PROJECT REPORT

on

SUPERMARKET SHOPPING ANALYSIS: MARKET SEGMENTATION, SEQUENCE CLUSTERING AND STATE PREDICTION USING MARKOV CHAINS AND HIDDEN MARKOV MODELS

SUBMITTED BY:

RITIK SHAH

DEPARTMENT OF MARKETING

ALLIANCE MANCHESTER BUSINESS SCHOOL

UNIVERSITY OF MANCHESTER

SEPTEMBER 2019

Acknowledgement

It is with gratitude that I express my sincere gratefulness to **Dr. Panagiotis Sarantopoulos, Assistant Professor,** Department of Marketing, Alliance Manchester Business School, University of Manchester, under whose able guidance and constant supervision this work has been accomplished. I thank him for taking time out of his busy schedule and aiding me with his suggestions, encouragement and cooperation, which in turn helped me, enhance the scientific merit of the present project work.

Contents

2 Project Summary

The following project is an attempt towards shopping state sequence analysis of consumers visiting a supermarket. The main purpose of this analysis is to identify the shopping state of customers and help the businesses to better understand its customers and therefore fulfill their needs effectively. This study takes into account the shopping data of 336 customers visiting a supermarket store for a period of 3 years. The project is divided into 3 parts: Market Segmentation, Clustering of Customers and prediction of shopping state using a Markov chain and a Hidden Markov Model.

Market segmentation is a decision-making tool for the marketing manager in the crucial task of selecting a target market for a given product and designing an appropriate marketing mix. Market Segmentation helps in better understanding of differences between the customers and shopping trips, which in turn form the basis of long-term competitive advantage in the selected targets segments. The main goal of this step is to divide the shopping trips (*baskets*) into segments followed by a detailed characterization of each segment on the basis of products bought. Further, we try to draw insightful relationship and differences between the segments in order to find the target segment.

Clustering and visualizing sequences is the next important step in order to identify customer with similar shopping patterns. Clustering can efficiently summarize and render the sequences into limited number of groups (*clusters*). A Markov chain and Hidden Markov Model (HMM) is fitted for each cluster. Markov models are used for predicting the future shopping state depending on the current state of the customer, by detecting the underlying latent structures. Further, we compute the prediction efficiency for a single model fitted for all the sequences and separate model fitted for individual clusters.

The goal of the project is to carry out a detailed analysis of shopping trips of customers to a supermarket. Identifying the target segment and predicting the future shopping state of customer are two major results derived from this project.

3 Introduction

The purpose of marketing is to match the genuine needs and desires of consumers with the offers of suppliers particularly suited to satisfy those needs and desires. This matching process benefits consumers and suppliers, and drives an organization's marketing planning process. The most crucial part of marketing is acting according to the needs of the customer. Almost all the businesses today are keen to practice data mining to better understand the customers. This project involves the study of customers going to supermarket and their shopping state. Supermarket analysis is used to discover patterns or correlations with in the set of items. The major aim is to analyze the habits of buyers to find out the correlation between one item to another. The correlations can help the marketer to promote sales strategy by items frequently purchased by the consumers. Facilitating easy access to the products for which the customer shows interest can help in enhancing the shopping experience. Therefore, to find out the target segment and analyzing the shopping state of the customer is the research priority.

To analyze the huge amount of transaction data we use K-means clustering algorithm. This algorithm finds the best possible clusters of products in the supermarket. Further, we formulate the stability and profile plot in order to characterize the segments. The transaction data can be considered as a time series sequential data for the customers. After research and experiment, Markov models can be used to model the time time-series dataset. HMM is followed by Viterbi Algorithm to predict the hidden state for each observation. This results can help supermarkets to understand the needs of their customers and act accordingly.

The outline of the report is as follows. In Section 4 we provide the descriptive analysis of data. Next, is the process of Segmentation of shopping trips in Section 5. We continue in Section 6, by analyzing the time series transaction data and clustering the customers. In Section 7 & 8, we use Markov chains and Hidden Markov Model (HMM) to predict the shopping state of customers. Finally, we formulate our work and focus on future prospects.

4 Descriptive Analysis

The project is based on the analysis of shopping sequences of 336 consumers from a supermarket. A total of 67,664 baskets were purchased by the consumers over a time period of 3 years (3-01-2012 to 30-09-2014). The dataset is divided into two sub-datasets, scans & cards.

Scans Dataset

The dataset describes the products bought by customers during the given time period. Each row corresponds to a single product transaction and the columns with the details of card number, the date of transaction, basket number, category of product bought, barcode, units of the category bought and the euros spent. The dataset has 548634 observations (*rows*) and 7 variables (*columns*).

Cards Dataset

The dataset describes the each customer visiting the supermarket. The dataset has 336 rows and 12 columns. Each row corresponds to a customer, followed by 12 columns mentioning the details as follows: card number (*user*), first transaction date, last transaction date, total number of transactions, total number of units bought, total number of euros spent, total number of days between first and last transaction, days per transaction, age, gender, p_code, household size of the customer.

A detailed descriptive analysis was carried out for both scans and cards dataset to understand the customers and their shopping statistics. The results obtained were as follows:

Table 1-Descriptitve Analysis

Maximum number of products were sold on:	*22/12/2012* *Products Sold:1446*
Minimum number of products were sold on:	*20/07/2014* *Products Sold: 148*
Product sold in maximum quantity:	*vegetables* *Quantity: 62925*
Products sold in minimum quantity:	*food_cupboard_other* *Quantity:4856*
Number of unique barcodes generated:	*13711*
Day for maximum shopping:	*Saturday*

	Number of products:132750
Day for minimum shopping:	*Sunday* *Number of products: 3107*
Month with highest revenue:	*December 2012*
Month with lowest revenue:	*August 2014*
Average of days per transaction for336 customers:	*5.54 days*
Highest number of transactions done by a customer:	*720*
Lowest number of transactions done by a customer:	*82*
Number of females shopping:	*226*
Number of males shopping:	*110*
Average age of the people shopping:	*50.61 years*
Maximum number of units purchased by a customer:	*10060.19*
Minimum number of units purchased by a customer:	*354.221*
Maximum amount spent by a customer:	*19437.25 Euros*
Minimum amount spent by a customer:	*913.505 Euros*

4.1 Participation Plot

The participation plot shows the participation each category in the total number of transactions.

