analyzed with respect to the volume, velocity, and variety of the data.

5.4. Software

R packages for class prediction are listed in Table 3. Several general classification packages are listed, including caret and RWeka, which both provide fully fledged environments for designing, building, and testing class prediction implementations. Basic discriminant analysis and SVM techniques are listed, along with implementations of tree methods and ensemble methods. The basic two class logit and probit models are included in the base R libraries. The mlogit and mprobit libraries are given for the multinomial logit and probit techniques. An example is given in Fig. 4, which shows a classification tree created with "rpart" for a bank marketing dataset.[3] The dependent variable is a response to a promotion (yes,no). The independent variables used to build the tree include duration of time the customer has held a bank account, job, martial status, month, and day of the month. At each tree branch, yes values are branched to the left and no values are branched to the right. The values displayed at each terminal node are the predicted response, the probability of a "yes" response, and the proportion of training instances captured at the node. For example, node 7 gives people whose last contact with the bank was at least 646 seconds long and who are married. These customers constitute 3% of the customer base and have a 62% chance of a "yes" response to the promotion.

5.5. Conclusions

There has been an explosion of interest in class prediction algorithms as of late. Competition websites such as

[3] Dataset available at https://archive.ics.uci.edu/ml/datasets/bank+marketing

Table 3
Segmentation and grouping packages in R.

CRAN	Reference	Description
asa	Culp, Johnson, and Michailides (2006)	Stochastic gradient boosting classification.
base	None	glm() implements logistic regression with family = binomial, binary logit with family = binomial(link = “logit”), and binary probit with family = binomial (link = “probit”).
Boruta	Kursa and Rudnicki (2010)	Feature selection using random forests to remove unimportant features.
caret	Kuhn (2008)	Access to class prediction algorithms, feature selection methods, and cross-validation procedures to evaluate performance. caretEnsemble allows an ensemble of caret classifiers.
class	Venables and Ripley (2003)	Contains a range of methods including SOMs, learning vector quantization and k-nearest neighbours.
e1071	Dimitriadou, Hornik, Leisch, Meyer, and Weingessel (2017)	Contains functions to implement SVMs and includes a choice of kernels.
FSelector	Cheng, Wang, and Bryant (2012)	Feature selection, including correlation, entropy, and chi-squared based methods.
MASS	Venables and Ripley (2003)	Contains lda() routine for linear discriminant analysis and qda() routine for quadratic discriminant analysis.
MCLUST	Fraley and Raftery (2003)	Gaussian model based clustering models and algorithms.
mlogit	Müllner (2013)	Implements a multinomial logit model.
mprobit	Müllner (2013)	Implements a multinomial probit model.
neuralnet	Gnther and Fritsch (2010)	Train and implement backpropogation neural networks.
randomForest	Liaw and Wiener (2002)	Creates random forest ensembles of weak classifiers.
ROSE	Lunardon, Menardi, and Torelli (2014)	Implements sampling methods for imbalanced class prediction.
rpart	Therneau and Atkinson (2017)	Functions to build classification (and regression) trees.
RWeka	Hall et al. (2009)	An R interface to Weka, a comprehensive data cleaning, feature selection, and prediction package.

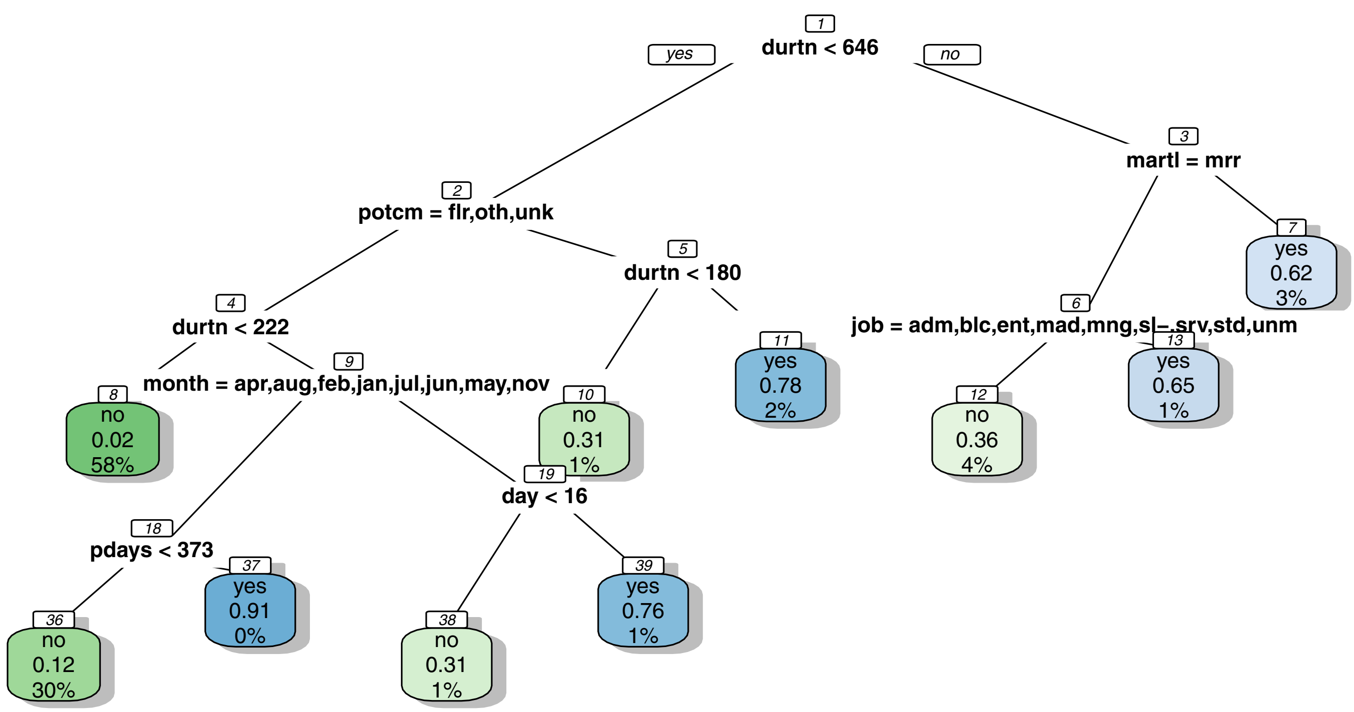

Fig. 4. Classification tree for bank marketing promotion (yes/no).

https://www.Kaggle.com and https://www.TopCoder.com, along with conferences such as KDD, regularly host class prediction competitions in which academics, industry practitioners, and hobbyists compete against each other in order to gain the winning prediction on a test dataset with models built from supplied training data. The rapid growth of data science as a field and its relative immaturity has contributed to this trend, with people trying to hone their skills through competition. Class prediction methods have been utilized in marketing from the 1960s onwards and have been applied to a range of problems, particularly in the areas of customer churn and predication. With recent academic interest in big data and analytics, there has been something of a revival in this area in marketing, spurred on by Neslin et al. (2006) and other recent work.

