American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
http://asrjetsjournal.org/

Detection of Expenditure Trends in the Telecommunication Sector

Ayse Humeyra Bilge[a], Arif Selcuk Ogrenci[b] , Huseyin Carpanali[c],

Esra Agca Aktunc[d], Fatma Atas[e]* , Tarkan Ozmen[f], Burak Erkan Kaya[g]

[a,b,c,g] *Kadir Has University, Istanbul 34083, Turkey*

[d] *Yeditepe University, Istanbul 34755, Turkey*

[e,f] *Turkcell Technology Research and Development Company, Istanbul, Turkey*

[a]*Email: ayse.bilge@khas.edu.tr,* [b]*Email: ogrenci@khas.edu.tr,* [c]*Email: huseyin.carpanali@khas.edu.tr*

[d]*Email: esra.agca@khas.edu.tr,* [e]*Email: fatma.atas@turkcell.com.tr,* [f]*Email: tarkan.ozmen@turkcell.com.tr*

[g]*Email: burak.kaya@khas.edu.tr*

Abstract

In the telecommunication sector, particularly in the cellular phone service area, customer expenditures have been in the areas of voice, short messages, and internet usage, leading to a pattern of more or less regular monthly bills. Recently, telecommunication companies started to associate retail stores to their billed commercial activities, resulting in unusual variations in the monthly payment sequences of their customers. In the present work we propose a method for detecting retail expenditure in monthly bills. We then code the information of the discretized version into a binary hierarchical tree and we classify them as positive or negative with respect to complaint potential.

Keywords: Retail; Telecommunication Sector; Hierarchical Clustering; Distance-Based Clustering.

1. Introduction

Nowadays, many GSM operator companies, worldwide operate in many sectors such as healthcare, payment, retail and automotive, as well as telecommunication services. These new sectors unrelated to telecommunications contribute to the formation of different marketing strategies for operator companies. This new strategy was initiated to allow the variety and the number of customers to change positively. This positive change affects operator companies that adopt new marketing strategies in all business settings. On the other hand, products sold within the scope of retailing, develop problems in systems such as complaint estimation and churn analysis. Telecommunication companies may adopt a variety of retail strategies.

---

* Corresponding author.

For example, Telefonica, Deutsche Telekom, Orange, and AT&T that are among the largest operator companies in Europe and USA, sell their customers' products related to the telecommunication sector, such as phones, tablets, and computers. A fee is charged to the invoice of the customer who purchases the product through this post-sales payment plan. The collection of the price of the products sold via the GSM invoice causes unusual changes in the monthly invoice amount of the customers. On the other hand, Turkcell, which is a major communications company established in Turkey, offers a variety of products and services in technology, retail, health, and payment system areas. In particular, a recent retail service called Pasaj launched in 2021, offers many products used in daily life, such as phones, computers, tablets, and white goods. Since the postpaid commercial activities of all retail sales made through the Pasaj are associated with the GSM invoice, unusual changes have started to occur in the monthly payment amounts of customers.

Turkcell has approximately 26 million customers in all sectors it serves. Customers are divided into two segments, with a portfolio of approximately 8 million corporate and 18 million individual customers. Monthly bills are issued four times a month, leading to a load of about 7 million bills each time. After receiving their bills, customers forward problems that they encounter in their invoices to their customer representatives, who are required to handle quite a large number of complaints in a short period. Thus, an efficient estimation of complaints is necessary for customer satisfaction. In the present work, as an initial step for complaint estimation we develop a method for detecting anomalies in the total amount of the bill and identifying the ones arising from retail expenditures.

After customers purchase products within the scope of retailing offered by Turkcell, installments created for the product are reflected on the customers' invoices. This leads to a sudden increase in the bill and unusually high invoices for the period of the instalments. Thus, retail expenditure results in a large upward step in the time series of bills with otherwise relatively low standard deviation. In this study, we apply this idea to the 9-month billing data of the corporate group to detect retail expenditure. The method we propose is the first step towards more refined screening. For example, retail expenditure that usually has an installment plan results in a plateau of high bills, while an unusual telecommunication expenditure such as roaming results in a single peak. The method proposed in this paper cannot distinguish these two types of anomalies. Nevertheless, anomalies due to unusually high telecommunication usage are less common and our method is still useful in detecting retail expenditures, which seem to be becoming a popular trend.