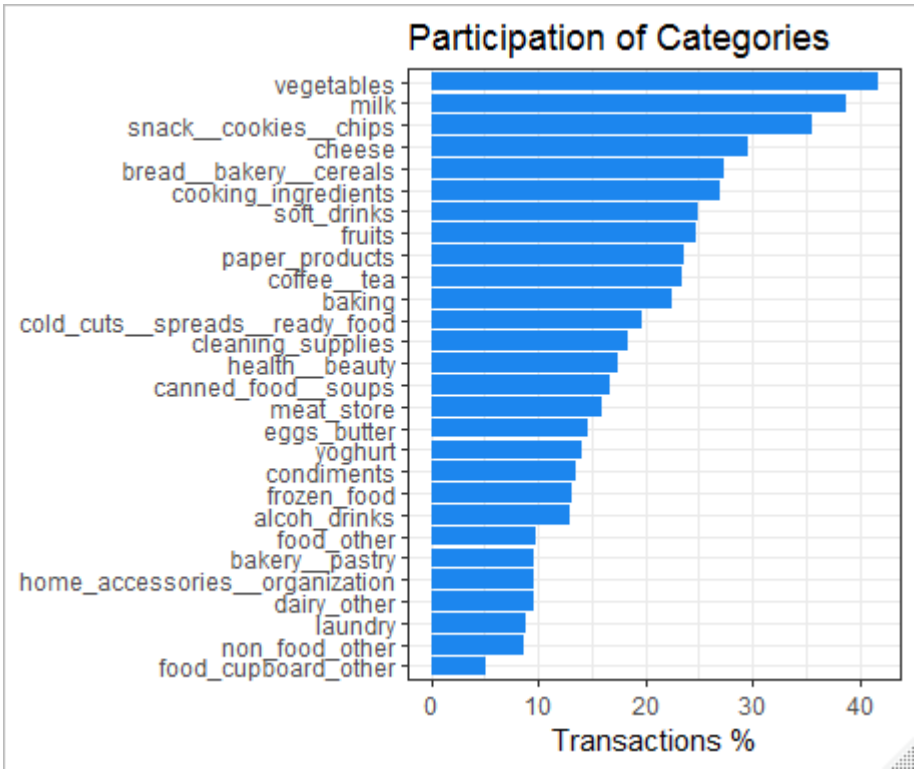


Figure 1-Participation plot of the categories

From the above plot it can be inferred that vegetables is the part of maximum number of transactions i.e. 41.75%, while food_cupboard_other category has the least participation i.e. 5.20%. Percentage participation each category in total number of transactions is given below:

Table 2-Participation percentage of Categories

Category	Percentage	Category	Percentage
alcoh_drinks	12.913809	food_other	9.762946
bakery_pastry	9.697919	frozen_food	13.179830
Baking	22.478718	fruits	24.813786

bread_bakery_cereals	27.366103	health_beauty	17.532218
canned_food_soups	16.753370	home_accessories_organization	9.650627
cheese	29.690825	laundry	8.825964
cleaning_supplies	18.352447	meat_store	16.002601
coffee_tea	23.523587	milk	38.810889
cold_cuts_spreads_ready_food	19.784524	non_food_other	8.596891
condiments	13.627631	paper_products	23.563490
cooking_ingredients	26.956727	snack_cookies_chips	35.605344
dairy_other	9.629936	soft_drinks	24.937929
eggs_butter	14.710925	vegetables	41.751892
food_cupboard_other	5.203653	yoghurt	14.057697

4.2 Principle Component Analysis

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

Results & Conclusion from Principle Component Analysis

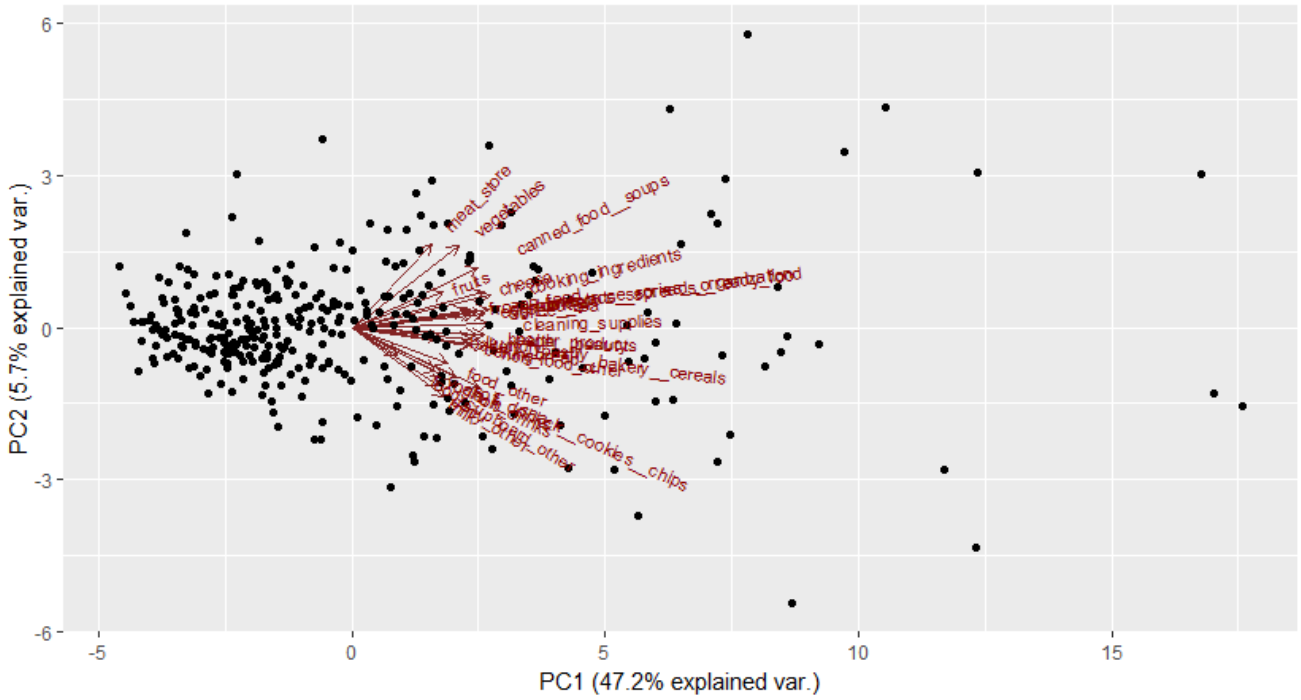


Figure 2-Principle Component 1 and 2 for scans dataset

The black dots in the above figure represent customers (*336*) and the red lines represent the product categories (*28*). From the above figure, a conclusion can be drawn that there are no unique customers for a particular category and they overlap each other. For a supermarket analysis a 2-D plot of consumers' product preferences typically does not contain a clear group of consumers. Rather consumer preferences are spread across entire plot. Therefore it can be stated that consumers buy versatile products from the supermarket and come in all shapes and forms.

5 Market Segmentation

Market Segmentation is the process of grouping consumers or shopping trips (*baskets*) into naturally existing or artificially created segments which have similar product preferences or characteristics. The most popular partitioning method is K-means clustering algorithm. This algorithm divides the 67664 baskets into subsets such that baskets assigned to same market segment are similar to one another as possible, while baskets belonging to different market segments are as dissimilar as possible. The representative of market segment is referred to as the centroid. This iterative algorithm always converges and leads to segmentation solution.