A few salient points can be elicited from this review. First, there is no one “silver bullet” when it comes to choosing a prediction technique. Algorithm performance varies across dataset type and characteristics. When implementing a new prediction application, it is advisable to not only examine past work on similar datasets, but also to cast a wide net and look at a range of different algorithms. Second, ensemble methods, which combine multiple classifiers into a single classifier tend to perform very well on a range of datasets and if computational power allows, make good starting points for algorithm implementation. Third, for large complex datasets with noisy data, some sort of feature selection may be required to keep the signal to noise ratio high and keep computational costs reasonable. Fourth, judging algorithms solely on prediction performance on a holdout dataset is taking a rather narrow view of performance. A very tiny incremental increase in prediction percentage may only improve profit by a tiny amount and does not take account of other factors, such as the algorithm runtime, the cost of implementing the algorithm, algorithm robustness, and the interpretability of the results. For example, “The Netflix Prize”, though not strictly a class prediction exercise, was one of the first large-scale online data competitions. A prize of $1,000,000 was awarded to the winning entry. Competition entrants needed to predict movie review scores for users based on their review scores for other movies. The final solution was an amalgamation of the many different techniques tried and shared among competition entrants. However, despite the fanfare, the winning solution was never actually implemented. An entry in the Netflix tech blog (Amatriain & Basilico, 2012) notes that “the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment” and that with a move towards online streaming content the style and format required for the recommendations had changed, which illustrates that empirical prediction performance is only one aspect of the overall system needs of a marketing analytics application.

6. Overall discussion

We have presented a review of marketing analytics that primarily covers the topics of visualization, segmentation and grouping, and class prediction. These topics were chosen as not only are they core to marketing strategy and have a long history in academic marketing, but are also of interest to researchers in expert systems, data mining, statistics, and operations research. There is a commonality throughout all three areas. In the 1960s and early 1970s there were a number of papers that took methodology from statistics and psychology and applied it to managerial marketing problems of positioning, segmentation, and response prediction. A core group of researchers in applied psychometrics and measurement, including J. Douglas Carroll, Ronald Frank, Paul Green, and Donald Morrison developed methods in all three of the fundamental areas of visualization, segmentation, and class prediction. This was part of an overall growth in interest in quantitative marketing in the 1960s, spearheaded by researchers such as Frank Bass and Andrew Ehrenberg who produced pioneering work in the areas of product diffusion (Bass, 1969) and consumer choice (Bass, 1974; Ehrenberg, 1959). The rapid growth of quantitative marketing in this era was spurred on by the availability of computational tools and data, an initiative by the Ford Foundation to equip business faculty with skills in mathematics and quantitative analysis, and the founding of the Marketing Science Institute to support the application of scientific techniques to marketing (Winer and Neslin, 2014, pp. 10–15).

There has been a steady stream of methodological work in the intervening period, but the last ten years have seen an explosion of interest in marketing analytics for several reasons. The first of these is the availability of data. Many of the papers cited in this review give extensions of basic methods, but are designed to deal with large, complex modern datasets, such as online reviews (including text), web-logs, CRM data, and brand scanner data. Much of this work is cross disciplinary, with researchers in different fields working with one another and citing one others’ work. This has partly been brought about by increased computational sophistication in statistics and increased interest in statistical methodology in expert systems and data mining.

Marketing models and methods do not exist in a vacuum. Most are implemented within computational software or systems. In fact, many advances in marketing analytics methodology and many innovative applications are published in the expert and intelligent systems literature, including in this journal. Our hope is by publishing this review in an expert systems journal and positioning it at the intersection of multiple fields, we have achieved several things. First, we have provided a historical context and managerial marketing insight to researchers in the expert systems field. Second, as much applied marketing segmentation research is now carried out in the expert systems field, we will have made researchers in the marketing domain aware of this work. Third, by including more theoretical work from statistics and operations research, some of this work can filter through to applied researchers.

Given the challenges and sophistication of modern data analysis problems, it is our view that this interdisciplinary approach will continue and strengthen over the coming decades, as solving practical problems will require a range of computational, statistical, and business skills. While marketing analytics implementations can achieve strong return on investment for businesses, success is correlated with a strong analytics infrastructure and culture, elements that need buy-in from multiple areas of a company, including marketing, IT, and senior decision makers (Germann, Lilien, & Rangaswamy, 2013).