2. Literature Survey

The incorporation of the installments of retail sales to billed services is an emerging commercial application, that has not been studied widely in the literature. In the telecommunication area, phone companies sell mobile phones and related equipment to their customers by reflecting installments to their bills [1]. The introduction of a new service item to the bills may give rise to customer complaints and it is desirable to discriminate between customers that make different types of expenditure, hence obtaining a segmentation of customers with respect to the selected criterion. The key element of customer segmentation is a preliminary step for the estimation of customer complaints and cluster analysis is commonly used for segmentation. Cluster analysis consists of grouping similar observations or data points into clusters based on their similar characteristics. It is used in

pattern analysis, and decision-making, including data mining, document retrieval, and pattern classification. Clustering can be based on the "dissimilarity" of the objects to be classified. For this it is necessary to start with a measure of dissimilarity and group objects based on their distances from each other [2]. Alternatively, "hierarchical clustering" consists of subdividing the population based on selected criteria [3]. Hierarchical clustering solutions, also known as "dendograms" provide a view of the data at different levels of abstraction [4].

For obtaining hierarchical clustering solutions, agglomerative algorithms are generally used (3,5,6,7) The method of the agglomerative algorithms is to assign each object to its cluster and then merge pairs of clusters until the whole tree is constructed (4,7)

Other than that, partitional algorithms (2,8,13) can also be used to achieve hierarchical clustering solutions by using a sequence of repeated bisections. Various authors (13,16) stated that using partitional clustering algorithms for clustering large datasets is convenient due to their requirement of relatively less computational power. The roots of hierarchical clustering go back to as early as the 1950s, with the single linkage method [17]. This method finds clusters at a series of increasing distance thresholds which can be shown as (d_1, d_2, d_3, …). The clusters at level d_i are found by dividing the samples into sets that contain all segments of length d_i or less. The sets are called clusters of level d_i [18]. Early hierarchical clustering works can be seen in biological taxonomy around the 1950s and 1960s [5], which also benefited from single linkage clustering. A common goal of hierarchical clustering is to derive a partition.

In the present work we use a hierarchical clustering by refining the partitioning of the population into subgroups, according to criteria with a hierarchy of importance attributes.

3. Description of Data

In the present study, we work with a sample of 1 713 995 corporate customers. Data consists of their monthly bills for the period January 2021-September 2021. In this pool, customers with very high invoices appear as outliers, and for the clarity of visual presentations, in the present work we exclude 3060 customers whose monthly average bill is above 1000 TL.

Data consists of the time series of 9-month bills of more than 1.7 million customers. Instead of applying distance-based clustering methods to the full dataset, we use a hierarchical clustering based on 3 features extracted from the 9-month time series. These features are, the average value of monthly bills (denoted by m), the standard deviation of the monthly bills, divided by the mean (denoted by s), and the difference between the maximum and the minimum of the monthly bills, i.e, the range, divided by the mean (denoted by r). A scatter plot of these features is given below.

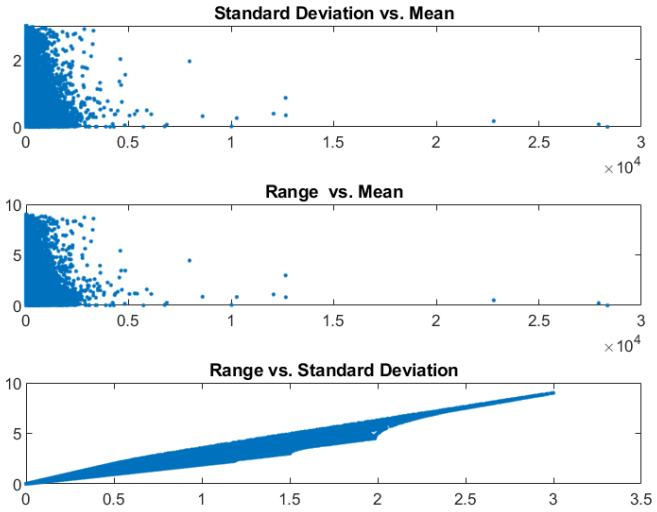

Figure 1: Scatter plot of the mean, the normalized standard deviation, and the normalized difference between maximum and minimum values of monthly bills, the normalized range.