5.1 Separation, Stability & Profiling of Segments

Once the shopping trips have been grouped into segments, each of these segments have to be profiled and described in detail. Before profiling, an important task is to determine the optimal number of segments. The key here is to repeat the extraction process for different number of segment and then select the number of segments with most stable solution. Once we choose the stable solution, profiling and describing the selected segment solution help users to understand each of the segments, and select which one(s) to target. When one or more target segments have been chosen, profiling and describing segments inform the development of the customized marketing mix. For the current dataset, we repeat the extraction process for clusters varying from 3 to 8. Stability, positioning and profile diagrams are plotted for every segmentation solution.

Results & Conclusion from Market Segmentation

Analyzing the plots for different segmentation solutions we conclude that segmentation solution with 5-clusters is the most stable and can be characterized efficiently.

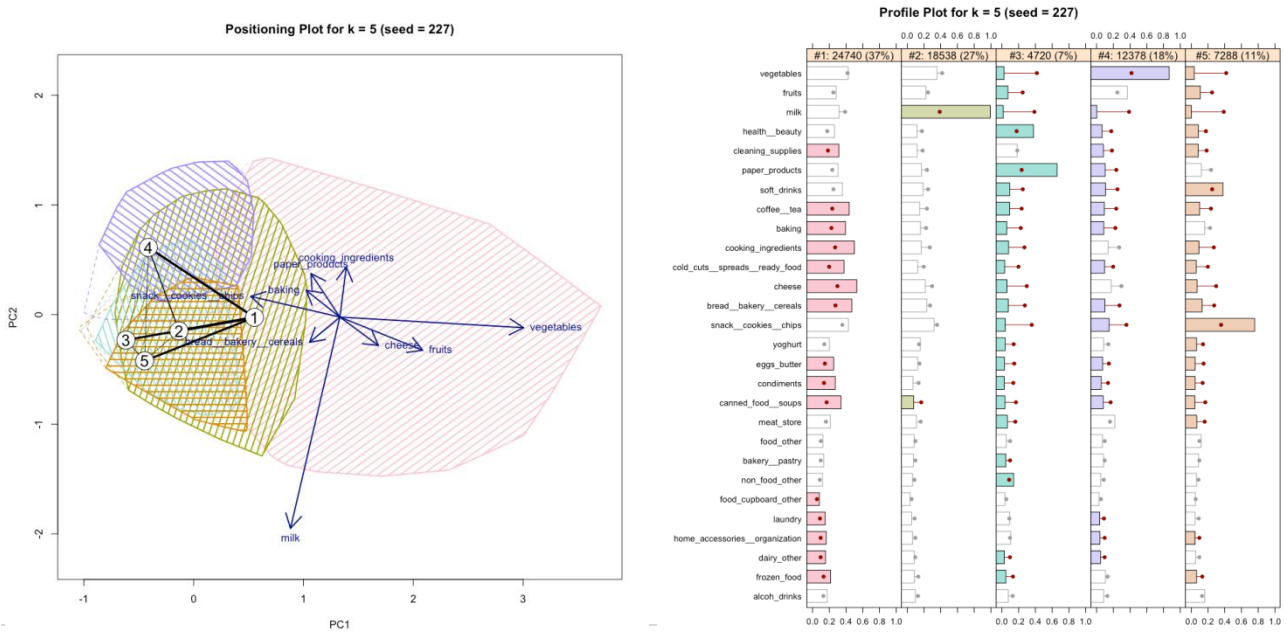


Figure 3-Positioning and Profile Plot for 5-Cluster Segmentation Solution

The clusters are characterized as follows:

- **Cluster 1:** General Items
- **Cluster 2:** Milk
- **Cluster 3:** Paper and Health Products
- **Cluster 4:** Vegetables and Fruits
- **Cluster 5:** Snacks and Drinks

The profile plot above for 5-segments shows that the segment with *highest number of shopping trips, 37%* (*target segment*) includes all the products, followed by *27% (target segment-2) shopping trips* having milk as the major product while only *7% shopping trips* having paper and health products as the major product. In order to determine the target segments in detail we will do the Segment Level Stability Analysis (SLS_A) in the next section. Segmentation results show that for a supermarket chain, consumers come in all shapes and forms and the target segment include customers with diversified descriptor variables and products.

5.2 Segment Level Stability across Solutions (SLS_A)

This method is used to analyze the stability of segments for a segmentation solution. The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing different numbers of segments. High values of segment level stability across solutions (SLS_A) serve as indicators of market segments occurring naturally in the data, rather than being artificially created. Natural segments are more attractive to organizations because they actually exist, and no managerial judgement is needed in the artificial construction of segments. For the current dataset the SLS_A plot is as follows:

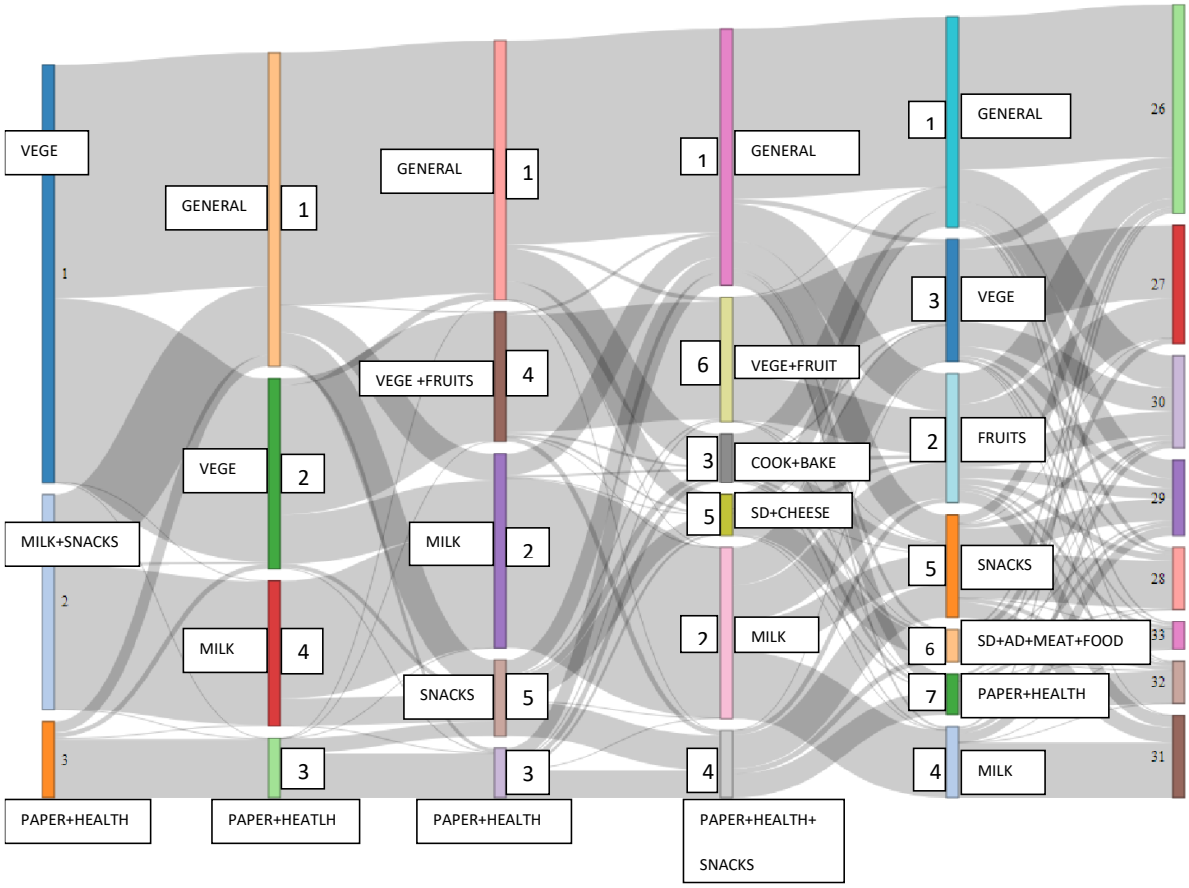


Figure 4-Segment Level Stability Plot across different Segmentation Solutions

Each column in the Figure-4 represents a segmentation solution with a specific number of segments. The number of segments extracted increases from left to right. The column on the far left represents the segmentation solution with three segments. The column on the far right represents the segmentation solution with eight segments. The lines between segments indicate movements of segment members between segments. Thick lines between two segments indicate that many segment members are retained (despite the number of segments increasing). Thick lines represent stubborn market segments, market segments which re-

occur across segmentation solutions, and therefore are more likely to represent natural segments. Segments which have many lines coming in from the left and branching into many lines to their right, suffer from changing segment membership across calculations with different numbers of segments. Such segments are more likely to be artificially created during the segment extraction process.

Results & Conclusion from SLS_A

Looking at 5-cluster segmentation solution Cluster 1, 2 and 4 are quite stable, since the movement of observation is when moving from 4-segment solution to 6-segment solution is very limited. Looking at the segment profile plot, it can be seen that the shopping trips in this segments majorly include general, milk and vegetables & fruits as their categories. From Figure-4, it is also obvious that Cluster-5 and 3 (*Snacks & Drinks, Paper & Health*) demonstrates low segment level stability within the solution. Rather, it represents a grouping of shopping trips (*baskets*) the algorithm was forced to extract because we asked for five segments.

Determining the segment level stability within the segmentation solution (SLS_W)

Once we determine the most stable segmentation solution and stable segments by looking at the SLS_A plot, we can quantify the stability of the segments within the solution using the SLS_W boxplot. Figure-5 represents Jaccard coefficient on y-axis and segment number on x-axis. Segment having the highest J-coefficient is the most stable.

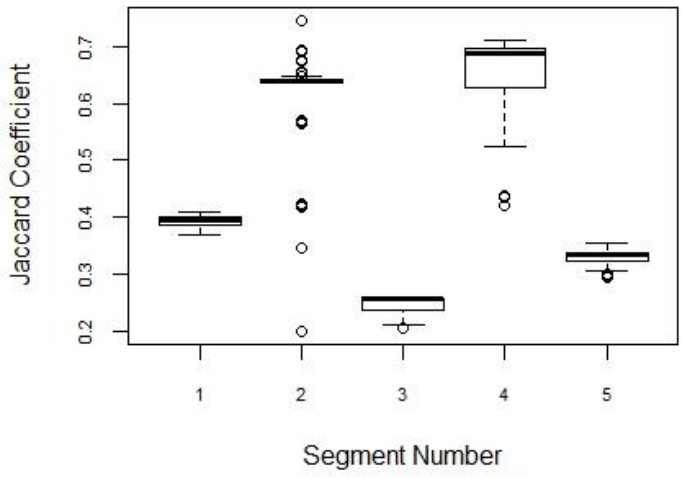


Figure 5-Segment Level Stability within 5-segment solution

The key conclusion that can be drawn is that Cluster 1, 2 and 4 are potential target segments because they show a potential sign of naturally existing market segment. From Figure-5 it can be seen that Cluster 4 is the most stable, followed by Cluster 2 and 1. Do not consider targeting Cluster 5 and Cluster 3 as there are very few

shopping trips (*baskets*) which include snacks & drinks, paper & health products only as the category.

6 Sequences Analysis

The supermarket is all about the shopping sequences of the customers visiting it. These shopping sequences can more appropriately be described as a *Time-Series Sequential Data*. An analytical approach and analysis of the shopping sequences can help in determining,

- The relationship of sequences with covariates.
- Group with similar patterns and typologies of sequences.
- Standard trajectories and consumer shopping behavior.
- Consumer needs and the future shopping state of the customer.

The time series dataframe for 336 sequences is prepared from the first and last date of transaction. The day on which no transaction is made is filled with the previous or last occurring shopping state. Therefore, a dataset with 336 rows and 1002 columns (*3-01-2012 to 30-09-2014*) is prepared under the name time_step.

Results from Sequence Analysis

Market Segmentation results in a total of 5-clusters segmentation solution. The legends for the clusters are as follows-