Several authors have noted a disconnect between academics and practitioners in marketing, due to increased specialization and siloing of research (Reibstein, Day, & Wind, 2009), academic incentives that reward publications in academic journals rather than broader commercial impact (Lilien, 2011), and a hesitancy on the part of academics to engage with practitioners (Baron, Richardson, Earles, & Khogeer, 2011). However, outside of marketing, there is less worry about this disconnect, possibly as large numbers of Ph.D. graduates in areas such as statistics and expert systems go into industry, which helps narrow any academic-practitioner disconnect. We predict that there will be increasingly close interactions between academics and practitioners in marketing analytics. There are several reasons for this. First, increasing numbers of Ph.D. graduates are going into analytics/big data jobs in industry, where technical skills are at a premium, thus leading to an overlap in professional networks between academia and industry. Second, there has been a concerted effort from academics to engage with businesses and solve industry problems. Examples of this include the Wharton Customer Analytics Initiative, which works with industry partners to provide datasets and associated research opportunities for researchers, and well regarded practice prizes from the INFORMS and INFORMS marketing science communities, which are designed to reward research that has strong real world outcomes and impact (Lilien, Roberts, & Shankar, 2013). Third, business schools are increasingly emphasizing analytics at all levels of the curriculum and a range of pedagogical material has been developed to meet this need. For marketing analytics, in addition to classic books on marketing models (Lilien, Kotler, & Moorthy, 1991) and marketing engineering (Lilien & Rangaswamy, 2006), there are recent books on implementing marketing analytics using R (Chapman & Feit, 2015), building spreadsheet models for marketing analytics (Winston, 2014), and on marketing strategy aspects of data analytics (Palmatier & Sridhar, 2017). Fourth, as much of the technology associated with big data and analytics is new, there is a scramble to learn new techniques and methods, both from practitioners and academics. This has lead to a range of meet-up groups that are targeted to people wishing to learn new technologies, which attract both practitioners and academics. This phenomenon, along with the growth of data science competitions, which frequently feature marketing data, and the use of open-source software such as R and Python, has lead to a vibrant marketing analytics community, which encompasses both practitioners and academics. This bodes well for the future.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abbasi, A., France, S., Zhang, Z., & Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering, 23*(3), 447–462.
Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*, 734–749.
Akkucuk, U. (2004). *Nonlinear mapping: Approaches based on optimizing an index of continuity and applying classical metric MDS on revised distances, Ph.D. thesis*. Rutgers University, Newark, NJ.
Akkucuk, U., & Carroll, J. D. (2006). PARAMAP vs. Isomap: A comparison of two non-linear mapping algorithms. *Journal of Classification, 23*(2), 221–254.
Alsabti, K., Ranka, S., & Singh, V. (1998). *An efficient k-means clustering algorithm, Working paper*. Syracuse University. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.44.518
Amatriain, X., & Basilico, J. (2012). Netflix recommendations: Beyond the 5 stars. https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429.
Andrews, R. L., & Manrai, A. K. (1999). MDS maps for product attributes and market response: An application to scanner panel data. *Marketing Science, 18*(4), 584–604.
Arabie, P., Carroll, J. D., DeSarbo, W., & Wind, J. (1981). Overlapping clustering: A new method for product positioning. *Journal of Marketing Research, 18*(3), 310.
Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistical Surveys, 4*, 40–79.
Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. *Expert Systems with Applications, 111*, 11–34.
Baglama, J. (2016). IRLBA: Fast partial singular value decomposition method. In *Handbook of Big Data* (pp. 125–136)). Boca Raton, FL: CRC Press.
Ballings, M., & den Poel, D. V. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications, 39*(18), 13517–13522.
Ballings, M., & den Poel, D. V. (2015). CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research, 244*(1), 248–260.
Banerjee, A., Krumpelman, C., Basu, S., & Mooney, R. J. (2005). Model-based overlapping clustering. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining* (pp. 532–537). New York, NY: ACM Press.
Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics, 49*(3), 803–821.
Baron, S., Richardson, B., Earles, D., & Khogeer, Y. (2011). Marketing academics and practitioners: Towards togetherness. *Journal of Customer Behaviour, 10*(3), 291–304.
Barron, J., & Hollingshead, J. (2002). Making segmentation work. *Marketing Management, 11*(1), 24–28.
Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science, 15*(5), 215–227.
Bass, F. M. (1974). The theory of stochastic preference and brand switching. *Journal of Marketing Research, 11*(1), 1–20.
Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36*(1), 105–139.
Bell, D. R., & Song, S. (2007). Neighborhood effects and trial on the internet: Evidence from online grocery retailing. *Quantitative Marketing and Economics, 5*(4), 361–400.
Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing, 12*(1), 17–30.
