This is the peer reviewed author accepted manuscript (post print) version of a published work that appeared in final form in:

Bogomolov, Timofei, Korolkiewicz, Malgorzata W & Bogomolova, Svetlana 2020 'Identifying patterns in fresh produce purchases: the application of machine learning techniques', *Handbook of Research on Big Data Clustering and Machine Learning* ch. 18, pp. 378-408.

This un-copyedited output may not exactly replicate the final published authoritative version for which the publisher owns copyright. It is not the copy of record. This output may be used for non-commercial purposes.

The final definitive published version (version of record) is available at:
https://doi.org/10.4018/978-1-7998-0106-1.ch018

Persistent link to the UniSA Research Outputs Repository:
https://hdl.handle.net/11541.2/139914

General Rights
Copyright and moral rights for the publications made accessible in the UniSA Research Outputs Repository are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the persistent link identifying the publication in the UniSA Research Outputs Repository

If you believe that this document breaches copyright, please contact us and provide details. We will remove access to the work immediately and investigate your claim.

Identifying Patterns in Fresh Produce Purchases:

The Application of Machine Learning Techniques

Timofei Bogomolov
University of South Australia, Australia

Małgorzata W. Korolkiewicz
University of South Australia, Australia

Svetlana Bogomolova
University of South Australia, Australia

ABSTRACT

In this chapter, machine learning techniques are applied to examine consumer food choices, specifically purchasing patterns in relation to fresh fruit and vegetables. This product category contributes some of the highest profit margins for supermarkets, making understanding consumer choices in that category important not just for health, but also for economic reasons. Several unsupervised and supervised machine learning techniques, including hierarchical clustering, latent class analysis, linear regression, artificial neural networks, and deep learning neural networks are illustrated using the Nielsen Consumer Panel Dataset, a large and high-quality source of information on consumer purchases in the United States. The main finding from the clustering analysis is that households that buy less fresh produce are those with children – an important insight with significant public health implications. The main outcome from predictive modelling of spending on fresh fruit and vegetables is that contrary to expectations, neural networks failed to outperform linear regression models.

Keywords: Nielsen Consumer Panel Dataset, Machine Learning, Hierarchical Clustering, Latent Class Analysis, Multiple Linear Regression, Artificial Neural Network, Deep Learning Neural Network

INTRODUCTION

Recent advances in technology have led to more data being available than ever before, from sources such as climate sensors, transaction records, scanners, cellphone GPS signals, social media posts, digital images, and videos, just to name a few. This phenomenon is referred to as Big Data, allowing researchers, governments, and organizations to know much more about their operations, thus leading to decisions that are increasingly based on data and analysis, rather than experience and intuition (McAfee & Brynjolfsson, 2012).

Big Data is typically defined in terms of its variety, velocity, and volume. Variety refers to expanding the concept of data to include unstructured sources such as text, audio, video, or click streams. Velocity is the speed at which data arrives and how frequently it changes. Volume is the size of the data, which for Big Data typically means large, given how easily terabytes to zettabytes of information are amassed in today's marketplace.

When it comes to consumer behavior and decisions, consumer data makes it possible to track individual purchases, to capture the exact time at which they occur, and to track purchase

histories of individual customers. This data can be linked to demographics, advertising exposure, or credit history. Hence, researchers now have access to much more consumer data with greater coverage and scope, but also much less structure or much more complex structure than ever before. Traditional econometric modelling generally assumes that observations are independent, grouped (panel data), or linked by time. However, the Big Data we now have available may have more complex structure, and the goal of modern econometric modelling could be to uncover exactly what the key features of this dependence structure are (Einav & Levin, 2014). Developing methods that are well suited to that purpose is a challenge for researchers.

This chapter examines consumer food choices, in particular, purchasing patterns in relation to fresh fruit and vegetables. Consumption of fresh fruit and vegetables makes an important contribution to society in multiple ways. Increased consumption of fruit and vegetables can have a significant positive effect on population health (Mytton, Nnoahim, Eyles, Scarborough, & Mhurchu, 2014; World Health Organization 2015). Strong sales of fresh produce support primary production, contributing to rural and regional economies and farmers' livelihoods (Bianchi & Mortimer, 2015; Racine, Mumford, Laditka, & Lowe, 2013). Fruit and vegetable categories in supermarkets contribute some of the highest profit margins, compared to other product categories (e.g., packaged food), making these categories very important for supply-chain members. Therefore, better understanding and prediction of patterns of consumer purchases of fresh fruit and vegetables could have a substantial positive effect on a range of health, economic, commercial, and social outcomes.

Traditionally, consumer research into fresh fruit and vegetables has relied on consumer surveys, where consumers report their attitudes and intentions to buy fresh produce and barriers to doing so (Brown, Dury, & Holdsworth, 2009; Cox et al., 1996; Péneau, Hoehn, Roth, Escher, & Nuessli, 2006; Finzer, Ajay, & Ali, 2013; Erinosho, Moser, Oh, Nebeling, & Yaroch, 2012). The results were inherently biased by the indirect link between what consumers say in surveys and their actual behavior. When fresh produce purchases were examined, they were often based on self-reports, which typically are influenced by social desirability bias (Norwood & Lusk, 2011) and memory failures, resulting in over- or under-reporting of purchases (Ludwichowska, Romaniuk, & Nenycz-Thiel, 2017). Overcoming these limitations, this chapter draws on a more reliable Consumer Panel Dataset, which is one of the Nielsen datasets made available to marketing researchers around the world at the Kilts Center for Marketing, the University of Chicago Booth School of Business. Since participating households routinely scan all their purchases, Nielsen Consumer Panel Dataset provides a complete and accurate account of their spending on fresh fruit and vegetables across all grocery outlets.

Leveraging off this unique and high-quality dataset, consumer purchasing decisions in relation to fresh fruit and vegetables are investigated using both unsupervised and supervised machine learning techniques, including clustering, artificial neural networks, and deep learning neural networks. By exploring the Nielsen Consumer Panel Dataset, the authors aim to illustrate the extent to which machine learning methods can lead to models that uncover previously unknown patterns of consumer behavior and that capture consumer purchasing decisions better. The specific objectives of this chapter are:

- To illustrate the application of unsupervised (clustering) and supervised (linear regression, artificial neural networks, and deep learning neural networks) machine learning techniques to household purchase data in fresh fruit and vegetables; and

- To highlight the advancement in knowledge and interpretation that could be obtained by the application of these machine learning techniques.

In economic terms, improved understanding of how consumers buy fresh produce could help growers and retailers to increase sales of fresh fruit and vegetables. This, in turn, would have a positive flow on effect on:

- Population health – fruit and vegetables are the healthiest of the foods, and research shows that people do not eat enough of them;
- Economic outcomes, especially for agrarian economies, including the United States and Australia;
- Social outcomes, by providing more jobs and income to food-growing communities and supply-chain members;
- The environment, by reducing consumption of meat products in favor of fruit and vegetables.

The structure of the chapter is as follows: the Nielsen Consumer Panel Dataset is introduced first, followed by an analysis of consumer purchasing decisions. The analysis begins with an exploration of consumer characteristics using clustering techniques. Groupings of households obtained from clustering are then compared in terms of their purchasing behavior, and observed differences are formally assessed using a non-parametric version of MANOVA. The second stage of the analysis incorporates household groupings resulting from clustering into supervised predictive modelling of consumer spending on fresh fruit and vegetables. The next section discusses main outcomes and contributions, followed by some suggestions for future research directions and a conclusion.

NIELSEN CONSUMER PANEL DATASET

In this chapter, patterns of customer purchases of fruit and vegetables are studied using the Nielsen Consumer Panel Dataset, which consists of a representative panel of households that continually provide information about their purchases using in-home scanners to record all their purchases intended for in-home use. Consumers provide information about their households and what products they buy, as well as when and where they make purchases. The dataset covers 13 years (2004 to 2016), with more than a billion purchases from up to 60,000 consumer households in any given year. It provides rich information about household purchasing patterns that allows researchers to study questions that cannot be addressed using other forms of data. For example, the dataset also includes purchases from retailers who traditionally do not cooperate with scanner data-collection companies. In addition, due to the national coverage (the entire United States divided into 52 major markets), there is wide variation in household location, seasonal patterns in availability of fresh produce, and demographics, adding to the richness of the captured consumer information. There are many examples of applications of this dataset in the economic and marketing literature, ranging from studies on tax effects, geography, price policies, in-store promotion and advertising to disaster relief sales, pharmaceutical products, liquor, etc. For a list of working papers that leverage Nielsen Consumer Panel Dataset, see https://www.chicagobooth.edu/research/kilts/datasets/nielsen/working-papers. Yet, studies that use the Nielsen Consumer Panel Dataset and that focus on fresh fruit and vegetable purchases are relatively rare. A few exceptions are studies focused on organic purchases (e.g., Kim, Seok, &

Mark, 2018; Kim, Seok, Mark, & Reed, 2018; Nelson, Fitzgerald, Tefft, & Anderson, 2017), and a study of retirement influence on fresh fruit and vegetable purchases (Hinnosaar, 2018).

The Nielsen Consumer Panel Dataset consists of the following files: Panelists, Trips, Purchases, Retailers, Products, Product Extra Attributes, and Brand Variations. Panelist files contain demographic, geographic, and product ownership information, which is updated annually. Demographic and product ownership variables are recorded for the entire household and the head of the household, as well as demographics for some other household members. Household describing variables (used in this analysis) include household size, income, presence and age of children in household, employment, education, marital status, occupation, type of residence, and race. There is also spatial information – ZIP code, state, county, region – as well as basic indicators of household assets – kitchen appliances, TV items, Internet connection.

Trips files provide summary information about each of the shopping trips made by panelists, such as the household ID, date, retail chain shopped, and the total amount spent for the trip. Since 2007, the number of shopping trips has varied between 10 and 11 million per year, and on average, there are about 180 shopping trips per household per year.

Purchases files provide information about the specific products a household bought on a shopping trip, including the product ID, the quantity, price paid, and any perceived deals. Specifically, there is a flag if a purchase was seen by the panelist as a deal, and the coupon value if the panelist used a coupon. For each year, there are more than 60 million transactions recorded.

Each Products file contains detailed product information for each unique product code, such as description, brand description, multipack and size. Products come from 10 Nielsen-defined food and non-food departments, approximately 4.2 million unique product codes in total for all years. The departments are dry grocery, frozen foods, dairy, deli, packaged meat, fresh produce, non-food grocery, alcohol, general merchandise, and health and beauty aids. Some products contain additional characteristics (e.g., flavor, variety, packaging, salt content, organic claim). These are recorded in Products Extra Attributes files for approximately one million products. Finally, the Retailers file contains an anonymized list of all retailers, each classified into one of 66 channel types (e.g., grocery, online shopping, discount store, hardware, drug store), from which products were purchased.

For the purposes of analyses presented in this chapter, only supermarket shopping is considered. That is, only shopping trips to retailers classified as "grocery" are analyzed, accounting for 66% of all shopping trips recorded. A single year (2013) was selected for the analysis, with data from other years (2014 and 2015) used for validation of outcomes. Given the authors' focus on fresh fruit and vegetables, products of interest belong to the "Fresh Produce" category.

ANALYSIS OF SUPERMARKET PURCHASING DECISIONS BY HOUSEHOLDS

The primary unit of analysis in this chapter is the household. In the field of shopper behavior, household level analysis is the most common, because many grocery purchases, especially from supermarkets, are made on behalf of the entire household. Purchases and trip information were therefore aggregated by household to produce the following variables to describe households' shopping behavior:

- Average number of shopping trips per month (num_trips).
- Average monthly number of items bought across all product categories (purchases).

- Average monthly spending on all product categories (spending).
- Average monthly number of purchases perceived as deals for all product categories (deals).
- Average monthly number of purchases labelled as organic for all product categories (organic).
- Average monthly number of items bought in the Fresh Produce category (fr_purchases).
- Average monthly spending on Fresh Produce (fr_spending).
- Average monthly number of Fresh Produce purchases labelled as organic (fr_organic).
- Average monthly number of perceived deals for Fresh Produce category (fr_deals).

Descriptive statistics for raw data from 2013 are reported in Table 1. On average, US households made 5.11 shopping trips to supermarket per month (median 4.25 trips). They bought 61.8 items (median 51.3) and spent on average $233.28 (median $189.14). That included 7.1 items (median 4.7) and $16.87 (median $11.63) on products from Fresh Produce category (fruit and/or vegetables). On average, households bought 0.8 (0.6) items labelled as organic, including 0.2 (0) items from Fresh Produce category. Also, households purchased on average 21.6 (10.5) items on a deal; only 1.6 (0.5) of them were from Fresh Produce category.