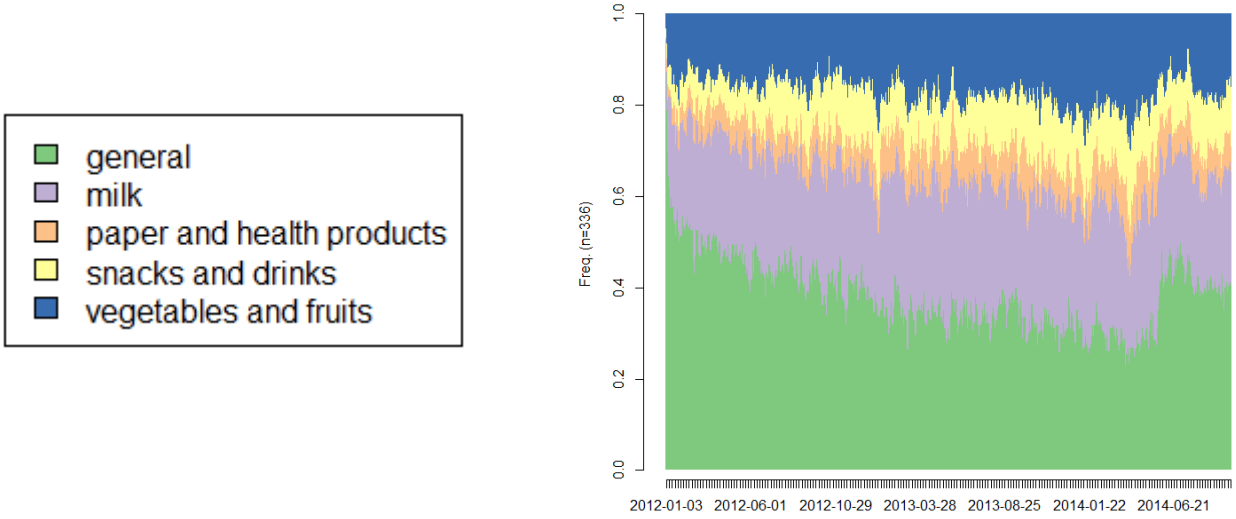


Figure 6-State Distribution Plot with Time

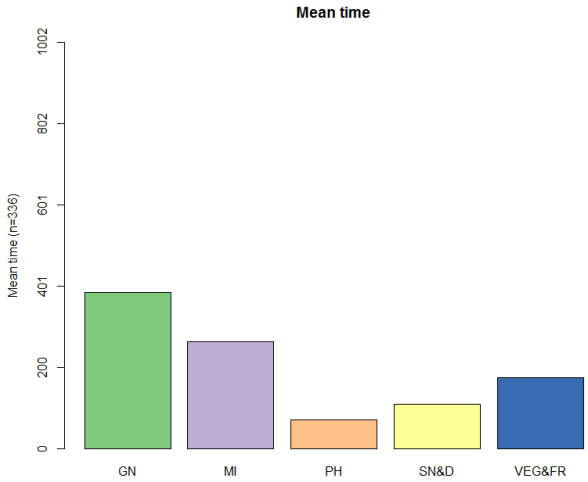


Figure 7-Mean time spent in each State by 336 Customers

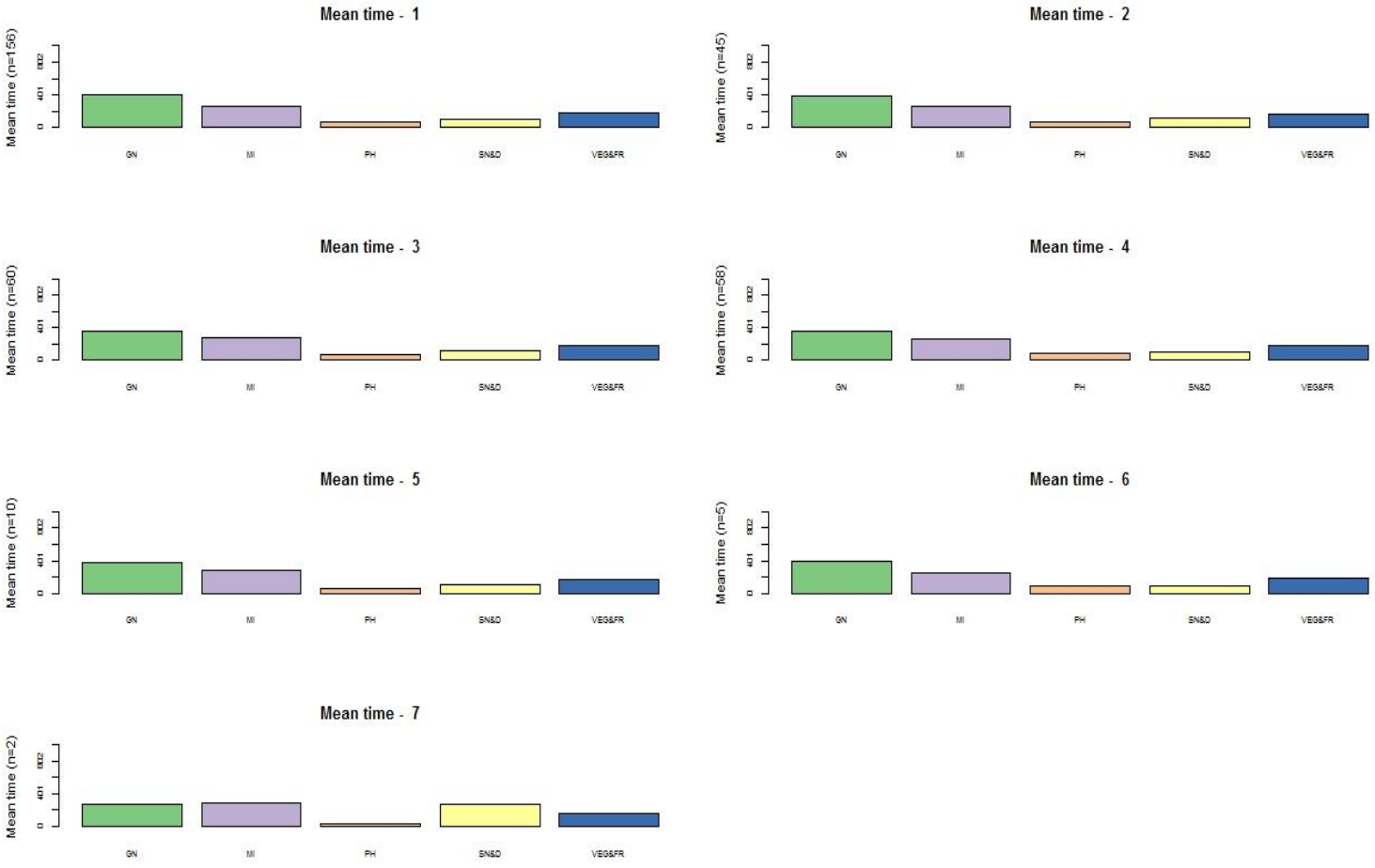


Figure 8-Mean time spent in each state with respect to Household Size

6.1 Entropy Analysis of Sequences

Entropy is a measure of diversity of states within a sequence. The entropy is 0 when all cases are in the same state and is maximal when we have the same proportion of cases in each state. The entropy can be seen as a measure of the diversity of states observed at the considered position. Plotting the transversal entropies can be useful to find out how the diversity of states evolves along the time axis.

Results & Conclusion from Entropy Analysis

The entropy for 336 sequences is summarized as follows:

Entropy
Min. :0.6132
1st Qu.:0.8541
Median :0.8846
Mean :0.8762
3rd Qu.:0.9095
Max. :0.9734

State distribution of sequence having the maximum entropy:

SEQ	GN	MI	PH	SN&D	VEG&FR
211	321	184	175	177	145

State distribution of sequence having the minimum entropy:

SEQ	GN	MI	PH	SN&D	VEG&FR
44	682	183	36	33	68

From the above data it can be seen that sequence 44 has a large number of shopping state as General, therefore has the least diversity i.e. entropy. On the other hand, sequence 211 has all the categories equally distributed, therefore has the maximum diversity i.e. entropy. The overall distribution of entropy can be seen in Figure-9.