Bezdek, J. C., Ehrlich, R., & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2), 191–203.
Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R*: 10 (2nd). New York, NY: Springer.
Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(4), 993–1022.
Bock, K. W. D., & den Poel, D. V. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications, 39*(8), 6816–6826.
Bodea, T., & Ferguson, M. (2014). *Segmentation, revenue management and pricing analytics* (1st). New York, NY: Routledge.
Boone, D. S., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International Journal of Research in Marketing, 19*(3), 287–301.
Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd). New York, NY: Springer.
Bradlow, E. T., Bronnenberg, B., Russell, G. J., Arora, N., Bell, D. R., Duvvuri, S. D., et al. (2005). Spatial models in marketing. *Marketing Letters, 16*(3), 267–278.
Bramer, M. A. (1982). A survey and critical review of expert systems research. In *Introductory Readings in Expert Systems* (pp. 3–29)). New York, NY: Gordon and Breach, Science Publishers, Inc.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Brida, J. G., Disegna, M., & Osti, L. (2012). Segmenting visitors of cultural events by motivation: A sequential non-linear clustering analysis of Italian Christmas market visitors. *Expert Systems with Applications, 39*(13), 11349–11356.
Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing, 36*(Supplement C), 93–100.
Brodbeck, D., & Girardin, L. (2003). Visualization of large-scale customer satisfaction surveys using a parallel coordinate tree. In T. Munzner, & S. North (Eds.), *Ieee symposium on information visualization 2003* (pp. 197–201).
Brown, S. (1996). Art or science?: Fifty years of marketing debate. *Journal of Marketing Management, 12*(4), 243–267.
Brunsdon, C., & Comber, L. (2015). *An introduction to R for spatial analysis and mapping* (1st). London, UK: Sage.
Brusco, M. J., Cradit, J. D., Steinley, D., & Fox, G. L. (2008). Cautionary remarks on the use of clusterwise regression. *Multivariate Behavioral Research, 43*(1), 29–49.
Brusco, M. J., Cradit, J. D., & Tashchian, A. (2003). Multicriterion clusterwise regression for joint segmentation settings: An application to customer value. *Journal of Marketing Research, 40*(2), 225–234.
Bucklin, L. P. (1967). The concept of mass in intra-urban shopping. *Journal of Marketing, 31*(4), 37–42.
Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*(2), 201–215.
Burez, J., & den Poel, D. V. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626–4636.
Burger, P. C., & Schott, B. (1972). Can private brand buyers be identified? *Journal of Marketing Research, 9*(2), 219–222.
Caliski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1–27.
Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling (Presidential address for Psychometric Society). *Psychometrika, 41*(4), 439–463.
Carroll, J. D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 48*(2), 157–169.
Carroll, J. D., & Chang, J.-J. (1964). A general index of nonlinear correlation and its application to the problem of relating physical and psychological dimensions. *American Psychologist, 19*(7), 540.
Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart–Young" decomposition. *Psychometrika, 35*(3), 283–319.
Carroll, J. D., & Green, P. E. (1995). Psychometric methods in marketing research: Part I, conjoint analysis. *Journal of Marketing Research, 32*(3), 385–391.
Carroll, J. D., & Green, P. E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research, 34*(2), 193–204.
Casabayó, M., Agell, N., & Sánchez-Hernández, G. (2015). Improved market segmentation by fuzzifying crisp clusters: A case study of the energy market in Spain. *Expert Systems with Applications, 42*(3), 1637–1643.
Chapman, C., & Feit, E. M. (2015). *R for marketing research and analytics* (1st). New York, NY: Springer.
Chaturvedi, A., & Carroll, J. D. (2001). Deriving market structures via additive decomposition of market shares (application of three-way generalized SINDCLUS).
Chaturvedi, A., & Carroll, J. D. (2006). CLUSCALE ("CLUstering and multidimensional SCAL[E]ing"): A three-way hybrid model incorporating overlapping clustering and multidimensional scaling structure. *Journal of Classification, 23*(2), 269–299.
Chaturvedi, A., Carroll, J. D., Green, P. E., & Rotondo, J. A. (1997). A feature-based approach to market segmentation via overlapping k-centroids clustering. *Journal of Marketing Research, 34*(3), 370–377.
Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). Multivariate analysis of mixed data: The PCAmixdata R package. arXiv:1411.4911
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.
Chen, L. (2006). *Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis*. University of Pennsylvania Ph.D. dissertation.
Chen, L., & Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association, 104*(486), 209–219.
Chen, Y., Fu, C., & Zhu, H. (2008). A data mining approach to customer segment based on customer value. In J. Ma, Y. Yin, J. Yu, & S. Zhou (Eds.), *2008 fifth international conference on fuzzy systems and knowledge discovery: 4* (pp. 513–517). Picastaway,NJ: IEEE.
Cheng, T., Wang, Y., & Bryant, S. H. (2012). FSelector: A Ruby gem for feature selection. *Bioinformatics, 28*(21), 2851–2852.
Chiang, M. M.-T., & Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *Journal of Classification, 27*(1), 3–40.
CMO (2016). CMO Survey Report: Highlights and Insights, Aug. 2016. https://cmosurvey.org/wp-content/uploads/sites/11/2016/08/The_CMO_Survey-Highlights_and_Insights-Aug-2016.pdf.