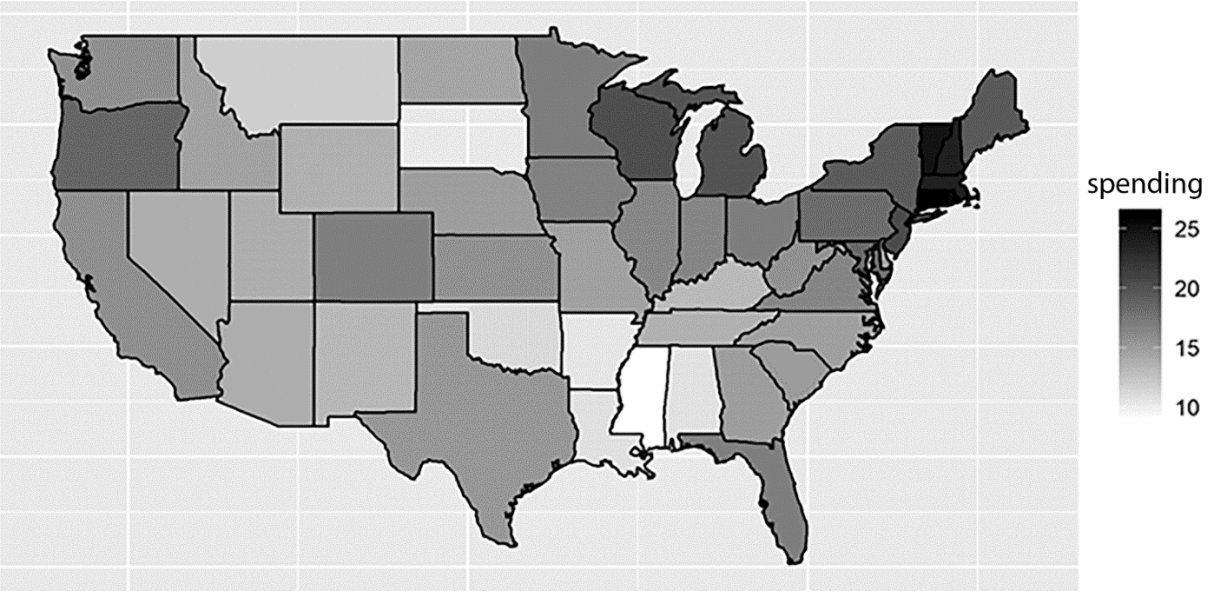
Table 1. Descriptive statistics for variables representing shopping behavior of households

Key	Mean	Std. Dev.	Median	IQR	Min	Max
num_trips	5.11	3.49	4.25	3.88	1.00	38.33
purchases	61.78	46.74	51.25	55.67	1.00	701.08
spending	$233.28	$183.65	$189.14	$211.15	$0.89	$3935.32
organic	0.84	2.68	0.17	0.58	0.00	91.91
deals	21.62	29.09	10.50	29.75	0.00	501.42
fr_purchases	7.08	7.27	4.73	6.67	0.00	132.58
fr_spending	$16.87	$17.60	$11.63	$15.30	$0.00	$712.52
fr_organic	0.19	0.71	0.00	0.13	0.00	43.73
fr_deals	1.61	2.90	0.50	2.00	0.00	84.17

Figure 1 shows average monthly spending on items from Fresh Produce category for different states. Spending pattern correlates with overall wealth of households – coastal states tend to have higher household incomes, and as a result, they might be spending more on all products, including fruit and vegetables.

Next, unsupervised and supervised learning techniques are employed to understand household shopping behavior further. More specifically, clustering is used to explore household demographic information and to create groupings based on similar household characteristics. The resulting household clusters are then compared in terms of their shopping behavior using variables defined in this section. This is followed by predictive modelling of household spending on fresh fruit and vegetables using selected supervised learning techniques. In the case of both unsupervised and supervised learning, multiple methods of varying complexity are employed and compared in relation to outcomes of interest. The purpose of employing multiple methods is to illustrate the impact (if any) of different approaches to the same task on results from that task, and to explore the extent to which more advanced machine learning techniques can be expected to produce better outcomes or predictions with panel data and in the context of studying consumer behavior.

Figure 1. Average monthly expenditure on Fresh Produce in 2013



Exploratory Analysis of Nielsen Consumer Data Using Clustering Techniques

Unsupervised learning, more specifically, clustering techniques, are employed (a) to discover distinct clusters based on household characteristics, and (b) to determine how shopping patterns overall and in fresh fruit and vegetables differ between those clusters of households.

Clustering has been applied across many disciplines to place objects into groups or clusters suggested by available data so that objects in a given cluster are more similar (in some sense) than those placed in different clusters. The objective can be to find a hidden structure in the data or to reduce its dimensionality. Partitioning data into clusters may be an aim in itself or an input into further analysis.

Clustering is considered unsupervised learning, since there are no pre-defined classes into which to assign objects. In fact, the number of classes may need to be established as a part of the clustering process. Most commonly used clustering techniques are distance-based. As the name suggests, a distance metric of some kind (e.g., Euclidean distance) is used to separate observations into clusters with small within-cluster distances and large separation from other clusters. The algorithms involved are relatively simple to apply and interpret, but they suffer from a lack of well-defined rules on how to form clusters and how to assess the quality of clustering solutions. Despite numerous measures and criteria having been proposed in the literature, a decision what “similar enough” or “good enough” might mean in a particular application remains highly subjective. A comprehensive review of distance-based clustering methods and the underlying theory can be found in Hastie, Tibshirani, and Friedman (2009).

Distance-based clustering can be classified into two types: hierarchical clustering and partitional clustering. Hierarchical clustering comes in two forms, agglomerative and divisive. Agglomerative clustering starts with each observation in its own separate cluster; the algorithm then proceeds to find the most similar data points and to group them into larger clusters. In contrast, divisive clustering works in the other direction by first starting with a single cluster

containing all observations and then dividing most dissimilar observations into separate, smaller clusters. Divisive clustering is more complex than agglomerative clustering, as it requires an additional clustering method as a sub-routine to divide existing clusters at each step. The divisive approach has not been studied in the literature as extensively as agglomerative methods, and it is rarely used in practice (Hastie et al., 2009; Jobson, 1992). Partitional clustering assumes there are central features around which the data can be clustered in non-overlapping partitions with the number of centers defined a priori, and it has the advantage of being relatively computationally inexpensive. One disadvantage of partitional clustering is that it requires the expected number of clusters to be known beforehand. In contrast, hierarchical clustering allows the researcher to decide the number of clusters in the data based on the analysis of a dendrogram, which is a tree-like graphical depiction of the structure produced by the algorithm. This makes hierarchical clustering a more flexible approach. Hierarchical clustering is also preferred when there is a perceived hierarchy in the data and for smaller datasets due to its computational complexity. In terms of distance-based clustering, this chapter focuses on hierarchical clustering, as it can be applied to categorical data.

However, clustering techniques are not limited to distance-based methods. There is also a range of model-based clustering techniques that rely on knowing the probability distribution of the data. Cases are assigned to clusters based on assessments of probability rather than a distance measure to find similarities between cases. Latent class analysis (LCA) is an example of a model-based clustering method, also referred to as a finite mixture model. As discussed in McCutcheon (1987) and elsewhere (e.g., Vermunt & Magidson, 2002), the method assumes that the population is composed of different unobserved groups, or latent classes, and the joint density of observed variables is a mixture of the class-specific density. The simplest form of a latent class model assumes that observed data come from a mixture of multivariate multinomial distributions. Cluster membership is decided based on probability calculated using the maximum likelihood method. As the method relies on distributional assumptions, it is possible to use formal tests or goodness-of-fit indices to decide the number of clusters. In that respect, the focus is on modelling the latent structure behind the data rather than looking for similarities, as is the case with distance-based cluster analysis. LCA is sometimes referred to as a "soft clustering" method, as it allows objects to belong to more than one class, and thus it offers more flexibility than a "hard clustering" method such as hierarchical clustering, where objects must be split into distinct clusters.

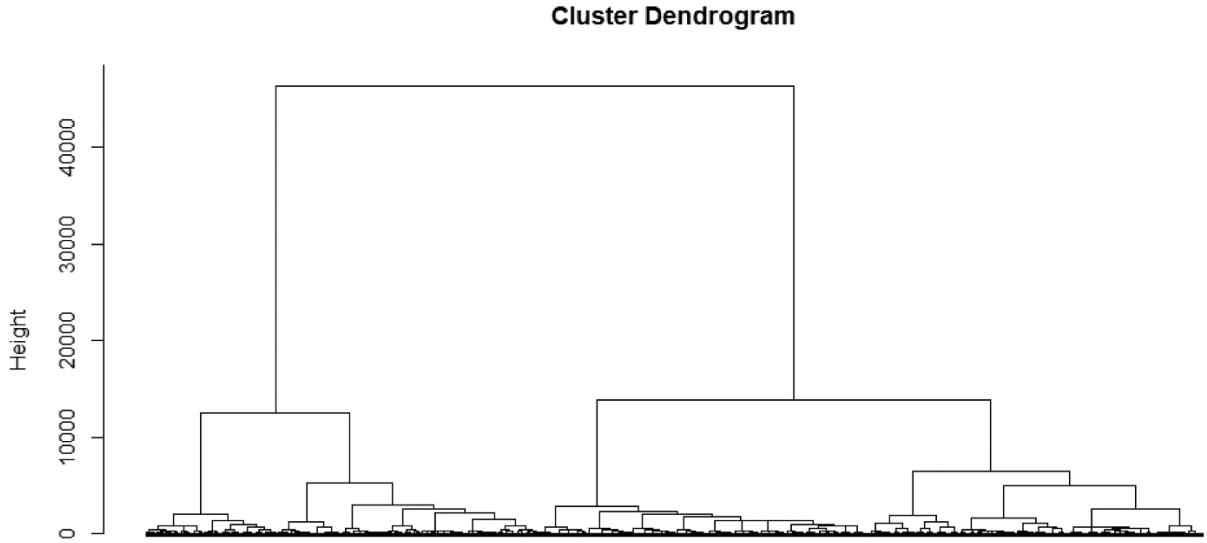
As argued in Anderlucci and Hennig (2014), cluster analysis is not a well-defined problem, in that differing meanings can be assigned to similar objects assigned to the same clusters. In distance-based clustering, the aim is to find well-separated clusters with low within-cluster distances whereas the model-based approach aims to uncover latent structure in the data. In both cases, "true" clustering of the data may be sought; however, there is not much literature to guide users as to which approach is most suitable for their particular application. In this chapter, both distance-based clustering (hierarchical) and model-based clustering (LCA) are applied to illustrate their use in the context of studying purchasing decisions by households. The authors' interest in LCA arises in part from its growing popularity facilitated by the availability of easily implemented LCA-based algorithms in statistical analysis software packages. Another reason for considering a model-based approach is the question of whether resulting partition represents true clustering of the data. It has been pointed out in the clustering literature (e.g., Ultsch & Lötsch, 2017) that by virtue of their design, commonly used algorithms including hierarchical clustering can impose structure on data even when there is actually none. Also, while a hierarchical cluster

analysis always ensures that the most similar observations are in the same clusters, often there are situations where a cluster should be split and its members re-allocated to other clusters. However, no such step exists in the algorithm, and clusters created by hierarchical cluster analysis are commonly less homogeneous than clusters formed by other algorithms. This issue does not arise with LCA, and in addition, some researchers (Anderlucci & Henning, 2014; Vermunt & Magidson, 2002) have found the latent class approach relatively more successful at recovering the true cluster structure when such structure is known, albeit with artificial datasets. Other advantages of LCA clustering include being able to incorporate variables of different scale types, add covariates to the analysis, and overcome sparseness in the data (Vermunt & Magidson, 2002).

Both hierarchical clustering and LCA clustering are now applied to produce clusters of households based on their demographic characteristics. In both cases, the shopping behavior of the resulting clusters is compared. Similarities and differences between the two clustering solutions are also discussed.

Hierarchical Clustering

Figure 2. Cluster dendrogram for 2013 data



To investigate household purchasing behavior, clusters were first formed using agglomerative hierarchical clustering based on following household variables defined by Nielsen and stored in Panelist files: region_code, household_income, household_size, type_of_residence, household_composition, age_and_presence_of_children, race, and marital_status. As all these are categorical variables, they were converted into dichotomous dummy variables and then used to create a matrix of Manhattan distances. Ward's method of minimizing within-cluster variation was used to form clusters, as it produced the most interpretable clusters. The Manhattan distance metric is often chosen, as it is less sensitive to outliers (Jobson, 1992). It is also preferred for high-dimensional and categorical data, and a mathematical justification for using it with Ward's linkage is given in Strauss and von Maltitz (2017).