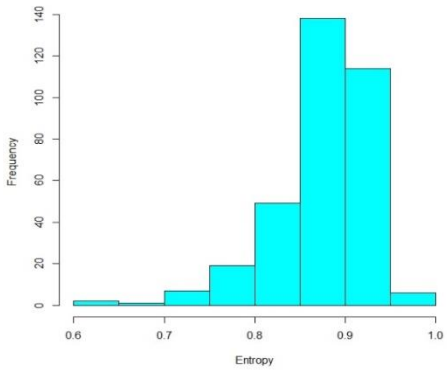


Figure 9-Entropy Distribution of the Sequence

From Figure-9 it can be observed that all the sequences have a high entropy and the maximum number of sequences lie in the range of 0.85-0.90. This entropy distribution concludes that the shopping sequences are diverse in nature.

We also plot the entropy distribution with the covariates of the customers in Figure 10, 11, 12 and 13.

Figure 10-Age vs Entropy Distribution

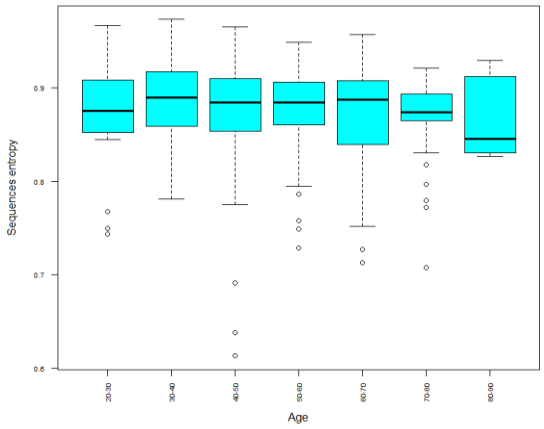


Figure 11-Household Size vs Entropy Distribution

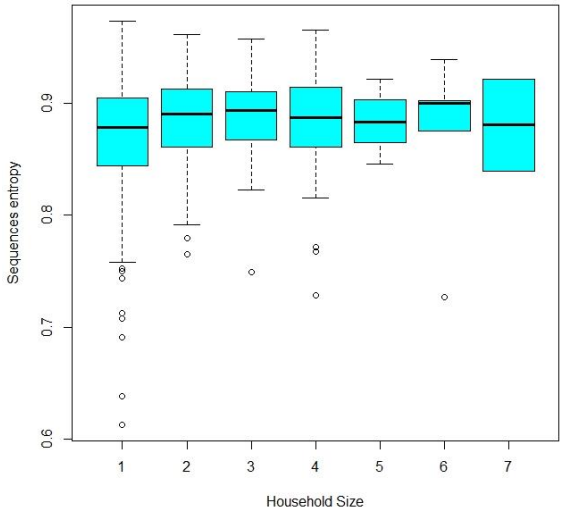


Figure 12-Euros Spent vs Entropy Distribution

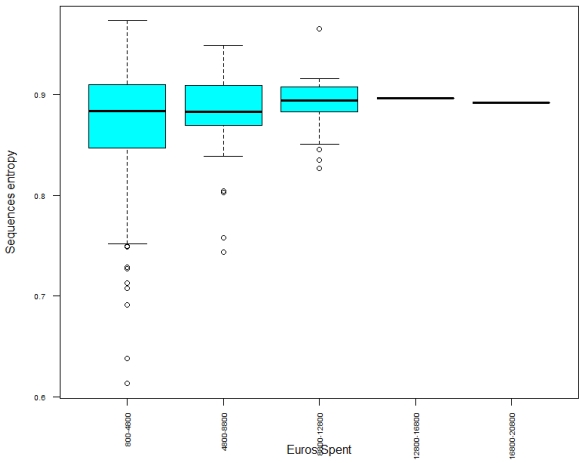
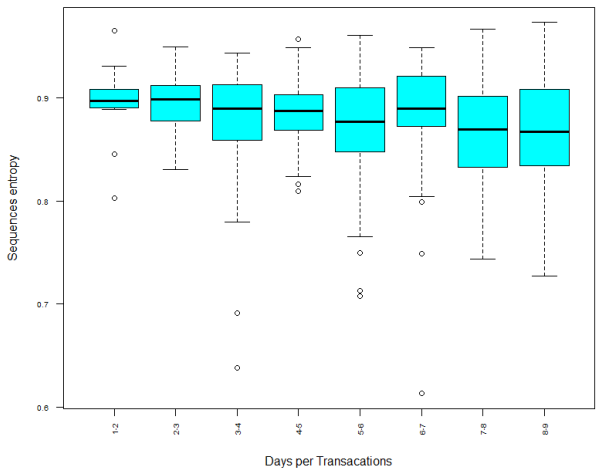


Figure 13-Days per Transaction vs Entropy Distribution



From the above figures the entropy variation forms a visible pattern with the days per transaction covariate. We can see that as the days per transaction increase the entropy decreases, which is expected because the customers shopping more frequently from the supermarket tend to change their state more frequently and thus increasing the diversity.

6.2 Clustering of Sequences

Clustering is an exploratory data analysis method aimed at finding automatically homogeneous groups or clusters in the data. Once we have identified the target segments for categories, the next important step is to understand every consumer and their needs. As discussed before, the consumers in supermarket come in all shapes and forms, therefore it is very difficult to cluster them on the basis of type of products purchased. A better and more efficient way to group customers is by analyzing their shopping trajectory i.e. the shopping sequences. The method has typically been used in combination with Optimal Matching (*OM*) distances to identify distinct groups of sequences with similar patterns; that is, to define a typology of sequences. OM distances is an optimum way to measure the cost required to change one sequence to another by shifting of a state or insertion and deletion of a state. Sequence with least distances are grouped together.

Results & Conclusion for Sequence Clustering

Sequences were clustered from 3 to 8 clusters. In order to determine the best clustering solution the results obtained were characterized using histograms and tukeys difference plot. A 3-cluster solution results in the best characterization of groups of customers. Figure-14 & 15 show the histogram and tukeys plot respectively.