Commenges, H., & Giraud, T. (2017). Introduction to the SpatialPosition package. https://cran.r-project.org/web/packages/SpatialPosition/vignettes/SpatialPosition.html.
Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications, 42*(22), 8403–8412.
Coussement, K., & den Poel, D. V. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications, 36*(3, Part 2), 6127–6134.
Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.
Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science, 24*(4), 595–615.
Culp, M., Johnson, K., & Michailides, G. (2006). ada: An R package for stochastic boosting. *Journal of Statistical Software, 17*(2).
Dahlhaus, E. (2000). Parallel algorithms for hierarchical clustering and applications to split decomposition and parity graph recognition. *Journal of Algorithms, 36*(2), 205–240.
Dekimpe, M. G., & Hanssens, D. M. (2000). Time-series models in marketing:: Past, present and future. *International Journal of Research in Marketing, 17*(2), 183–193.
DeSarbo, W. S. (1982). GENNCLUS: New models for general nonhierarchical clustering analysis. *Psychometrika, 47*(4), 449–475.
DeSarbo, W. S., & Cho, J. (1989). A stochastic multidimensional scaling vector threshold model for the spatial representation of "pick any/n" data. *Psychometrika, 54*(1), 105–129.
DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification, 5*(2), 249–282.
DeSarbo, W. S., & Edwards, E. A. (1996). Typologies of compulsive buying behavior: A constrained clusterwise regression approach. *Journal of Consumer Psychology, 5*(3), 231–262.
DeSarbo, W. S., & Hoffman, D. L. (1987). Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research, 24*(1), 40–54.
Desarbo, W. S., & Manrai, A. K. (1992). A new multidimensional scaling methodology for the analysis of asymmetric proximity data in marketing research. *Marketing Science, 11*(1), 1–20.
DeSarbo, W. S., Young, M. R., & Rangaswamy, A. (1997). A parametric multidimensional unfolding procedure for incomplete nonmetric preference/choice set data in marketing research. *Journal of Marketing Research, 34*(4), 499–516.
D'Haen, J., den Poel, D. V., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications, 40*(6), 2007–2012.
Dhalla, N. K., & Mahatoo, W. H. (1976). Expanding the scope of segmentation research. *Journal of Marketing, 40*(2), 34–41.
Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2017). Support Vector Machines: The interface to libsvm in package e1071. https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.
Dixon, P., & Palmer, M. W. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science, 14*(6), 927–930.
Douard, J.-P., Heitz, M., & Cliquet, G. (2015). Retail attraction revisited: From gravitation to purchase flows, a geomarketing application. *Recherche et Applications en Marketing (English Edition), 30*(1), 110–129.
Douzas, G., & Bacao, F. (2017). Self-organizing map oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications, 82*, 40–52.
Duda, R. O., Hart, P. E., & Stork, D. G. (2002). *Pattern classification* (2nd). New York, NY: John Wiley & Sons.
Dwyer, P. (2012). Inferring brand proximities from user-generated content. *Journal of Brand Management, 19*(6), 467–483.
Ehrenberg, A. S. C. (1959). The pattern of consumer purchases. *Journal of the Royal Statistical Society.Series C (Applied Statistics), 8*(1), 26–41. doi:10.2307/2985810.
Eliashberg, J., & Manrai, A. K. (1992). Optimal positioning of new product-concepts: Some analytical implications and empirical results. *European Journal of Operational Research, 63*(3), 376–397.
Elkan, C. (2003). Using the triangle inequality to accelerate k-means. In T. Fawcett, & N. Mishra (Eds.), *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 147–153). Palo Alto, CA: AAAI Press.
Elrod, T. (1988). Choice map: Inferring a product-market map from panel data. *Marketing Science, 7*(1), 21–40.
Farley, J. U. (1964). "Brand loyalty" and the economics of information. *The Journal of Business, 37*(4), 370–381.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188.
Forsyth, R. (1984). The architecture of expert systems. In *Expert systems: principles and case studies* (pp. 9–17). Cambrdige, UK: CRC Press.
Fraley, C., Raftery, A., & Wehrens, R. (2005). Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics, 14*(3), 529–546.
Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631.
Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation,and discriminant analysis software: MCLUST. *Journal of Classification, 20*(2), 263–286.
France, S. L., & Carroll, J. D. (2007). Development of an agreement metric based upon the Rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. . In P. Perner (Ed.), *Proceedings of MLDM 2007* (pp. 499–517). Heidelberg, Germany: Springer.
France, S. L., & Carroll, J. D. (2009). Visualizing the competitive structure of online auctions. In P. Perner (Ed.), *Proceedings of ICDM 2009* (pp. 103–116). Heidelberg, Germany: Springer.
France, S. L., & Ghose, S. (2016). An analysis and visualization methodology for identifying and testing market structure. *Marketing Science, 35*(1), 182–197.
France, S. L., Vaghefi, M. S., & Zhao, H. (2016). Characterizing viral videos: Methodology and applications. *Electronic Commerce Research and Applications, 19*(Supplement C), 19–32.
Frank, R. E., & Green, P. E. (1968). Numerical taxonomy in marketing analysis: A review article. *Journal of Marketing Research, 5*(1), 83–94.
Frank, R. E., Massy, W. F., & Morrison, D. G. (1965). Bias in multiple discriminant analysis. *Journal of Marketing Research, 2*(3), 250–258.
Galili, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics, 31*(22), 3718–3720.
Gensch, D. H., & Recker, W. W. (1979). The multinomial, multiattribute logit choice model. *Journal of Marketing Research, 16*(1), 124–132.
Germann, F., Lilien, G. L., & Rangaswamy, A. (2013). Performance implications of deploying marketing analytics. *International Journal of Research in Marketing, 30*(2), 114–128.
Giordani, P., Ferraro, M. B., & Giordani, M. P. (2015). Package fclust. https://cran.r-project.org/web/packages/fclust/.
Green, P., Maheshwari, A., & Rao, V. (1969). Dimensional interpretation and configuration invariance in multidimensional scaling: An empirical study. *Multivariate Behavioral Research, 4*(2), 159–180.
Green, P. E., & Carmone, F. J. (1969). Multidimensional scaling: An introduction and comparison of nonmetric unfolding techniques. *Journal of Marketing Research, 6*(3), 330–341.
Green, P. E., & Carroll, J. D. (1988). A simple procedure for finding a composite of several multidimensional scaling solutions. *Journal of the Academy of Marketing Science, 16*(1), 25–35.
Green, P. E., Carroll, J. D., & Carmone, F. J. (1977). Design considerations in attitude measurement. In *Moving ahead with attitude research. Chicago, Illinois: American Marketing Association* (pp. 9–18).
Green, P. E., Frank, R. E., & Robinson, P. J. (1967). Cluster analysis in test market selection. *Management Science, 13*(8), 387–400.
Green, P. E., & Rao, V. R. (1969). A note on proximity measures and cluster analysis. *Journal of Marketing Research, 6*(3), 359–364.
Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*(2), 103–123.
Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing, 54*(4), 3–19.
Griffith, D. A. (1982). A generalized Huff model. *Geographical Analysis, 14*(2), 135–144.
Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications, 100*, 1–16.
Grover, R., & Srinivasan, V. (1987). A simultaneous approach to market segmentation and market structure. *Journal of Marketing Research, 24*(2), 139–153.
Grover, R., & Srinivasan, V. (1989). An approach for tracking within-segment shifts in market shares. *Journal of Marketing Research, 26*(2), 230–236.
Grover, R., & Srinivasan, V. (1992). Evaluating the multiple effects of retail promotions on brand loyal and brand switching segments. *Journal of Marketing Research, 29*(1), 76–89.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182.
Gnther, F., & Fritsch, S. (2010). Neuralnet: Training of neural networks. *The R Journal, 2*(1), 30–38.
Haines, G. H., Simon, L. S., & Alexis, M. (1972). Maximum likelihood estimation of central-city food trading areas. *Journal of Marketing Research, 9*(2), 154–159.
Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *Journal of Marketing, 32*(3), 30–35.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletter, 11*(1), 10–18.
Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52*(2), 112–118.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd). New York, NY: Springer.
Hauser, J. R., & Shugan, S. M. (1983). Defensive marketing strategies. *Marketing Science, 2*(4), 319–360.
Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Queue, 10*(2), 30–55.
Hennig, C. (2015). fpc: Flexible procedures for clustering. https://cran.r-project.org/web/packages/fpc/fpc.pdf.
Hernández, T. (2007). Enhancing retail location decision support: The development and application of geovisualization. *Journal of Retailing and Consumer Services, 14*(4), 249–258.
Hernández, T., & Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail & Distribution Management, 28*(8), 357–367.
Ho, Y., Chung, Y., & Lau, K.-N. (2010). Unfolding large-scale marketing data. *International Journal of Research in Marketing, 27*(2), 119–132.
Hoch, T. (2015). An ensemble learning approach for the Kaggle taxi travel time prediction challenge. In A. Martínez-Usó, J. Mendes-Moreira, L. Moreira-Matias, M. Kull, & N. Lachiche (Eds.), *Proceedings of the 2015th international conference on ecml pkdd discovery challenge - volume 1526* (pp. 52–62). Aachen, Germany: CEUR-WS.org.