As a first step, the algorithm was executed for a few small random samples to study the dendrograms and thus to determine the number of clusters to be produced. The dendrogram

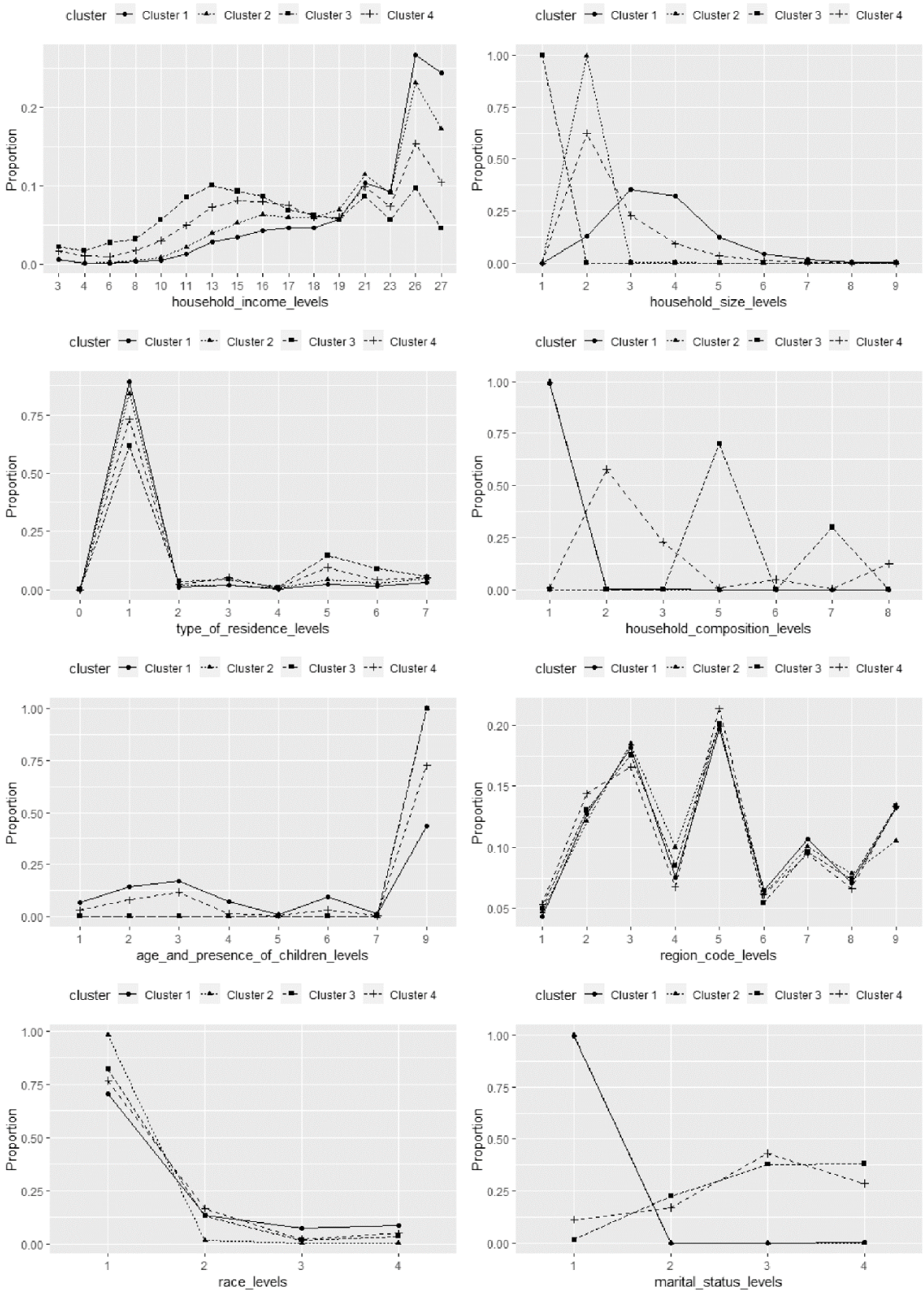
shown in Figure 2 suggested between 3 and 6 clusters. A 4-cluster solution was selected as the most optimal. Next, hierarchical clustering solutions for 4 clusters over different sub-samples of the dataset were investigated and shown to be very similar, and so the four-cluster solution was obtained. Household characteristics were then used to understand cluster composition better. Based on the distributions of proportions depicted in Figure 3, clusters can be characterized as follows:

- Cluster 1 predominantly consists of married couples with 2-3 children or other dependents. It appears that this cluster includes "traditional" families with children as well as so-called crowded nests (married couples living with adult children). It is the largest cluster, and it accounts for 31% of households. This cluster has the highest household size and the highest proportion of households with top income levels. It has a slightly lower proportion of White/Caucasian people and a slightly higher proportion of Asian people than other clusters. This cluster could be named *Families*.
- Cluster 2 has second highest proportion of households with top levels of income, and it consists mostly of White/Caucasian married couples without children or other dependents. This cluster is the same size as Cluster 1, and it accounts for 30-31% of households. It is comprised of DINKs (double income, no kids) including empty nesters (50+ couples whose adult children have left home). This cluster could be named *DINKs*.
- Cluster 3 is composed of single-member households, that is, males or females living alone, never married or widowed, divorced, or separated; predominantly with no kids. It accounts for 25% of households, and it consists mainly of females living alone (approximately 73%). This cluster has the lowest level of income and a somewhat higher proportion of multi-family housing than other clusters. Race distribution in Cluster 3 is similar to that in Clusters 1 and 4. This cluster could be named *SINKs* (single income no kids).
- Cluster 4 is the smallest cluster, accounting for 12% of households. It is the third in terms of incomes and the second in terms of household size (2-4 members). Most households in this cluster are headed by females. All marital status categories are represented, with the highest proportion in the separated/divorced category. This cluster includes single parents with one or more children plus other remaining family and relationship groups (e.g., unmarried couples). This cluster could be named *Single Parents Plus*.

All clusters have very few differences in terms of region, residence type, or race distribution (except Cluster 2 as mentioned above). Hence, the major driving characteristics are household composition and marital status.

Figure 4 shows the average values of variables characterizing shopping behavior for each cluster relative to the average for all households in the dataset. As the distributions of the shopper behavior variables are highly right-skewed, log transformation was applied in all cases. Transformed values were then standardized to produce *z* scores (mean = 0, standard deviation = 1). A bar near zero indicates a cluster average close to the overall average, while a bar with a positive or negative value indicates a deviation from the overall average in terms of higher or lower spending or the number of items purchased.

Figure 3. Household characteristics by cluster in the hierarchical solution



When interpreting, it is important to look at the relative size of the bars within each cluster; a disparity between spending patterns overall and on fresh produce could indicate useful patterns. Also, positive and negative deviations have different interpretations. For example, a high bar in the positive direction for total spending and a low positive bar in fresh produce spending means lower purchases of fruit and vegetables in relation to the total basket size. On the other hand, a high negative bar for the total spending and a low negative bar for the fresh produce spending means that the cluster has a higher proportion of fresh produce in the overall basket.

Many of the variables considered here are likely highly correlated, e.g., a high number of shopping trips in general is associated with a high number of items in all categories and high spending. Hence, the most important finding should come from observing disparities between these variables, e.g., if a household buys well below the average number of items overall but close to an average number of items with deals, it means frugal shopping and utilizing deals more frequently.

Based on Figure 4, the clusters can be described as follows:

- Cluster 1 – *Families* are *big spenders.* They shop with average or even slightly lower frequency, but they buy more than the average number of items and spend even more than the average amount across all product categories. That means they are prepared to pay higher prices for products. This is confirmed by a relatively low number of purchases of deals, compared to the overall shopping. They do the largest total shopping of all clusters, explained by their larger household size. At the same time, while their spending on Fresh Produce is above average, it is not as high as for the total grocery shopping, and it is below that for *DINKs*.
- Cluster 2 – *DINKs* are *shopaholics*. They are households that shop frequently, spend above the average overall, and buy more than the average number of items. They sometimes take advantage of deals, particularly on Fresh Produce, but they do not bother with organics. Their spending and number of items from the Fresh Produce category is the highest among all clusters and above their spending in all categories. *DINKs* really love their smashed avocado!
- Cluster 3 – *SINKs* are *economical shoppers*. They make less than the average number of trips to the supermarket, buy much less than the average number of items, and spend much less than average (explained by their small household size). They take advantage of deals. Their spending on Fresh Produce is relatively in line with the overall shopping and spending. They do not bother with buying organics.
- Cluster 4 – *Single Parents Plus* are the *least health-conscious shoppers*. They are the closest households to an elusive average. They make an average number of shopping trips, they buy close to an average number of items overall and from the Fresh Produce category, and they spend close to an average amount. At the same time, they show the highest disparity between overall shopping and shopping on Fresh Produce – they buy a smaller amount of fruit and vegetables compared to their overall shopping. In the same way, they buy fewer products with deals compared to their overall shopping; hence, they do not take an advantage of price promotions. They are also not particularly interested in buying organic.

In summary, *Families* and *Single Parents Plus* buy relatively lower proportions of fresh fruit and vegetables out of total grocery spend. That is, the proportion of Fresh Produce in their shopping baskets is lower than that in the baskets of *DINKs* or *SINKs*. This is extremely worrying, as both clusters represent household with children. From the public health perspective, these are the households that should be buying more fresh fruit and vegetables to support children's healthy development and to instill healthy dietary patterns in the next generation.

Figure 4. Normalized shopping behavior (means) by cluster in the hierarchical solution

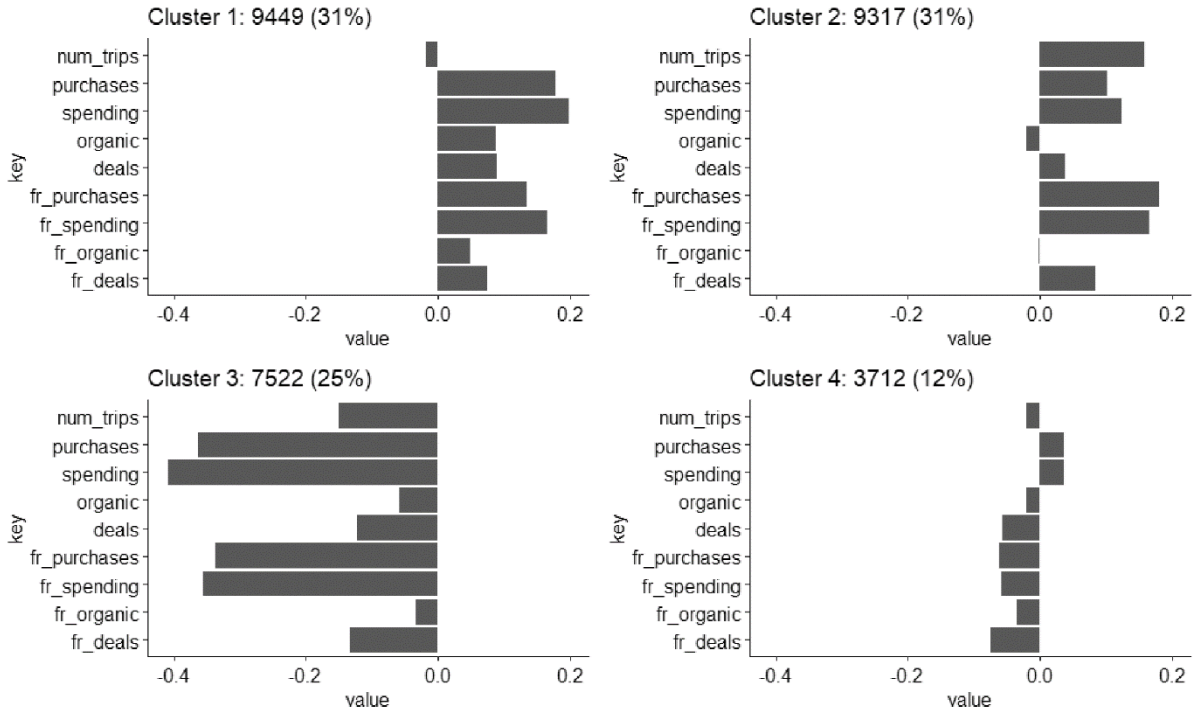


Table 2. Descriptive statistics for shopping behavior variables based on hierarchical clustering

	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)
Cluster	Families		DINKs		SINKs		Single Parents Plus	
N	18,660 (31%)		18,726 (31%)		15,379 (25%)		7,586 (13%)	
num_trips	5.1 (3.6)	4.2 (3.9)	5.6 (3.6)	4.8 (4.2)	4.6 (3.1)	3.8 (3.4)	5 (3.5)	4.1 (3.7)
Purchases (item)	73.2 (55.1)	61 (67.7)	65.7 (44.3)	57.7 (56.7)	43 (30.6)	36.5 (36.8)	62 (45.8)	52.3 (54.4)
Spending ($)	277 (212)	228 (249)	252 (179)	216 (217)	157 (122)	129 (136)	233 (176)	194 (205)
Organic (item)	1.1 (3.3)	0.2 (0.8)	0.8 (2.6)	0.2 (0.6)	0.7 (2)	0.1 (0.5)	0.8 (2.2)	0.1 (0.6)
Deals (item)	26.2 (35.3)	12.3 (36.8)	22.4 (27.9)	11.8 (32.2)	15.5 (20.1)	8.1 (21.7)	20.7 (28.3)	9.9 (29.1)
fr_purchases (item)	8 (8.1)	5.4 (7.4)	8.1 (7.8)	5.6 (7.6)	5 (5.2)	3.4 (4.3)	6.5 (6.4)	4.4 (6)
fr_spending ($)	19.5 (20.4)	13.6 (17.5)	19 (18.3)	13.5 (17)	11.7 (12.2)	8.3 (10.4)	15.4 (15.1)	11 (13.8)
fr_organic (item)	0.2 (0.9)	0 (0.2)	0.2 (0.7)	0 (0.2)	0.2 (0.6)	0 (0.1)	0.2 (0.5)	0 (0.1)
fr_deals (item)	1.8 (3.2)	0.6 (2.3)	1.8 (3.2)	0.6 (2.3)	1.2 (2.1)	0.4 (1.6)	1.4 (2.5)	0.4 (1.8)

Managerially, it is also important to look at the absolute spending and the items bought, not just standardized figures, to ensure the differences between the clusters are substantial enough to

develop actionable strategies for each cluster. Summary statistics in Table 2 are now discussed to gain further insights into the shopping behavior of households. Due to the right-skewness of the distributions for all variables of interest, median might be a better centrality measure. Hence, median is reported along mean values, and it is used to describe typical behavior. It should also be noted that reported standard deviations and interquartile range (IQR) values indicate that there is a lot of variability between individual households in each cluster.