Figure 14-Distribution of categories across 3-clusters

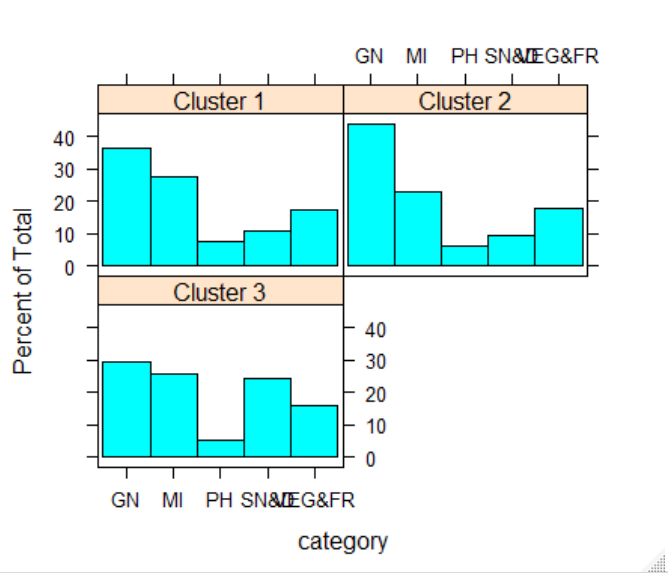
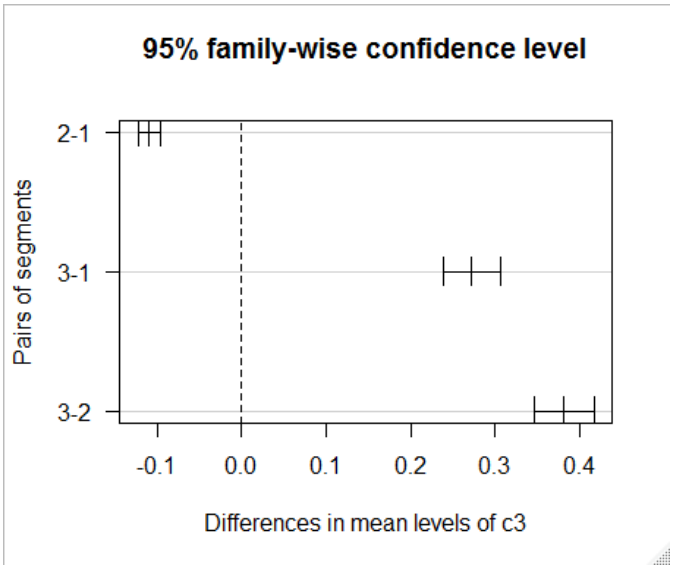


Figure 15-Tukeys honest significant differences between 3-clusters



From the above figures the clusters can be labelled as follows:

- **Cluster-1**: Customers buying products from all the categories.
- **Cluster-2**: Customers majorly buying products from General category.
- **Cluster-3**: Customer majorly buying products from Snacks and Drinks category.

We also plot Tukeys honest significant differences for 3-cluster solution based on age, household size and entropy. The plots are shown in Figure 16, 17 and 18.

Figure 16-Household Size

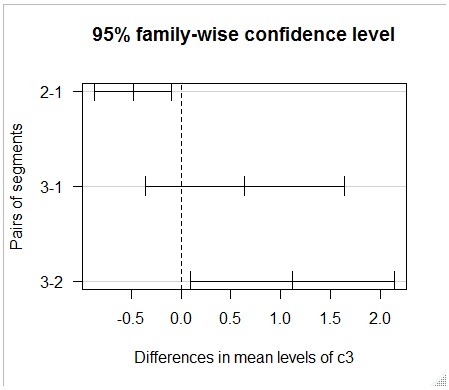


Figure 17-Age

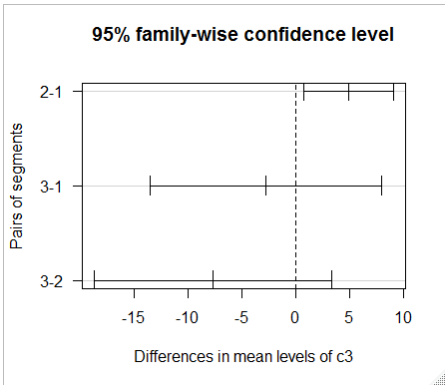
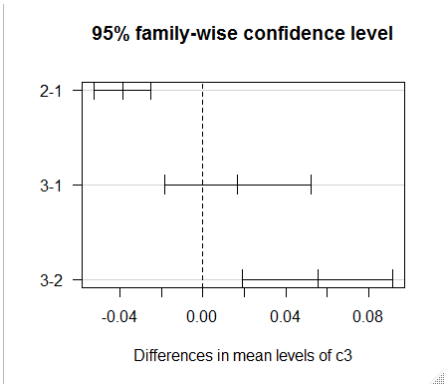


Figure 18-Entropy



From the above plot it is clear that the mean household size of people in cluster 2 is less than cluster 1 and 3. Customers of all age are part 3-clusters and there is no significant difference between the clusters on the basis of age. Cluster 2 has the least entropy among all the clusters which also can be seen from Figure-12 as the sequences have a large number of *General* state reducing the diversity.

7 State prediction using Markov Chains

Markov model is a stochastic model used to model randomly changing systems. The probability of the future event in the sequence depends only on the current state attained and not on the events that occurred before it. The **markovchain** package in R is useful for analyzing discrete time Markov chains, homogeneous and simple inhomogeneous ones as well as continuous time Markov chains. We use the **markovchainFit()** function, which returns fitted Markov chain for a given sequence or set of sequences. The model generates a 2-D transition matrix representing probability of transition from one observable state to another. Four methods are available of markovchain package: maximum likelihood, maximum likelihood with Laplace smoothing, Bootstrap approach, maximum a posteriori. For the shopping sequence dataset we go implement the model with the maximum likelihood method. Further, the trained model on 336 sequences is used to predict the latest shopping state of the customers depending on the last 2nd shopping state.

Results & Conclusion from Markov Chain Fit

Firstly, we determine the baseline efficiency by formulating the confusion matrix. Baseline efficiency is calculated by considering the most frequent outcome as the predicted state. The baseline efficiency comes out to be **40.47%**. The prediction accuracy from the Markov chain model comes out to be **83.631%**. Markov Chain gives a high accuracy for the shopping sequences and can be a very effective way to determine the shopping state of consumers visiting the supermarket.

Separate Markov models were fitted for the clusters of sequences. The prediction efficiency for the 3-clusters are as follows: **1-84.44%, 2-80%, 3-100%**. The weighted mean comes out to be 83.631% which is exactly similar to the prediction accuracy for the model of pooled sequences. Therefore, Markov model generated for the clusters and pooled sequences is similar.

8 State prediction using Hidden Markov Models (HMMs)

Hidden Markov model is a variation of Markov model in which the sequence data consist of observed states, which are regarded as probabilistic functions of hidden states. A discrete first order hidden Markov model for a single sequence is characterized by the following:

- **Observed state sequence**- with observed states
- **Hidden state sequence**- with hidden states
- **Initial probability vector**- the probability of starting from a particular hidden state.
- **Transition Matrix**- 2-D matrix representing the probability of moving from one hidden state at time t-1 to a hidden state at time t. We only consider homogeneous HMMs, where transition probabilities are constant over time.
- **Emission Matrix**- 2-D matrix representing the probability of a particular hidden state emitting a particular observed state.