Hoffman, D. L., & Franke, G. R. (1986). Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research, 23*(3), 213–227.
Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing, 34*(1), 154–172.
Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30.
Hruschka, H. (1986). Market definition and segmentation using fuzzy clustering methods. *International Journal of Research in Marketing, 3*(2), 117–134.
Hruschka, H., & Natter, M. (1999). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research, 114*(2), 346–353.
Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., et al. (2015). Telco churn prediction with big data. In S. B. Davidson, & Z. Ives (Eds.), *Proceedings of the 2015 ACM SIGMOD international conference on management of data SIGMOD '15*. New York, NY, USA: ACM.
Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304.
Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
Huff, D. L. (1964). Defining and estimating a trading area. *Journal of Marketing, 28*(3), 34–38.
Huff, D. L. (2003). Parameter estimation in the Huff model. *ESRI, ArcUser*, 34–36.
Hummel, M., Edelmann, D., & Kopp-Schneider, A. (2017). CluMix: Clustering and visualization of mixed-type data. https://pdfs.semanticscholar.org/1e65/755051c4b749fac17a23ff93924157acacdd.pdf.
Iacobucci, D. (1996). *Networks in marketing* (1st). Thousand Oaks, CA: Sage.
Iacobucci, D., Grisaffe, D., & DeSarbo, W. (2017). Statistical perceptual maps: Using confidence region ellipses to enhance the interpretations of brand positions in multidimensional scaling. *Journal of Marketing Analytics, 5*(3), 81–98.
IDC (2016). IDC's worldwide semiannual big data and analytics spending guide.
IDG (2016). IDG 2016 Data & Analytics Executive Summary. http://www.idgenterprise.com/resource/research/tech-2016-data-analytics-research/.
Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314.
Inselberg, A., & Dimsdale, B. (1987). Parallel coordinates for visualizing multidimensional geometry. In *Computer Graphics 1987: Proceedings of CG International 87* (pp. 25–44). Tokyo: Springer Japan.
Jain, D., Bass, F. M., & Chen, Y.-M. (1990). Estimation of latent class models with heterogeneous choice probabilities: An application to market structuring. *Journal of Marketing Research, 27*(1), 94–101.
Ji, S., & Ye, J. (2009). Linear dimensionality reduction for multi-label classification. In H. Kitano (Ed.), *Proceedings of the 21st international jont conference on artifical intelligence IJCAI'09* (pp. 1077–1082). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
Jockers, M. (2017). Package 'syuzhet'. https://cran.r-project.org/web/packages/syuzhet.
Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.
Kamakura, W. A., & Russell, G. J. (1989). A probilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research, 26*(4), 379–390.
Kamgar-Parsi, B., Gualtieri, J. A., Devaney, J. E., & Kamgar-Parsi, B. (1990). Clustering with neural networks. *Biological Cybernetics, 63*(3), 201–208.
Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881–892.
Karypis, G. (2002). CLUTO-a clustering toolkit. http://www.glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/manual.pdf.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis* (1st). Hoboken, NJ: Wiley-Interscience.
Keim, D. A., Hao, M. C., Dayal, U., & Hsu, M. (2002). Pixel bar charts: A visualization technique for very large multi-attribute data sets. *Information Visualization, 1*(1), 20–34.
Kim, E., Kim, W., & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems, 34*(2), 167–175.
Klemz, B. R., & Dunne, P. M. (2000). Exploratory analysis using parallel coordinate systems: Data visualization in N-dimensions. *Marketing Letters, 11*(4), 323–333.
Koga, H., Ishibashi, T., & Watanabe, T. (2007). Fast agglomerative hierarchical clustering algorithm using locality-sensitive hashing. *Knowledge and Information Systems, 12*(1), 25–53.
Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. *Computer, 46*(5), 44–50.
Kotler, P. T., & Keller, K. L. (2015). *Marketing management* (15th). Hoboken, NJ: Pearson.
Kraemer, G. (2017). Package 'dimred'. https://cran.r-project.org/web/packages/dimRed/.
Kriegel, H. P., Kroger, P., Pryakhin, A., & Schubert, M. (2005). Effective and efficient distributed model-based clustering. In J. Han, B. W. Wah, V. Raghavan, X. Wu, & R. Rastogi (Eds.), *Fifth IEEE international conference on data mining (ICDM'05)* (pp. 1–8). Picastaway,NJ: IEEE.
Kruskal, J. B. (1964). Multidimensional scaling for optimizing a goodness of fit metric to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.
Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (1st). Newbury Park, USA: Sage Publications.
Kuha, J. (2004). AIC and BIC. *Sociological Methods & Research, 33*(2), 188–229.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5).
Kursa, M., & Rudnicki, W. (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11), 1–13.
Lazer, W. (1963). Life style concepts and marketing. In S. A. Greyser (Ed.), *Toward scientific marketing: Proceedings of the winter conference of the American marketing association: 12* (pp. 130–139). Boston, MA: American Marketing Association.
Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1–18.
Lee, A. J. T., Yang, F.-C., Chen, C.-H., Wang, C.-S., & Sun, C.-Y. (2016). Mining perceptual maps from consumer reviews. *Decision Support Systems, 82*, 12–25.
Lee, J. A., & Verleysen, M. (2009). Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing, 72*(7–9), 1431–1443.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research, 48*(5), 881–894.
de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software, 31*(3).
Leisch, F. (2006). A toolbox for K-centroids cluster analysis. *Computational Statistics & Data Analysis, 51*(2), 526–544.
Levine, J. H. (1979). Joint-space analysis of "pick-any" data: Analysis of choices from an unconstrained set of alternatives. *Psychometrika, 44*(1), 85–92.
Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings* (pp. 4–15)). Heidelberg, Germany: Springer.
Li, X. (1990). Parallel algorithms for hierarchical clustering and cluster validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(11), 1088–1092.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.
Lilien, G. L. (2011). Bridging the academic-practitioner divide in marketing decision models. *Journal of Marketing, 75*(4), 196–210.
Lilien, G. L., Kotler, P. K., & Moorthy, S. (1991). *Marketing models* (1st). Upper Saddle River, NJ: Prentice Hall.
Lilien, G. L., & Rangaswamy, A. (2006). *Marketing engineering: Computer-assisted marketing analysis and planning, revised second edition* (2nd). Bloomington,IN: Trafford Publishing.
Lilien, G. L., Roberts, J. H., & Shankar, V. (2013). Effective marketing science applications: Insights from the ISMS-MSI practice prize finalist papers and projects. *Marketing Science, 32*(2), 229–245.
Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–23.
Little, J. D. C. (1979). Decision support systems for marketing managers. *Journal of Marketing, 43*(3), 9–26.
Lovelace, R., Birkin, M., Cross, P., & Clarke, M. (2016). From big noise to big data: Toward the verification of large data sets for understanding regional retail flows. *Geographical Analysis, 48*(1), 59–81.
Lueks, W., Mokbel, B., Biehl, M., & Hammer, B. (2011). How to evaluate dimensionality reduction?. In B. Hammer, & T. Villmann (Eds.), *Proceedings of the workshop - new challenges in neural computation 2011: 5* (pp. 29–37).
Lunardon, N., Menardi, G., & Torelli, N. (2014). Rose: A package for binary imbalanced learning. *R Journal, 6*(1).
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: 1* (pp. 281–297). Oakland, CA, USA.
Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*(1), 36–43.
Mahajan, V., & Jain, A. K. (1978). An approach to normative segmentation. *Journal of Marketing Research, 15*(3), 338–345.
Maldonado, S., lvaro Flores, Verbraken, T., Baesens, B., & Weber, R. (2015). Profit-based feature selection using support vector machines general framework and an application for customer retention. *Applied Soft Computing, 35*(Supplement C).
McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions*: 382 (2nd). Hoboken, NJ: John Wiley & Sons.
McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification, 33*(3), 331–373.
McNicholas, P. D., Murphy, T. B., McDaid, A. F., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics & Data Analysis, 54*(3), 711–723.
Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*(3), 325–342.
Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16*(3), 379–407.
Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In *Clustering and classification* (pp. 341–375). Hackensack, NJ: World Scientific.
Montgomery, A. L. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science, 16*(4), 315–337.
Moon, S., & Kamakura, W. A. (2017). A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing, 34*(1), 265–285.
Moon, S., & Russell, G. J. (2008). Predicting product purchase from inferred customer similarity: An autologistic model approach. *Management Science, 54*(1), 71–82.