The results in Table 2 reinforce the conclusions from the standardized values analysis in Figure 4. An average US household shops at a supermarket 4.25 times a month (Table 1). Households from clusters *Families* and *Others* undertake almost the same number of shopping trips, 4.2 and 4.1 respectively. At the same time, two-person households in the *DINKs* cluster undertake 13% more, and single-person households in the *SINKs* cluster undertake 10% fewer shopping trips per month than the average household, 4.8 and 3.8 trips respectively.

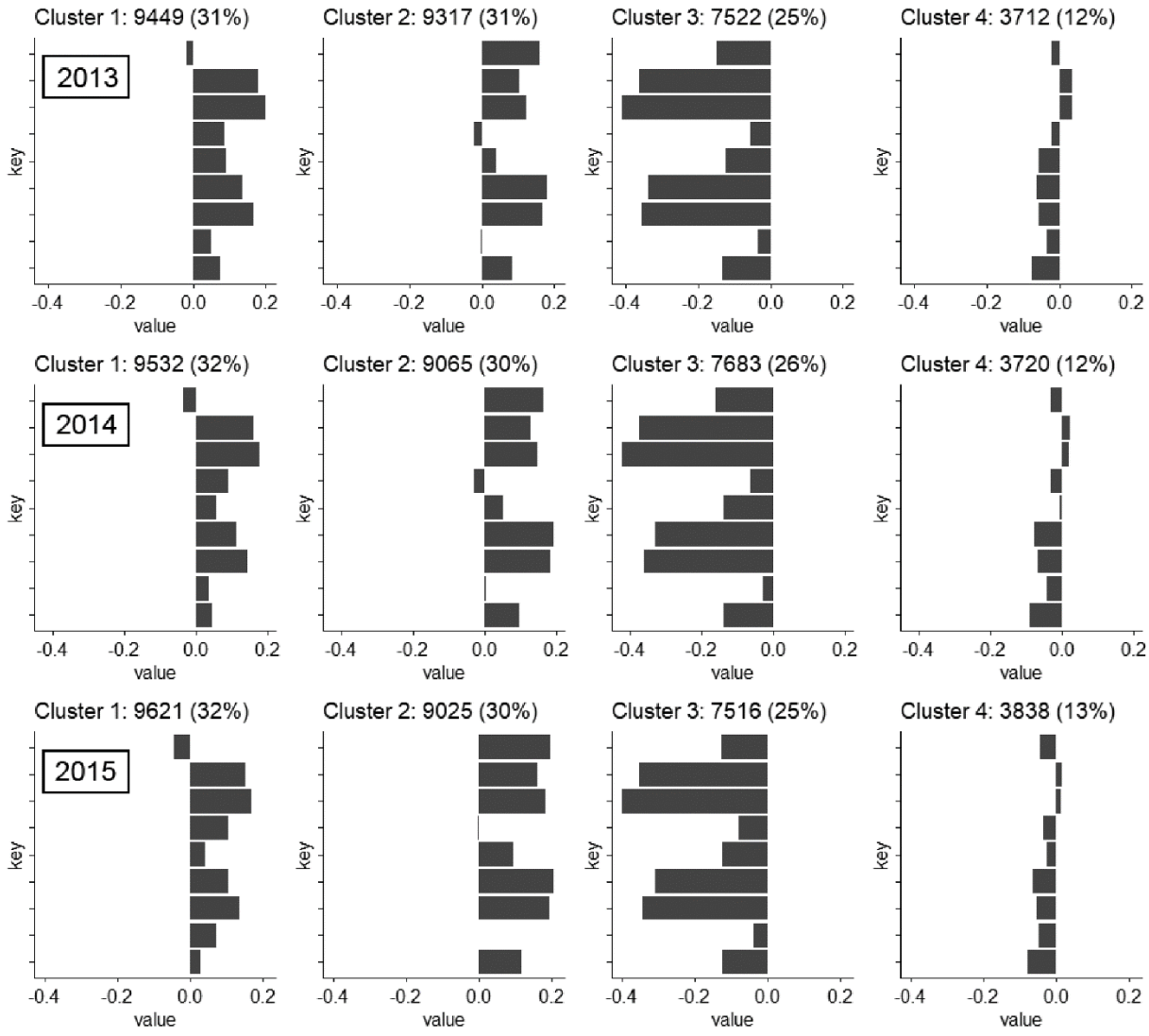
Shopping trip frequency does not always translate into purchasing activity. Despite a somewhat lower number of shopping trips, *Families* buy the highest number of items (median of 61) and spend the top dollar ($228), which is 20% above the national averages of 51.25 items and $189. Frequently shopping *DINKs* households buy 57.7 items and pay $216, which is about 13% above the national average. While *SINKs* households make just 10% fewer shopping trips, their spending is 30% below the national average, 36.5 items and $129 respectively. Households from the cluster *Other* are very close to the average household, with 52.3 items and $194.

In terms of Fresh Produce, a shopping trip typically results in between 3.4 (*SINKs*) and 5.6 (*DINKs*) Fresh Produce items being purchased, costing between $8.30 (*SINKs*) and $13.60 (*Families*). In terms of organic products, a typical household in clusters *Families* and *DINKs* buys one organic item for every five shopping trips. For clusters *SINKs* and *Single Parents Plus,* one organic product is typically purchased for every 10 shopping trips.

Looking at the ratio of Fresh Produce out of a total grocery basket items, it is clear that the two types of households with children (*Families* and *Single Parents Plus*) have lower proportions of fresh items (8.8% and 8.3% respectively) than the other two clusters (*DINKs* 9.8% and *SINKs* 9.4%). These results highlight a difference between clusters that is significant from the perspective of supply-chain managers and public health policy-makers. The overall picture also highlights that all households are buying relatively lower amounts of fresh produce, showing that less than 10% of their grocery baskets are fresh fruit and vegetables.

Real-life applications of clustering techniques suffer from non-replicability or instability of cluster solutions (see, for example, Ben-David, von Luxburg, & Pál, 2006; Leisch, 2015; Levine & Domany, 2001). To address this shortcoming in the present setting, the same clustering procedure was repeated independently for 30,000 panelists randomly sampled from the data for 2013, 2014 and 2015. Figure 5 shows the same nine shopping-related variables as in Figure 4, but for three different years. While there are some minor differences in cluster sizes and in the values of individual variables, the overall structure of clusters remains the same from year to year, which demonstrates the strong validity of the presented approach.

Figure 5. Spending behavior means by cluster for 2013, 2014, and 2015



The differences in the shopping behavior of different groups of customers presented in Table 2 offer managerially significant insights. However, it is necessary to investigate whether there are statistically significant differences among the four clusters. The classical parametric MANOVA (multivariate analysis of variance) cannot be used for these data, as all nine shopping behavior variables are right skewed, and they do not follow a normal distribution. Hence, they do not meet the assumption of multivariate normality required for MANOVA. A nonparametric test for the comparison of multivariate data samples (Burchett, Ellis, Harrar, & Bathke, 2017) was employed to test statistical significance on the global level, as well as to identify statistically significant response variables and factor levels, while controlling the familywise error rate.

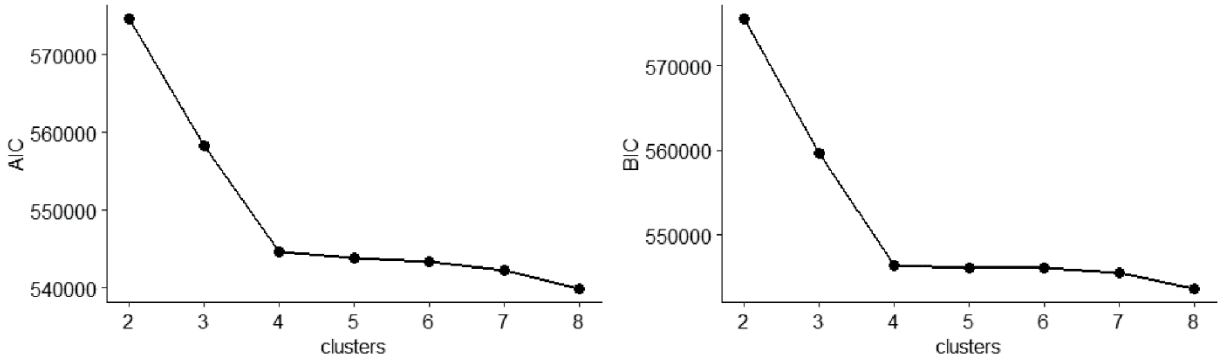
Global test results were highly significant (p-value = 0), indicating that there is a statistically significant difference among the clusters in relation to the nine spending pattern variables, namely the number of shopping trips plus the number of purchases made, the total amount spent, the number of purchases perceived as deals, and the number of organic products bought, both

overall and in the fresh produce category. As the global hypothesis of no difference among clusters was rejected, a detailed follow up analysis was carried out to test for differences for all possible permutations of clusters and shopping behavior variables. A null hypothesis of equality for all appropriate subsets of factor levels (clusters) and response variables (shopping behavior variables) was rejected, while maintaining the maximum overall Type I error rate at a significance level alpha of 0.05. It should, however, be noted that the large sample size may have been a deciding factor in producing statistical significance, and so not all the differences detected may be meaningful in practical terms.

Clustering Based on Latent Class Analysis

LCA is now performed to obtain another clustering solution based on the same demographic information about the households as before. As discussed above, LCA is a "soft clustering" method, where a probability of belonging to each cluster is calculated for each observation. This means that the same household could belong to more than one cluster, so the resulting clusters may overlap, and not be mutually exclusive. This characteristic of LCA, arguably, could be beneficial for the analysis of shopping behavior. Household needs could be changing over time, and/or the same households could display different characteristics in relation to their spending and consumption patterns, or preferences towards fresh produce could vary. With this flexibility in multiple cluster memberships, it is possible to see other clustering structures, which might offer different insights into shopping behavior.