Also, as mentioned before, the observation at time t is only dependent on the current hidden state, not on previous hidden states or observations. We can also fit the same HMM for multiple sequences. Each sequence will have its own hidden state sequence. The estimation process starts by giving initial values to the estimates. Good starting values are needed for finding the optimal solution in a reasonable time. In order to reduce the risk of being trapped in a poor local maximum, a large number of initial values should be tested. In order to find the best hidden state sequence, we use a dynamic programming method, the **Viterbi Algorithm.**

The hidden states are characterized on the basis of entropy of the sequences (*customers*). Further, we predict the last observable state using the generated model and calculate the prediction accuracy.

Results & Conclusion from the Hidden Markov Model

Hidden Markov Model is generated by varying the number of hidden states from 5 to 25. For each model we determine the prediction efficiency and the log likelihood. The results obtained when a single HMM is fitted for all the sequences are as follows:

Table 3-HMM with different hidden states and their prediction accuracy

Hidden States	Prediction Accuracy	Log Likelihood
3	45.5357%	-284417.6
4	69.3452%	-231236.8
5	72.0238%	-229176.6
7	70.8330%	-226707.1
10	78.5714%	-196633.0
20	80.9524%	-192542.6
25	78.2738%	-192297.5

A normal trend which can be observed from the above data is that as we increase the hidden states the prediction efficiency increases and is maximum at 20 hidden states. The prediction efficiency comes out to be 80.9524%.

It is also observed that Viterbi Algorithm gives different results for every run as it converges on a different point. The difference in results is not significant and can be seen in the table below.

Table 4-Prediction Accuracy of HMMs for different runs

Hidden States	Run 1	Run 2	Run 3	Run 4
5	70.8330%	71.4286%	67.2619%	72.0238%
20	78.8690%	80.9524%	76.7857%	79.1667%
25	74.4048%	78.2738%	75.8929%	76.7857%

We also determine the prediction efficiency for the clustering solution. A separate HMM is fitted for each cluster. This analysis is only done for 5 hidden states. The results are shown in the table below.

Table 5-Clusterwise prediction accuracy for 5 hidden state HMM

Run	Cluster 1	Cluster 2	Cluster 3	Weighted Mean
Run 1	72%	64%	90.9091%	70.2381%
Run 2	72.889%	62%	100%	70.7357%
Run 3	72%	70%	90.9091%	72.0238%
Run 4	72.444%	66%	90.9091%	71.1310%

From Table-4 and Table-5 it can be seen that the prediction accuracy of HMM for clusters and that for a pooled sequence is almost similar. Therefore, Hidden Markov model generated for the clusters and pooled sequences is similar.

8 Conclusion

The project successfully analyzes the three major aspects of shopping sequences in a supermarket.

a. ***Market Segmentation based on shopping trips & identifying the target segment***

   The shopping trips (*baskets*) can be divided into 5 clusters namely- general, milk, paper & health products, snacks & drinks, vegetables & fruits. General category has the highest number of basket followed by milk, vegetables & fruits, snacks & drinks and paper & health products at the last. Segment Level Stability Analysis (SLS) shows that milk, vegetables & fruits, general are stable or naturally occurring segments while snacks & drinks and paper & health products are vulnerable and artificially created segments.

b. ***Shopping state sequence analysis of the customers and clustering similar trajectories***

   Customers exist in all shapes and forms and therefore cannot be grouped with respect to type of category purchased. In order to group customers, we analyze their shopping sequences and group them on the basis of similar trajectories. We obtain 3-different clusters of customers namely- customers having all the categories evenly distributed in there state sequence, customers having general category as there major distribution in the state sequence and finally the customers having snacks & drinks category as a major state distribution. The size of 3-clusters is 225, 100, 11 respectively.

c. ***Future shopping state prediction using Markov chains and Hidden Markov models (HMMs)***

   Markov chains and Hidden Markov models predict the shopping state of customer based on his/her previous shopping state. Analyzing the need of customers is the major need in today's time. We successfully capture a prediction efficiency of 83.631% with Markov chains and 80.951% with HMM which is much higher as compared to the baseline efficiency.

9 Future Prospects

For future research, we suggest combining the current analysis with consumer behavior in order to develop an analysis that is capable of improving operational efficiency, satisfying the consumers and increasing the sales. A smart supermarket analysis can be developed which incorporates the effect of store environment and consumer motion inside the store. A more advance and varied algorithms can be developed, in order to support both the customers and retailers and maximizing the profit. Automating the sales process in a supermarket will be one of the major targets in future.

10 References

1. Dolnicar S, Leisch F (2017) Using segment level stability to select target segments in data-driven market segmentation studies. Mark Lett 28(3):423–436.
2. Ernst D, Dolnicar S (2018) How to avoid random market segmentation solutions. Journal of Travel Research 57(10): 69–82.
3. Bremaud P (1999). “Discrete-Time Markov Models." In Markov Chains, pp. 53-93. Springer.
4. Ching WK, Huang X, Ng MK, Siu TK (2013). “Higher-order markov chains." In Markov Chains, pp. 141-176. Springer.
5. In-Chul Jung, Young S. Kwon, Yung-Seop Lee (2012). “A Sequence Pattern Matching Approach to Shopping Path Clustering”, Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management.
6. Hui, S. K., Bradlow, E. T. and Fader, P. S. (2009), “Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping path and purchase Behavior”, Journal of consumer research, 36, 478-493.
7. Giorgio Alfredo Spedicato, Tae Seung Kang (2014). “The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R”.
8. Alexis Gabadinho, Gilbert Ritschard, Nicolas S. M□uller, Matthias Studer (2011). “Analyzing and Visualizing State Sequences in R with TraMineR”, Journal of Statistical Software, Volume 40, Issue 4.
9. Sara Dolnicar, Bettina Grün, Friedrich Leisch (2018 Edition). “Market Segmentation Analysis, Understanding It, Doing It, and Making It Useful”, Springer Open.
10. Satu Helske, Jouni Helske. “Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R”.
11. Himmelmann L (2010). HMM-Hidden Markov Models. R Package Version 1.0, URL http://CRAN.R-project.org/package=HMM.
12. Helske S (2017c). Visualization tools in the seqHMM package. URL https://cran.r-project.org/web/packages/seqHMM/vignettes/seqHMM_visualization.pdf.
13. Chia-Ruei Liu, Huei-Yuan Duan, Po-Wei Chen, Li-Hua Duan (2018). “Improve Production Efficiency and Predict Machine Tool Status using Markov Chain and Hidden Markov Model”, 8th International Conference on Computer Science and Information Technology (CSIT).
14. Oscar Gonzalez-Benito, Michael Greatorex, Pablo A. Munoz-Gallego (2000). “Assessment of potential retail segmentation variables An approach based on a subjective MCI resource allocation model”,

Journal of Retailing and Consumer Services 7 (2000) 171-179.

15. Venkatesh Shankar, J. Jeffrey Inman, Murali Mantrala, Eileen Kelley, Ross Rizley (2011). “Innovations in Shopper Marketing: Current Insights and Future Research Issues”, Journal of Retailing 87S (1, 2011) S29–S42.