Morey, L. C., Blashfield, R. K., & Skinner, H. A. (1983). A comparison of cluster analysis techniques withing a sequential validation framework. *Multivariate Behavioral Research, 18*(3), 309–329.
Morrison, D. G. (1967). Measurement problems in cluster analysis. *Management Science, 13*(12), B775–B780.
Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research, 6*(2), 156–163.
Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software, 53*(9), 1–18.
Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R* (1st). Boca Raton, USA: Chapman & Hall.
Neidell, L. A. (1969). The use of nonmetric multidimensional scaling in marketing analysis. *Journal of Marketing, 33*(4), 37–43.
Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20*(3), 1–13.
Neslin, S. A., Gupta, S., Kamakura, W. A., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2, Part 2), 2592–2602.
Oelke, D., Hao, M., Rohrdantz, C., Keim, D. A., Dayal, U., Haug, L. E., et al. (2009). Visual opinion analysis of customer feedback data. In J. Stasko, & J. J. van Wijk (Eds.), *2009 IEEE symposium on visual analytics science and technology* (pp. 187–194).
Olshavsky, R. W., MacKay, D. B., & Sentell, G. (1975). Perceptual maps of supermarket locations. *Journal of Applied Psychology, 60*(1), 80–86.
Olson, C. F. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing, 21*(8), 1313–1325.
Palmatier, R., & Sridhar, S. (2017). *Marketing strategy: Based on first principles and data analytics* (1st). Basingstoke, UK: Red Globe Press.
Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.
Park, S., Suresh, N. C., & Jeong, B.-K. (2008). Sequence-based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering, 65*(3), 512–543.
Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing, 69*(4), 167–176.
Pessemier, E. A., Burger, P. C., & Tigert, D. J. (1967). Can new product buyers be identified? *Journal of Marketing Research, 4*(4), 349–354.
Plummer, J. T. (1974). The concept and application of life style segmentation. *Journal of Marketing, 38*(1), 33–37.
Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148.
Puurula, A., Read, J., & Bifet, A. (2014). Kaggle LSHTC4 winning solution. arXiv preprint arXiv:1405.0546.
Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika, 43*(2), 145–160.
Reibstein, D. J., Day, G., & Wind, J. (2009). Guest editorial: Is marketing academia losing its way? *Journal of Marketing, 73*(4), 1–3.
ReVelle, C. S., & Eiselt, H. A. (2005). Location analysis: A synthesis and survey. *European Journal of Operational Research, 165*(1), 1–19.
Rexer, K. (2016). Rexler Analytics 2015 data science survey. http://www.rexeranalytics.com/data-science-survey.html.
Richman, M. B. (1986). Rotation of principal components. *Journal of Climatology, 6*(3), 293–335.
Ringel, D. M., & Skiera, B. (2016). Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science, 35*(3), 511–534.
Roberts, J. H., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing, 31*(2), 127–140.
Robertson, T. S., & Kennedy, J. N. (1968). Prediction of consumer innovators: Application of multiple discriminant analysis. *Journal of Marketing Research, 5*(1), 64–69.
Roig-Tierno, N., Baviera-Puig, A., Buitrago-Vera, J., & Mas-Verdu, F. (2013). The retail site location decision process using GIS and the analytical hierarchy process. *Applied Geography, 40*(Supplement C), 191–198.
Rosen, J. (2012). Who do online advertisers think you are? *The New York Times*.
Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science, 22*(3), 304–328.
Russell, G. J., & Kamakura, W. A. (1994). Understanding brand competition using micro and macro scanner data. *Journal of Marketing Research, 31*(2), 289–303.
Sallam, R. L., Howson, C., Idoine, C. J., Oestreich, T. W., Richardson, J. L., & Tapadinhas, J. (2017). Magic quadrant for business intelligence and analytics platforms. *Gartner RAS core Research Notes*. Gartner, Stamford, CT.
Schölkopf, B. (2001). The kernel trick for distances. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13, NIPS 2001* (pp. 301–307). Cambridge, MA: MIT Press.
Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*(2), 87–123.
Shmueli, G., & Jank, W. (2005). Visualizing online auctions. *Journal of Computational & Graphical Statistics, 14*(2), 299–319.
Shocker, A. D., & Srinivasan, V. (1974). A consumer-based methodology for the identification of new product ideas. *Management Science, 20*(6), 921–937.
Shugan, S. M. (1987). Estimating brand positioning maps using supermarket scanning data. *Journal of Marketing Research, 24*(1), 1–18.
Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software, 1*(3).
Silva-Risso, J. M., Bucklin, R. E., & Morrison, D. G. (1999). A decision support system for planning manufacturers' sales promotion calendars. *Marketing Science, 18*(3), 274–300.
Sisodia, R. S. (1992). Marketing information and decision support systems for services. *Journal of Services Marketing, 6*(1), 51–64.
Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing, 21*(1), 3–8.
Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing, 22*(4), 367–373.
Stanley, T. J., & Sewall, M. A. (1976). Image inputs to a probabilistic model: Predicting retail potential. *Journal of Marketing, 40*(3), 48–53.
Steinberg, D., & Colla, P. (2009). CART: Classification and regression trees. In *The Top Ten Algorithms in Data Mining* (pp. 179–201)). Boca Raton, FL: CRC Press.
Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*(1), 1–34.
Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological methods, 16*(1), 63–79.
Stoffel, K., & Belkoniene, A. (1999). Parallel k/h-means clustering for large data sets. In *Euro-Par'99 Parallel Processing: 5th International Euro-Par Conference Toulouse, France, August 31, – September 3, 1999 Proceedings* (pp. 1451–1454)). Berlin, Heidelberg: Springer.
Taylor, W. J. (1965). "Is marketing a science?" Revisited. *Journal of Marketing, 29*(3), 49–53.
Therneau, T. M., & Atkinson, E. J. (2017). An introduction to recursive partitioning using the RPART routines. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.
Thieme, R. J., Song, M., & Calantone, R. J. (2000). Artificial neural network decision support systems for new product development project selection. *Journal of Marketing Research, 37*(4), 499–507.
Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research, 51*(4), 463–479.
Torres, A., & Bijmolt, T. H. A. (2009). Assessing brand image through communalities and asymmetries in brand-to-attribute and attribute-to-brand associations. *European Journal of Operational Research, 195*(2), 628–640.
Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook* (pp. 667–685)). New York, NY: Springer US.
Tukey, J. W. (1977). *Exploratory data analysis* (1st). Reading, MA: Addison-Wesley.
Tuma, M. N., Decker, R., & Scholz, S. W. (2011). A survey of the challenges and pitfalls of cluster analysis application in market segmentation. *International Journal of Market Research, 53*(3), 391–414.
Urban, G. L., Johnson, P. L., & Hauser, J. R. (1984). Testing competitive market structures. *Marketing Science, 3*(2), 83–112.
Vellido, A., Lisboa, P. J. G., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications, 17*(4), 303–314.
Venables, W. N., & Ripley, B. D. (2003). *Modern applied statistics with S* (4th). New York, NY: Springer.
Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing, 68*(4), 106–125.
Venkatraman, V., Clithero, J. A., Fitzsimons, G. J., & Huettel, S. A. (2012). New scanner data for brand marketers: How neuroscience can help better understand differences in brand preferences. *Journal of Consumer Psychology, 22*(1), 143–153.
Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering, 25*(5), 961–973.
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243.
Wagner, W. P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert Systems with Applications, 76*, 85–96.
Waite, A. (2007). HR's role in audience segmentation: How employee segmentation can help HR create tailored programs for its different employee groups. *Strategic HR Review, 6*(2), 16–19.
Wansbeek, T., & Wedel, M. (1998). Marketing and econometrics: Editors' introduction. *Journal of Econometrics, 89*(1), 1–14.
Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.
Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121.
Wedel, M., & Steenkamp, J.-B. E. M. (1989). A fuzzy clusterwise regression approach to benefit segmentation. *International Journal of Research in Marketing, 6*(4), 241–258.