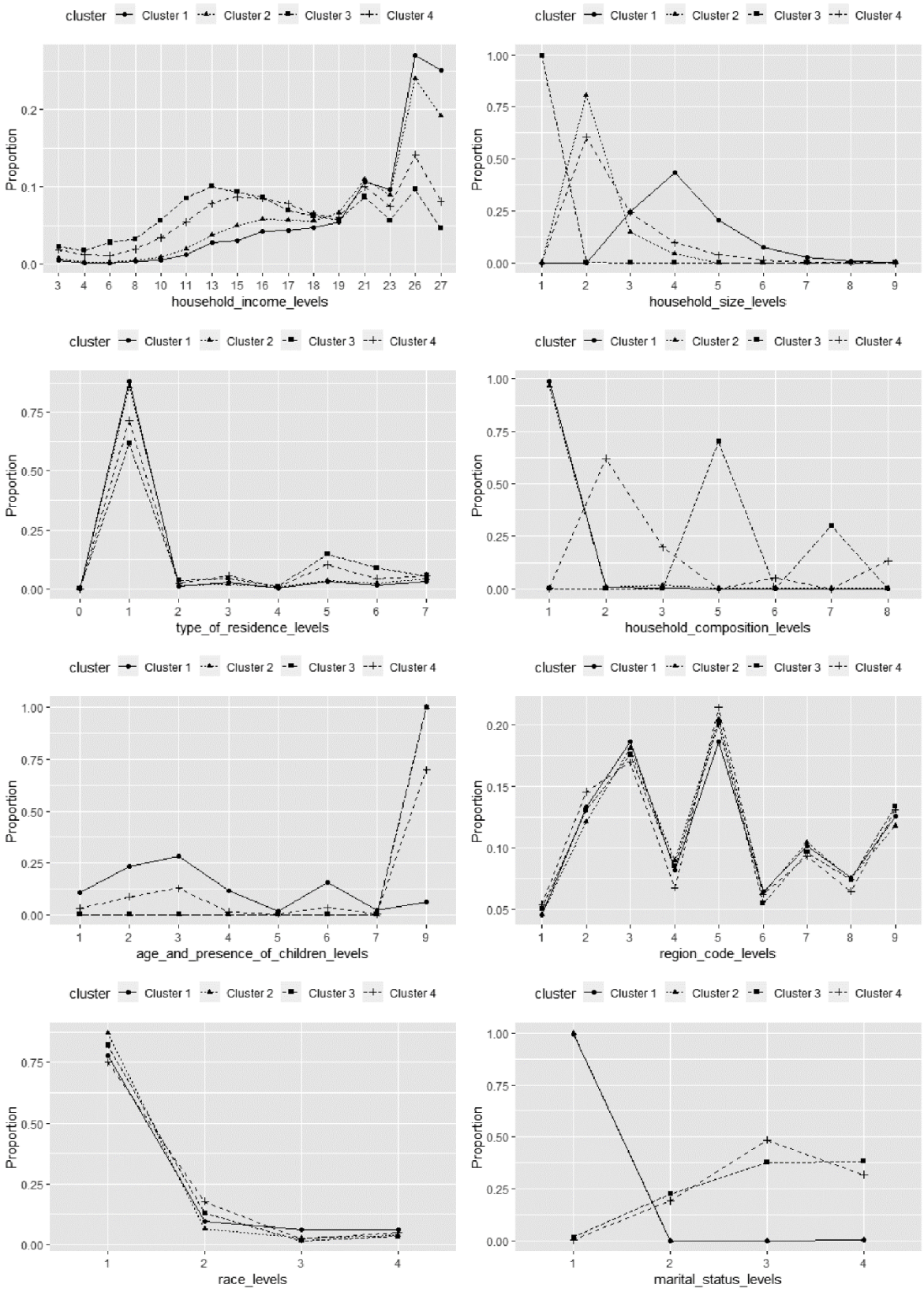
Figure 6. Scree plot based on model fit measures for different numbers of clusters in LCA



To investigate consumer purchasing behavior further, a range of LCA clustering solutions was considered and compared using the Bayesian information criterion (BIC; Schwartz, 1978) and the Akaike information criterion (AIC; Akaike, 1987). For further references regarding the choice of the number of latent classes see Nylund, Asparouhov, and Muthén (2007).

The results for solutions based on 2 to 8 clusters are shown in Figure 6. Both BIC and AIC statistics decrease as the number of clusters increases, indicating better model fit; however, the rate of decrease reduces significantly between 4 and 5 clusters. Therefore, a 4-cluster solution was selected for further analysis.

Figure 7. Household characteristics by cluster in the LCA solution



Household characteristics were again used to understand cluster composition better. Based on Figure 7, clusters produced by the LCA method can be characterized as follows:

- Cluster 1 predominantly consists of married couples with 2-3 children. It is a relatively small cluster, and it accounts for 19% of households. This cluster has the highest household size and the highest income level. It has a slightly lower proportion of White/Caucasian people and a slightly higher proportion of Asian people than other clusters. This cluster could be named *Families with Kids Under 18*.
- Cluster 2 predominantly consists of White/Caucasian married couples without children under 18. It is the largest cluster, and it accounts for 44% of households. This cluster has the second highest proportion of top income levels. In most cases, these are 2-person households but there is also a small proportion of households with one or two additional members. This suggests that this is a cluster of DINK households including the so-called empty nesters (parents whose grown up children have left home) plus crowded nesters (parents living with adult children). This cluster could therefore be named *DINKs & Crowded Nesters*.
- Households in Cluster 3 are single-member households with a high proportion of females living alone, never married, widowed, divorced, or separated, hence SINKs. It accounts for 26% of households. This cluster has the lowest level of income and a somewhat higher proportion of households who live in multi-family housing (e.g., house sharing) compared to other clusters. This cluster could be named *SINKs*.
- Cluster 4 is the smallest cluster, accounting for 11% of households. It is third in terms of income and second in terms of household size. It is represented by single parents (predominantly female) with one or more children. Although predominantly White/Caucasian, relative to the other clusters it has higher proportion of African Americans. This cluster could be named *Single Parents*.

All clusters are again similar in composition in terms of region, residence type (single family house), and race (except Clusters 2 and 4, as mentioned above). The major drivers behind cluster membership are again household composition and marital status. Shopping behavior characteristics of households in LCA-based clusters are discussed next.

Figure 8 shows average values of variables characterizing shopping behavior for each LCA-based cluster relative to the average for all households in the dataset. As in the hierarchical clustering solution, a bar near zero again indicates a cluster average close to the overall average, while a bar with positive or negative value indicates a deviation from the overall average in terms of higher or lower spending or the number of items purchased. As before, shopping behavior patterns are inferred from observing disparities in how variables deviate from the average (the sign and size of the bar) within each cluster.

Based on Figure 8, clusters can be described as follows:

- Cluster 1 – *Families with Kids Under 18* are *big spenders*. They shop less than average, but they buy more than the average number of items, and they spend more than the average amount across all product categories. They spend the most on Fresh Produce of all four clusters, but they come second after *DINKs & Crowded Nesters* when it comes to the number of Fresh Produce items. They buy a lot of Fresh Produce, and they pay

relatively higher prices than other clusters, which indicates a higher quality of fruit and vegetables. They take advantage of deals frequently, but they are also willing to spend more to buy organic products.

- Cluster 2 – *DINKs & Crowded Nesters* are *frequent shoppers.* They are households that shop more than average, spend above average overall, and buy more than the average number of items. They sometimes take advantage of deals, but they do not bother with organics. Their spending on items from the Fresh Produce category is second highest among all clusters, and they buy the largest number of Fresh Produce items. Also, they buy a higher proportion of fruit and vegetables than other clusters.
- Cluster 3 – *SINKs* are *economical shoppers*. They make fewer than the average number of trips to the supermarket, and they buy much fewer than average items and spend much less than average. Their spending on Fresh Produce is lower than the average, in line with their overall shopping and spending; however, the proportion of fruit and vegetables in their shopping is higher than the average household. They do not bother with deals or buying organics.
- Cluster 4 – *Single Parents* are the *least health-conscious shoppers*. They are households that are very close to the average. They do an average number of shopping trips, they buy close to an average number of items overall and from the Fresh Produce category, and they spend close to an average amount. At the same time, they show the highest disparity between overall shopping and shopping on Fresh Produce – they buy less fruit and vegetables compared to their overall shopping. They do not take advantage of price promotions overall, and even less so on Fresh Produce. They are also not interested in buying organic.

A comparison of the clusters shows the highest disparity between overall shopping and shopping for Fresh Produce category for *Families with Kids Under 18* and *Single Parents*. There is a much lower proportion of Fresh Produce items bought by these households compared to their counterparts from the *DINKs* and *SINKs* clusters and compared to the national average.

The summary statistics in Table 3 provide further insights into the shopping behavior of households in different clusters based on the LCA solution. Due to the lack of symmetry in the underlying distributions, median is again used to describe typical behavior in relation to the nine variables.

Compared to national averages of 4.25 shopping trips per month with 51.25 items, including 4.73 items from Fresh Produce category, and spending $189, including $11.63 on Fresh Produce items, *Families with Kids Under 18* shop at supermarkets on average 3.9 times per month, spend $242 for 64.9 items including $14 on 5.5 Fresh Produce items. That makes them the highest spending group – almost 30% above the national average.

Households from the *DINKs & Crowded Nesters* cluster shop in supermarkets more frequently – 4.7 times per month – and buy 57.1 items ($215) including 5.5 items ($13.40) from the Fresh Produce category. *SINKs* households are the lowest spenders – more than 30% below the national average – with 3.8 shopping trips per month, 36.5 items ($129) overall and 3.4 items ($8.30) from Fresh Produce category. *Single Parents* households are the closest to national averages, with 4.1 shopping trips per month, 51.8 items ($191), and 4.3 items ($10.80) from the Fresh Produce category.

Differences in shopping behavior of different groups of customers presented in Table 3 were assessed for statistical significance using the same non-parametric MANOVA approach as in the case of hierarchical clustering. The global hypothesis of no difference among the four clusters based on the nine shopping behavior variables was rejected (p-value = 0). The detailed follow-up analysis again showed statistically significant differences for all appropriate subsets of factor levels and response variables, while maintaining the maximum overall Type I error rate at a significance level alpha of 0.05.

Figure 8. Normalized shopping behavior (means) for LCA clusters

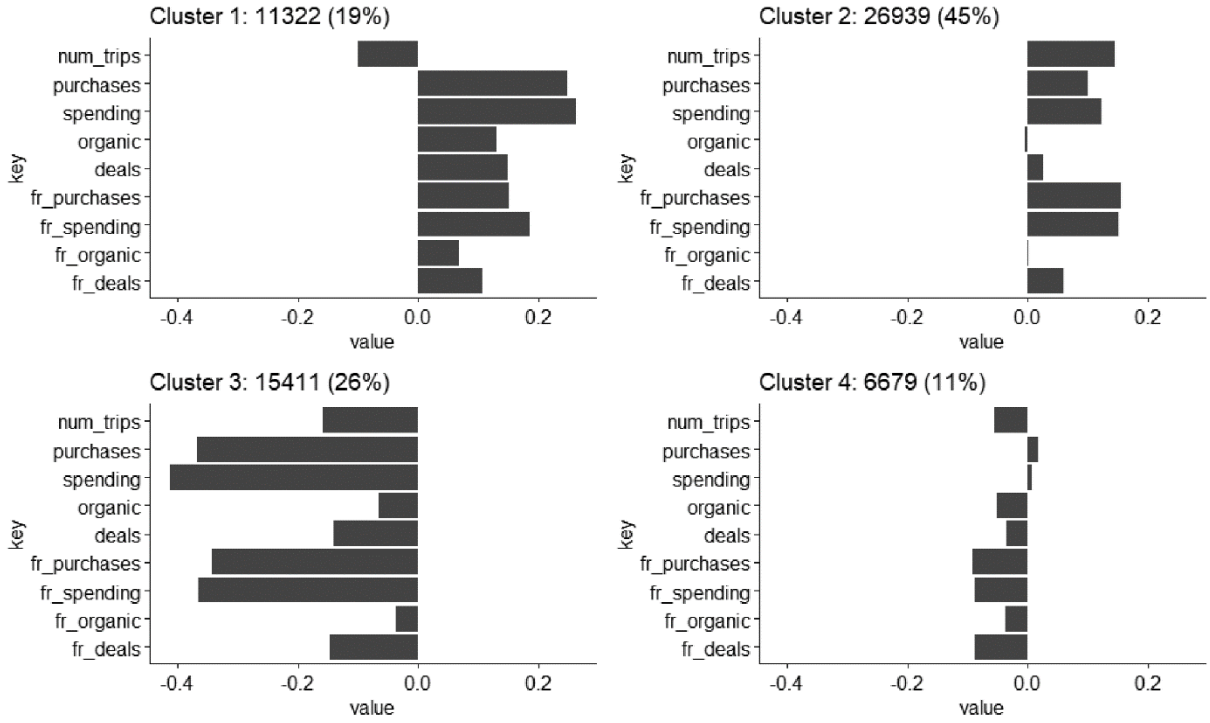


Table 3. Descriptive statistics for shopping behavior variables based on LCA clustering

	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)	Mean (Std. Dev)	Median (IQR)
Cluster	Families with Kids Under 18		DINKs & Crowded Nesters		SINKs		Single Parents	
N	11,322 (19%)		26,939 (45%)		15,411 (26%)		6,679 (11%)	
num_trips	4.8 (3.4)	3.9 (3.6)	5.6 (3.7)	4.7 (4.2)	4.6 (3.1)	3.8 (3.4)	4.9 (3.5)	4.1 (3.7)
Purchases (item)	76.9 (56.3)	64.9 (70.8)	66.2 (46.8)	57.1 (58.2)	43 (30.7)	36.5 (36.8)	61.7 (45.8)	51.8 (54.1)
Spending ($)	289 (213)	242 (259)	254 (188)	215 (222)	157 (122)	129 (136)	230 (174)	191 (200)
Organic (item)	1.2 (3.7)	0.2 (0.8)	0.8 (2.6)	0.2 (0.6)	0.7 (2)	0.1 (0.5)	0.7 (2.2)	0.1 (0.5)
Deals (item)	28.2 (36.6)	14 (39.7)	22.6 (29.3)	11.2 (32)	15.5 (20.2)	8.1 (21.7)	20.8 (28.4)	10 (29)
fr_purchases (item)	8.1 (8)	5.5 (7.6)	8 (7.9)	5.5 (7.4)	5 (5.2)	3.4 (4.3)	6.3 (6.2)	4.3 (5.8)
fr_spending ($)	19.9 (19.9)	14 (17.9)	19 (19.1)	13.4 (17)	11.7 (12.2)	8.3 (10.4)	15 (14.6)	10.8 (13.3)
fr_organic (item)	0.3 (0.9)	0 (0.2)	0.2 (0.7)	0 (0.2)	0.2 (0.6)	0 (0.1)	0.2 (0.5)	0 (0.1)
fr_deals (item)	1.9 (3.3)	0.7 (2.4)	1.8 (3.2)	0.6 (2.3)	1.2 (2.1)	0.4 (1.6)	1.4 (2.5)	0.4 (1.8)

As mentioned previously, LCA is a soft clustering method, in that a probability of belonging to each cluster is calculated for each observation. This means that clusters may overlap, because certain observations can have relatively high probabilities of belonging to multiple clusters. In the case considered here, most households have probability 1 or close to 1 of being in some cluster, so there is no doubt over their cluster membership. However, for about 3% of households, the LCA model produced probabilities that could place them in two clusters. For example, a household with probabilities (0.33, 0, 0.67, 0) could belong to Cluster 1 or 3.

There are two possible strategies to deal with this situation and to get cluster summary statistics. One is to treat households with questionable membership as outliers and to exclude them from the analysis. The other is to overinflate sample size by adding questionable households to all possible clusters. Both approaches were tried and the resulting clustering solutions compared to the one described above, where each household was assigned to a single cluster indicated by the highest of the probabilities. There was no discernible difference between the resulting clusters, and hence no significant differences for cluster interpretability.

Comparing the Results of Hierarchical and LCA Clustering

The LCA solution thus described bears much similarity to the hierarchical clustering solution. The major difference in cluster membership between LCA and hierarchical clustering appears when it comes to households comprising married couples with children. As shown in Table 4, in the hierarchical clustering solution most two-parents-with-children households were placed in one separate cluster. In LCA clustering solution, they were split, and parents living with adult children were grouped together with married couples without children (DINKs) to form the largest cluster. The remaining married couples with children under 18 were assigned to a separate (smaller) cluster. Single parent and single person (SINKs) households appear to have been treated in the same way by the two methods, being placed into separate clusters of similar sizes.

Hence, the main source of disparity appears to be related to characteristics and consequently shopping behaviors of married couple households that include other adult relatives, most likely adult children. According to the LCA method, these households are more similar to DINK households than families with children.

Table 4. Crosstabulation for cluster membership based on hierarchical and LCA clustering solutions

		LCA Clusters			
		Families with Kids Under18	DINKs & Crowded Nesters	SINKs	Single Parents
Hierarchical Clusters	Families	11,199	7436	3	22
	DINKs	34	18,680	0	12
	SINKs	0	11	15,368	0
	Single Parents Plus	89	812	40	6,645

In summary, managerially, it appears that both clustering methods – hierarchical and LCA, produce very similar four-cluster solutions, with largely similar shopping behaviors in respective clusters. Implications of these results are offered in the *Main Outcomes and Contributions* section.

Predictive Modelling of Household Spending on Fresh Fruit and Vegetables

This section describes an application of supervised machine learning techniques to the Nielsen Consumer Panel Dataset. Linear regression has been an important tool of data analysis since its introduction in the early 19th century. Traditionally it serves two purposes: (1) to make predictions for unobserved cases, e.g., out-of-sample or future observations; and (2) to identify the most important input variables (the strongest predictors in the model) through coefficient analysis. Linear regression remains one of the most popular types of data analysis today (Fox, Montgomery, & Lodish, 2004).

Artificial neural networks were introduced in the mid-1900s as an attempt to recreate human neural processing. A neural network is typically presented as three connected layers: an input nodes layer (independent variables or predictors), an output nodes layer (dependent variables or targets), and a hidden nodes layer between them. The number of nodes in the hidden layer can vary in a very wide range, and every node represents a multiple linear regression model with a possible addition of an activation function allowing it to control a process of results passing through. For more details, see Rojas (2013) or Hastie et al. (2009) and the references therein.

From the mathematical point of view, a neural network is a non-linear model resulting from a combination of many multiple linear regression models. Due to the very large number of coefficients in neural network models, their interpretations become much more problematic than linear regression models. However, their predictive ability seems to be superior to linear regression. Indeed, neural networks have been shown to be better predictive models than traditional predictive techniques for datasets where non-linear relationships might be expected or where relationships are simply unknown (Dasgupta, Dispensa, & Ghose, 1994; Sargent, 2001).

The most recent development in the area of machine learning is the introduction of so-called deep learning neural networks. These can be interpreted as neural networks with hierarchical architecture, created by adding multiple hidden layers of different sizes between input and output layers. Hence, a deep learning neural network becomes a non-linear model resulting from a combination of many neural networks. Deep learning techniques have demonstrated the ability to extract high-level, complex abstractions and data representations from large volumes of data in areas such as computer vision (Krizhevsky, Sutskever, & Hinton, 2012) and natural language processing (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), as well as customer behavior analysis (Shi, Xu, & Li, 2017).

In this chapter, all three types of supervised predictive models mentioned above (linear regression, neural network, and deep learning) are employed to predict household spending on products from the Fresh Produce category, that is, on fruit and vegetables, based on spending on other grocery product categories. For this purpose, transaction data was aggregated for individual trips in such a way that every observation is a vector of amounts spent in each of 119 product categories. The resulting dataset had 3.5 million observations (shopping trips) with the spending split across 119 product groups. Then, all observations were aggregated to calculate mean of monthly spending on each category for each household, and predictive models were run at the household level. The amount spent on fresh fruit and vegetables (variable Fresh Produce) was the target, and spending in all other product categories were predictors. Naturally, data were very sparse, as there were no purchases in many categories. All variables were log-transformed and normalized using min-max normalization, a typical process in machine learning that rescales all variables to take values between 0 and 1. This ensures that features measured on a larger scale do

not exert undue influence on the overall results. One more variable had been added as a predictor – clustering membership to serve as a proxy for demographic information about each household.

Ordinary Least-Squares Regression

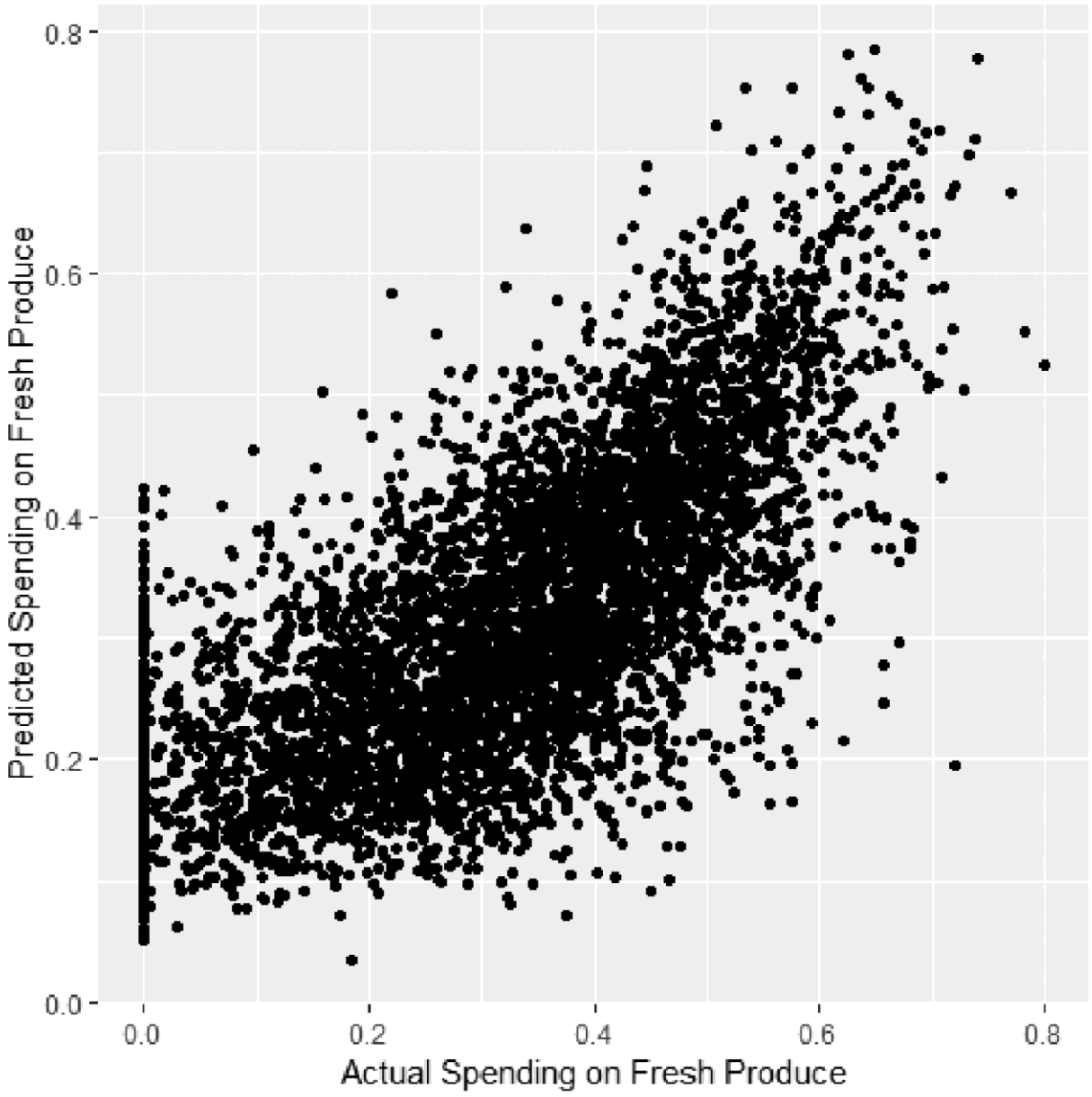
Ordinary least-squares regression was employed to predict the amount spent on Fresh Produce products by each household based on the information about the amount spent on products from other categories. More than 70% of the variables in the model were statistically significant due to the very large sample size. The model has an adjusted *R*-squared value of 0.57, which is considered a relatively good fit for consumer behavior data (Figure 9). There were no differences in prediction results from using different clustering solutions. Therefore, only results based on hierarchical clustering membership are reported below.

The model provides a few interesting insights into spending on Fresh Produce. The highest positive coefficient estimates, that is, product categories positively correlated with spending on Fresh Produce, were frozen unprepared meat, poultry, and seafood (0.2076), yogurt (0.1301), dried vegetables (0.1172), cheese (0.1135), salads and prepared food from the deli (0.0957), and canned vegetables (0.0875). Product categories with the highest negative coefficient estimates, that is, those negatively related to spending on Fresh Produce, were prepared food dry mixes (-0.0789), tobacco accessories (-0.0600), carbonated beverages (-0.0522), prepared frozen food (-0.0466), frozen pizza, snacks (-0.0437), and non-grocery types of products – shoe care, electronics, shaving. All these variables were highly statistically significant with *p*-values close to zero. These results seem logical: households that spend more on protein, dairy, and prepared vegetables – all healthful foods – also spend more on fresh fruit and vegetables. Conversely, households that spend more on unhealthy foods (carbonated beverages, frozen pizza) and tobacco spend less on fresh produce.

While the directional effects of these coefficient estimates are easy to interpret, one needs to be cautious in trying to assign practical significance to these estimates' values. The predictive model operates with normalized log-transformed values of monthly spending in each category, including the target, the Fresh Produce category. Hence, this is a log-log model, and the coefficient estimate interpretation should be given as an expected percentage change in the target variable when predictor value increases by some percentage. For example, an interpretation for a coefficient 0.2076 for frozen unprepared meat, poultry, and seafood, which is the largest coefficient, is as following: a 10% spending increase in this category will result in the spending on Fresh Produce equal to exp(0.2076 log(1.1)) = 1.019983, that is, almost a 2% increase in spending on Fresh Produce. Another large (negative) coefficient for carbonated beverages (-0.0522) means that a 10% increase in spending for this category is expected to decrease Fresh Produce spending by 0.5% – spending equal to exp(-0.0522 log(1.1) = 0.9950. These conversions across the largest coefficients suggest that most of these associations are small in value, and they might have limited managerial significance.

Coefficients for clustering membership were statistically significant but relatively low in value at 0.008 for Clusters 1 and 3 and 0.017 for Cluster 2. This means that with all other variables held constant, compared to a household in Cluster 4, a household in Clusters 1 or 3 would spend 0.8% more, whereas for a household from Cluster 2 the difference would be 1.7%.

Figure 9. Actual vs predicted plot based on the OLS regression model to predict spending on fresh produce



Artificial Neural Network

As the next step, the same prediction task was attempted using another form of supervised machine learning – artificial neural networks. A neural network with 122 input nodes (spending in 118 product categories and 4 cluster membership categories), 1 output node (spending in the Fresh Produce category) and one hidden layer of 180 nodes was created and trained over 40 epochs with a 20% validation split. The model used mean squared error as a loss function, and an RMSProp optimizer and a ReLU activation function for each layer. The model had 22,321 trainable parameters. The prediction result on the test sample was marginally worse than after the linear regression with *R*-squared = 0.54 (Figure 10).