Wedel, M., & Steenkamp, J.-B. E. M. (1991). A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation. *Journal of Marketing Research, 28*(4), 385–396.

Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: The Kohonen package. *Journal of Statistical Software, 21*(5), 1–19.

Wehrens, R., Buydens, L. M. C., Fraley, C., & Raftery, A. E. (2004). Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification, 21*(2), 231–253.

Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications, 39*(5), 5529–5533.

Wells, W. D. (1975). Psychographics: A critical review. *Journal of Marketing Research, 12*(2), 196–213.

Wells, W. D., & Tigert, D. J. (1971). Activities, interests and opinions. *Journal of Advertising Research, 11*(4), 27–35.

Wind, Y. (1970). Industrial source loyalty. *Journal of Marketing Research, 7*(4), 450–457.

Wind, Y. (1978). Issues and advances in segmentation research. *Journal of Marketing Research, 15*(3), 317–337.

Wind, Y., & Douglas, S. P. (1972). International market segmentation. *European Journal of Marketing, 6*(1), 17–25.

Winer, R. S., & Neslin, S. A. (2014). *The history of marketing science* (1st). Singapore: World Scientific Publishing Co. Pte. Ltd.

Winston, W. L. (2014). *Marketing analytics: Data-driven techniques with Microsoft Excel* (1st). Indianapolis, IN: Wiley.

Woodside, A. G., & Motes, W. H. (1981). Sensitivities of market segments to separate advertising strategies. *Journal of Marketing, 45*(1), 63–73.

Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., et al. (2010). Opinionseer: Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1109–1118.

Xiao, J., Xiao, Y., Huang, A., Liu, D., & Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems, 43*(1), 29–51.

Yankelovich, D. (1964). New criteria for market segmentation. *Harvard Business Review, 42*(2), 83–90.

Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. *Harvard Business Review, 84*(2), 122–131.

Yao, Z., Sarlin, P., Eklund, T., & Back, B. (2014). Combining visual customer segmentation and response modeling. *Neural Computing and Applications, 25*(1), 123–134.

Young, S., Ott, L., & Feigin, B. (1978). Some practical considerations in market segmentation. *Journal of Marketing Research, 15*(3), 405–412.

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30*(4), 451–462.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In C. E. Brodley (Ed.), *Proceedings of the twenty-first international conference on machine learning ICML '04* (pp. 116–123). New York, NY, USA: ACM.

Zhang, Y., Xiong, Z., Mao, J., & Ou, L. (2006). The study of parallel k-means algorithm. In M. Q. H. Meng, & Z. Wang (Eds.), *2006 6th world congress on intelligent control and automation: 2* (pp. 5868–5871). Piscataway, NJ: IEEE.

Zhao, W., Ma, H., & He, Q. (2009). Parallel k-means clustering based on MapReduce. In *Cloud Computing: First International Conference, CloudCom 2009, Beijing, China, December 1–4, 2009. Proceedings* (pp. 674–679). Berlin, Heidelberg: Springer.

Zoltners, A. A., & Sinha, P. (2005). The 2004 ISMS practice prize winnersales territory design: Thirty years of modeling and implementation. *Marketing Science, 24*(3), 313–331.