Figure 11 shows the history of the loss function for training and validation datasets over 40 epoch training process. The maximal improvements in the model – the lowest values of the loss function – were achieved after 10-15 epochs. After that, the neural network was overfitting the training dataset without any real improvements in the predictions on the validation dataset. Therefore, despite the expectation, the artificial neural network model did not produce better predictions than a much simpler multiple linear regression model.

Figure 10. Actual vs predicted plot for the one-layer artificial neural network to predict spending on fresh produce

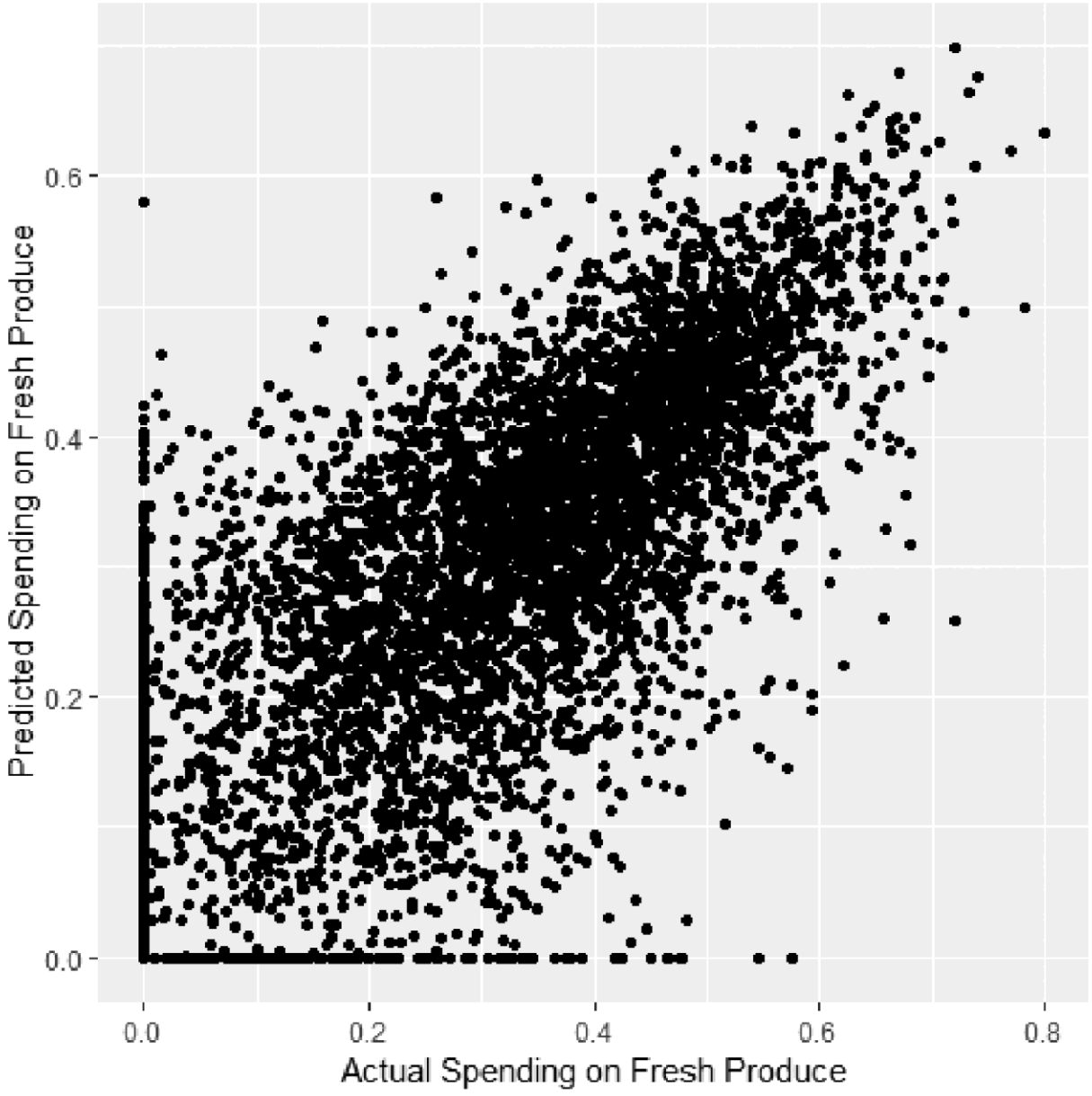
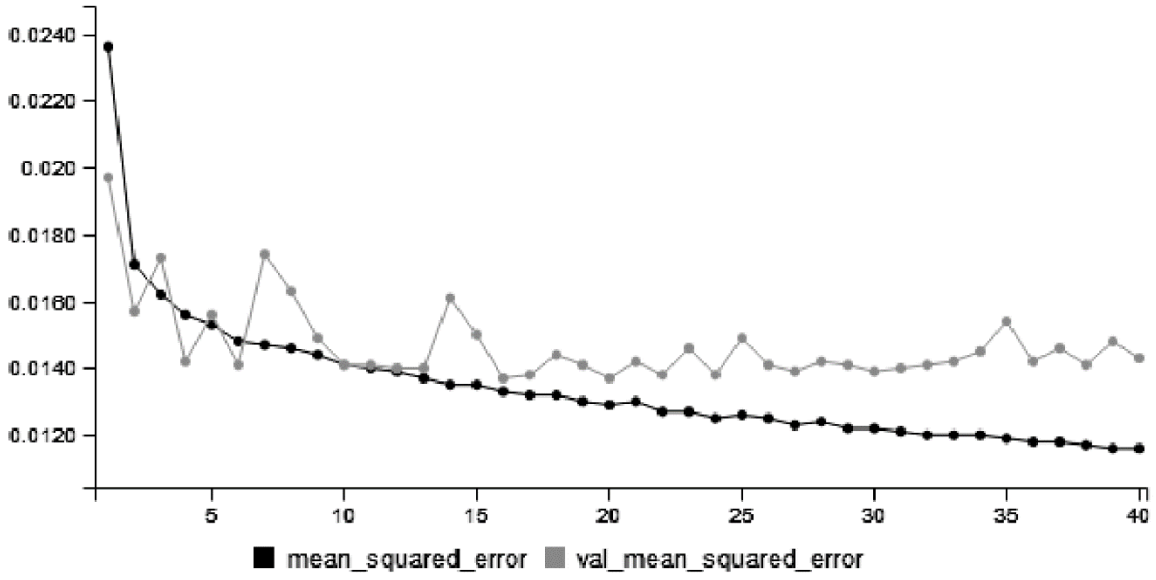


Figure 11. Loss function for training and validation datasets in the one-layer artificial neural network



Deep Learning Neural Network

A deep learning modeling technique was also attempted for the same dataset. A new neural network had the same 122 input and 1 output nodes as before, plus 10 hidden layers of 180 nodes each. That resulted in 315,541 trainable parameters. The model was trained over 40 epochs with a 20% validation split. Figure 12 shows results of predicting spending on Fresh Produce for households from the test sample, while Figure 13 reports a history of the loss function over the 40 epochs of the neural networks training process. The lowest values of the loss function on the validation dataset were achieved after 4 epochs only. After that, the model was overfitting the training dataset and actually reducing the effectiveness of predictions on the out-of-sample data. That resulted in a quite poor quality of predictions with an *R*-squared of 0.46, which is below than *R*-squared for the one-layer neural network and the OLS regression model fitted previously.

Figure 12. Actual vs predicted plot for the deep learning neural network

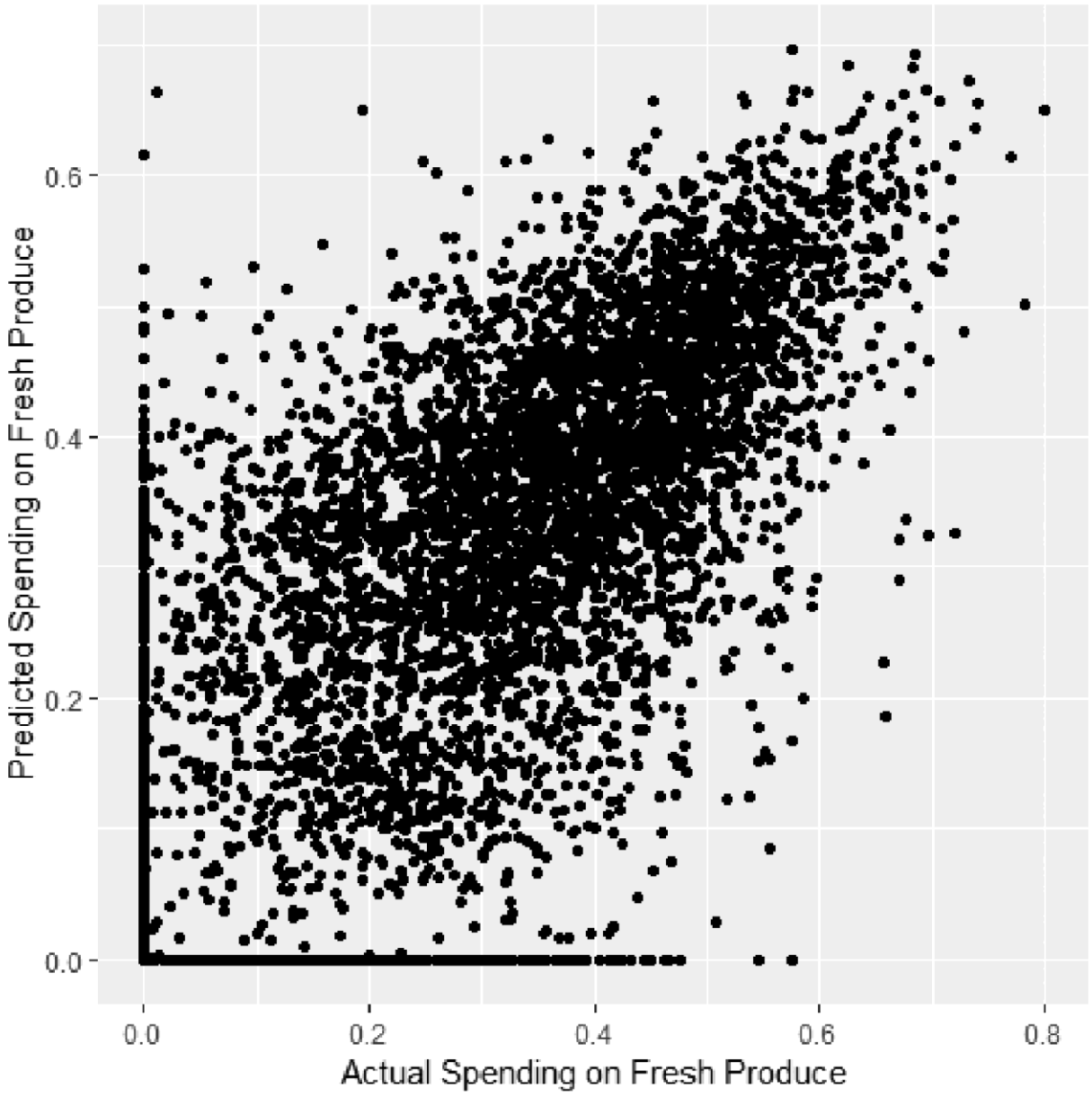
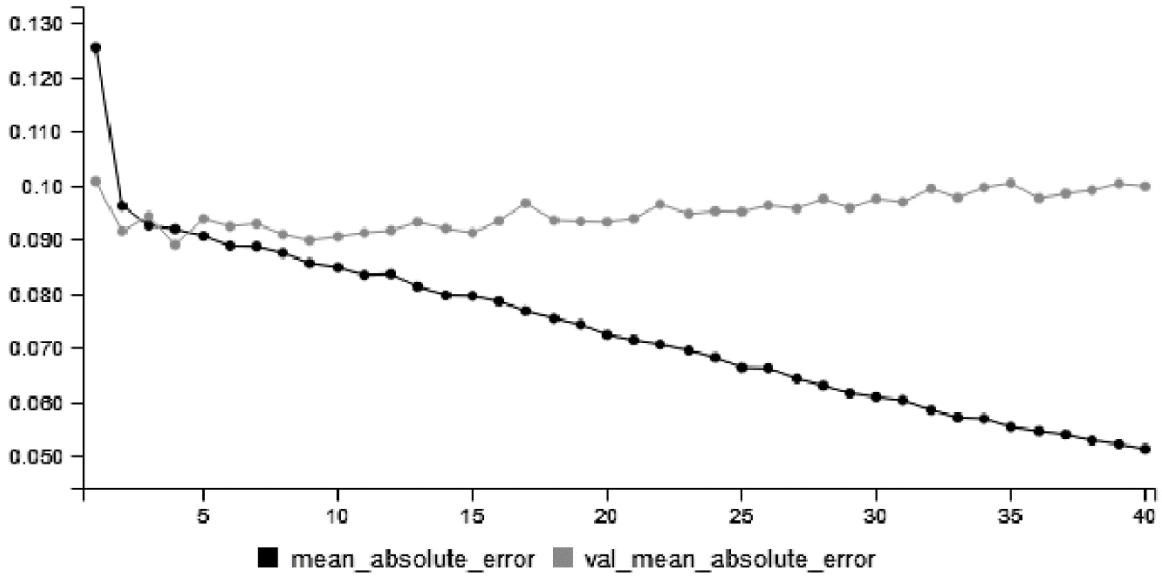


Figure 13. Loss function for training and validation datasets in deep learning neural network



MAIN OUTCOMES AND CONTRIBUTIONS

In this chapter, the authors aimed to illustrate the application of unsupervised and supervised machine learning techniques to consumer purchases panel data, with a particular focus on fresh fruit and vegetables.

First, the chapter has demonstrated a step-by-step application of two different clustering techniques: hierarchical (distance-based) and LCA (model-based). The key finding from the clustering analysis is that the types of households that buy less fresh produce are those with children – an important insight that carries serious public health implications. Further, the authors showed that in the context of consumer purchases in fresh produce, it is possible to develop a clustering solution that is replicable over time (three years) and across two different clustering methods (distance- and model-based). Such replication over years and across clustering methods, and the resulting robust clustering solution, makes an important contribution to the literature on clustering, where non-replicability of clusters is one of the major methodological challenges.

Second, the chapter has illustrated the application of three predictive modelling techniques – linear regression, artificial neural network, and deep learning, and it has compared the resulting models based on their predictive power. Despite common expectations, more advanced machine learning techniques (artificial neural network and deep learning neural network) were no more effective than a multiple regression model in predicting the amount spent on fresh fruit and vegetables based on household characteristics and spending in grocery categories other than fresh produce. This robust empirical evidence makes another valuable contribution to the field of machine learning, suggesting that in some circumstances (as presented here) more advanced techniques do not necessarily offer more insights or more predictive power. This finding is also useful for the field of marketing and econometrics to guide researchers' choices of analytical tools.

For the field of public health and agricultural economics, this chapter has offered a robust discovery of shopping patterns in fresh fruit and vegetables categories using scanner-based information, a major advancement on past studies dominated by survey-based information about consumer purchasing decisions. As such, this chapter has offered new insights into buying of fresh produce by US households, which is new knowledge that carries important implications for supply-chain and public health practitioners.

FUTURE RESEARCH DIRECTIONS

While promising, the analysis presented in this chapter offers only a glimpse into the wealth of insights that could be gained from Nielsen Consumer Panel Dataset regarding purchasing decisions made daily by US households. As discussed above, unsupervised learning in the form of clustering produced some results of practical significance; similarities and differences in spending on fresh produce were established for households grouped into clusters based on their demographic characteristics. Predicting the average amount spent on fresh produce using selected supervised learning techniques proved more challenging, and it could be subject of further study. The Nielsen dataset offers individual shopping trip information for all households, down to the exact contents of shopping baskets, which could be explored for patterns across time as well as cross-category purchase decisions.

Further replication of the observed patterns is needed. Specifically, the authors call for more studies that directly compare different clustering methods in a pursuit of developing procedures that provide more replicable clustering solutions. Similarly, while this chapter has demonstrated that in the context of consumer purchases of fresh fruit and vegetables, more advanced predictive models do not outperform the traditional models, more work is needed to compare these models on the same datasets to offer direct comparisons and to start building a body of evidence regarding their performance and predictive abilities in different contexts.

CONCLUSION

This chapter has illustrated the application of common machine learning techniques to a novel context of consumer purchases of fresh fruit and vegetables. The findings offer useful contributions to a number of academic disciplines: machine learning and statistics, marketing and econometrics, public health, and agricultural economics.

ACKNOWLEDGMENT

The authors acknowledge the support of the Ehrenberg-Bass Institute for Marketing Science, Business School, University of South Australia in gaining access to Nielsen Consumer Panel Dataset.

DISCLAIMER

1. Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.

2. The conclusions drawn from the Nielsen data are those of the researcher(s), and they do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332.

Anderlucci, L., & Hennig, C. (2014). The clustering of categorical data: a comparison of a model-based and a distance-based approach. *Communications in Statistics-Theory and Methods*, 43, 704-721.

Ben-David, S., von Luxburg, U., & and Pál, D. (2006). A sober look at clustering stability. In H. U. Simon & G. Lugosi (Eds.). *Proceedings of the 19th Annual Conference on Learning Theory* (pp. 5-19). New York, NY: Springer.

Bianchi, C., & Mortimer, G. (2015). Drivers of local food consumption: A comparative study. *British Food Journal, 117*, 2282-2299.

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research, 45*(1), 77-93.

Brown, E., Dury, S., & Holdsworth, M. (2009). Motivations of consumers that use local, organic fruit and vegetable box schemes in Central England and Southern France. *Appetite, 53*(2), 183-188.

Burchett, W., Ellis, A., Harrar, S., & Bathke, A. (2017). Nonparametric inference for multivariate data: The R package NPMV. *Journal of Statistical Software, 76*(4), 1-18.

Cox, D. N., Reynolds, J., Mela, D. J., Anderson, A. S., McKellar, S., & Lean, M. E. J. (1996). Vegetables and fruit: Barriers and opportunities for greater consumption. *Nutrition & Food Science, 96*(5), 44-47.

Dasgupta, C. G., Dispensa, G. S., & Ghose, S. (1994). Comparing the predictive performance of a neural network model with some traditional market response models. *International Journal of Forecasting, 10*(2), 235-244.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(1), 1-24.

Erinosho, T. O., Moser, R. P., Oh, A. Y., Nebeling, L. C., & Yaroch, A. L. (2012). Awareness of the Fruit and Veggies—More Matters campaign, knowledge of the fruit and vegetable recommendation, and fruit and vegetable intake of adults in the 2007 Food Attitudes and Behaviors (FAB) Survey. *Appetite, 59*(1), 155-160.

Finzer, L. E., Ajay, V. S., Ali, M. K., Shivashankar, R., Goenka, S., Pillai, D. S., . . . Prabhakaran, D. (2013). Fruit and vegetable purchasing patterns and preferences in South Delhi. *Ecology of Food and Nutrition, 52*(1), 1-20.

Fox, E. J., Montgomery, A. L., & Lodish, L. M. (2004). Consumer shopping and spending across retail formats. *The Journal of Business, 77*(S2), S25-S60.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hinnosaar, M. (2018, July 30), *The impact of retirement on the healthiness of food purchases*. doi:10.2139/ssrn.3235215

Jobson, J. D. (1992). *Applied multivariate data analysis. Volume II: Categorical and multivariate methods*. New York, NY: Springer.

Kim, G., Seok, J. H., & Mark, T. (2018, February 13). *New market opportunities and consumer heterogeneity in the U.S. organic food market*. doi:10.2139/ssrn.2916250

Kim, G., Seok, J. H., Mark, T., & Reed, M. R. (2018, January 1). *The price relationship between organic and non-organic vegetables in the U.S.: Evidence from Nielsen scanner data*. doi:10.2139/ssrn.3176082

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, & K.Q. Weinberger (Eds.), *Advances in neural information processing systems 25 (NIPS 25)* (pp. 1097-1105). Lake Tahoe, CA: Curran Associates.

Leisch, F. (2015). Resampling methods for exploring clustering stability. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 637-652). Boca Raton, FL: Chapman and Hall/CRC.

Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation, 13*, 2573-2593.

Ludwichowska, G., Romaniuk, J., & Nenycz-Thiel, M. (2017). Systematic response errors in self-reported category buying frequencies. *European Journal of Marketing, 51*, 1440-1459.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 60-68.

McCutcheon, L. A. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K.Q. Weinberger (Eds.), *Advances in neural information processing systems 26 (NIPS 2013)* (pp. 3111-3119). Lake Tahoe, CA: Curran Associates.

Mytton, O. T., Nnoaham, K., Eyles, H., Scarborough, P., & Mhurchu, C. N. (2014). Systematic review and meta-analysis of the effect of increased vegetable and fruit consumption on body weight and energy intake. *BMC Public Health, 14*(1), 1-11.

Nelson, E., Fitzgerald, J. M., Tefft, N., & Anderson, J. (2017, September 1). *US household demand for organic fruit*. doi:10.2139/ssrn.3081997

Norwood, F. B., & Lusk, J. L. (2011). Social desirability bias in real, hypothetical, and inferred valuation experiments. *American Journal of Agricultural Economics, 93*, 528-534.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: An Interdisciplinary Journal, 14*, 535-569.

Péneau, S., Hoehn, E., Roth, H. R., Escher, F., & Nuessli, J. (2006). Importance and consumer perception of freshness of apples. *Food Quality and Preference, 17*(1-2), 9-19.

Racine, E. F., Mumford, E. A., Laditka, S. B., & Lowe, A. (2013). Understanding characteristics of families who buy local produce. *Journal of Nutrition Education and Behavior, 45*(1), 30-38.

Rojas, R. (2013). *Neural networks: A systematic introduction*. Berlin, Germany: Springer Science & Business Media.

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches: Results from medical data sets. *Cancer: Interdisciplinary International Journal of the American Cancer Society, 91*(S8), 1636-1642.

Schwartz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

Shi, H., Xu, M., & Li, R. (2017). Deep learning for household load forecasting—A novel pooling deep RNN. *IEEE Transactions on Smart Grid, 9,* 5271-5280.

Strauss T., & von Maltitz M. J. (2017). Generalising Ward's method for use with Manhattan distances. *PLoS ONE, 12*(1), 1-21. doi:10.1371/journal.pone.0168288

Ultsch, A., & Lötsch, J. (2017). Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics, 66,* 95-104. doi:10.1016/j.jbi.2016.12.011

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class models* (pp. 89-106). Cambridge: Cambridge University Press.

World Health Organization (2015). Increasing fruit and vegetable consumption to reduce the risk of noncommunicable diseases. Geneva, Switzerland: Author.

ADDITIONAL READING

Agresti, A. (2012). *Categorical data analysis*. New York, NY: Wiley.

Athey, S. (2018). The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (pp. 507-547), Cambridge, MA: National Bureau of Economic Research. Retrieved from https://EconPapers.repec.org/RePEc:nbr:nberch:14009.

Hennig, C. (2015). Clustering strategy and method selection. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 703-737). Boca Raton, FL: Chapman and Hall/CRC.

Johnson, R. A., & Wichern, D. W. (2008). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice Hall.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives and prospects. *Science, 349*(6245), 255-260. doi:10.1126/science.aaa8415

Najafabadi, M. N., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, 2*(1), 1-21. doi:10.1186/s40537-014-0007-7

Von Luxburg, U., Wiliamson, R. C., & Guyon, I. (2012). Clustering: Science or art? In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *JMLR Workshop and Conference Proceedings* 27 (pp. 65-79). Washington, DC: MIT Press.

KEY TERMS AND DEFINITIONS

Artificial Neural Network (ANN): A predictive computer algorithm inspired by the biology of the human brain that can learn linear and non-linear functions from data. Artificial neural networks are particularly useful when the complexity of the data or the modelling task makes the design of a function that maps inputs to outputs by hand impractical.

Cluster Analysis: A type of an unsupervised learning that aims to partition a set of objects in such a way that objects in the same group (called a cluster) are more similar, whereas characteristics of objects assigned into different clusters are quite distinct.

Deep Learning: A type of machine learning based on artificial neural networks. It can be supervised, unsupervised, or semi-supervised, and it uses an artificial neural network with multiple layers between the input and output layers.

Hierarchical Clustering: The most common approach to clustering. The method proceeds sequentially, producing a nested assignment of objects into clusters. It is typically agglomerative, with cluster sizes increasing as the number of clusters decreases. At each step of the process, a clustering criterion based on a measure of proximity between groups must be computed to decide which groups of objects are to be joined together.

Latent Class Analysis (LCA): A statistical technique used in factor, cluster, and regression modelling, where constructs or latent classes are identified from multivariate categorical data and used for further analysis. The probability that a case belongs to a particular latent class is calculated using the maximum likelihood method. The resulting models can also be described as finite mixture models.

Machine Learning: A branch of artificial intelligence that focuses on data analysis methods that allow for automation of the process of analytical model building.

Partitional Clustering: A commonly used approach to clustering that begins with a preselected number of groups or clusters. An initial allocation of objects to clusters is followed by reassignment to new groups based on a measure of proximity between each object and each group. The process continues until all objects have been assigned to their closest groups. A commonly used partitioning method is the k-means algorithm.

Predictive Modelling: A process of using data mining or machine learning techniques to predict outcomes of interest. Once variables that are likely to influence the outcomes are identified and the relevant data is collected, a model is formulated and tested.

Supervised Learning: A machine learning task designed to learn a function that maps an input onto an output based on a set of training examples (training data). Each training example is a pair consisting of a vector of inputs and an output value. A supervised learning algorithm analyzes the training data and infers a mapping function. A simple example of supervised learning is a regression model.

Unsupervised Learning: A class of machine learning techniques designed to identify features and patterns in data. There is no mapping function to be learned or output values to be achieved. Cluster analysis is an example of unsupervised learning.