From this figure, one can see that the range and the standard deviation are closely correlated, as expected. On the other hand, except for a few cases, scatter plots of normalized standard deviation and normalized range versus the mean are accumulated. As there is no obvious clustering structure, we subdivide the spans of the mean (m), the normalized standard deviation (s), and the normalized range (r) into bins. This subdivision will be based on the frequency of distribution of these features, as given by the graph below.

As seen in Figure 2, sorted values also have no obvious jumps or plateaus, indicating the lack of an obvious clustering structure. The only exception is in the monthly mean, displaying a jump at 1.5 TL. This threshold separates inactive and active customers; actually, inactive customers constitute about 40% of the total pool.

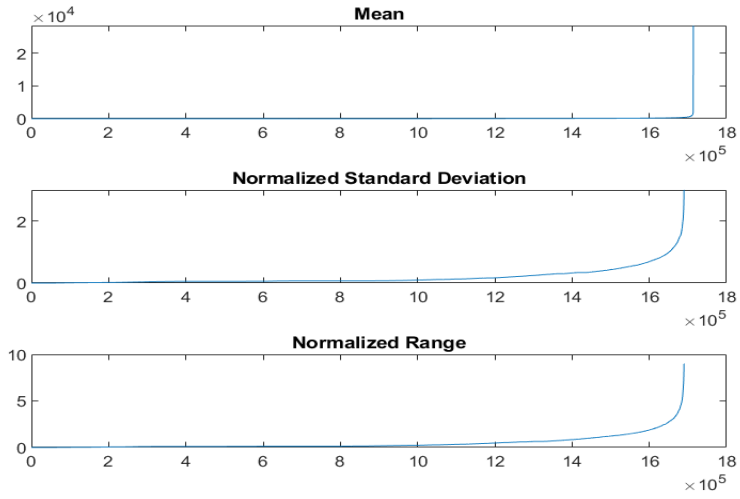

Figure 2: Sorted values of the mean, the normalized standard deviation, and the normalized range.

4. Hierarchical Clustering

In the present work, we divide the values of each of the 3 features, the average value of monthly bills (m), the standard deviation of the monthly bills, divided by the mean (s), and the difference between the maximum and the minimum of the monthly bills, divided by the mean (r), into 4 bins, by assigning 3 thresholds.

The subdivision of the monthly mean values of the bills is based on billing practices. Thresholds for the normalized standard deviation are chosen as 1, 2, and 3. The range of monthly bills is used as an indicator of the retail component of the bill, and thresholds are chosen accordingly. These values are presented in Table 1 below.

Table 1: Thresholds for the features.

	T_{min}	T_1	T_2	T_3	T_{max}
Mean	0	1.5	50	500	28 348
N. Std	0	1	2	2.5	3
N. Range	0	0.5	2	4	9

The hierarchy for clustering will be based first on the mean of 9-month bills, then their standard deviation and finally on the range. For this purpose, we sort data according to the mean. In Figure 3, below, we display monthly invoices, their means, standard deviations, and ranges. From this figure, one can see that average bills of about 1.5 TL form a plateau, justifying the choice of the threshold value T1=1.5 for the mean. In the same figure, we can see that there are invoices with large standard deviations and a large range in all mean invoice levels. But in cases where the mean of the bills is very low, high values of relative standard deviation and relative range are meaningless, in particular with respect to the potential for complaints. Thus, although we will keep track of these cases, they will be identified in the binary tree scheme.

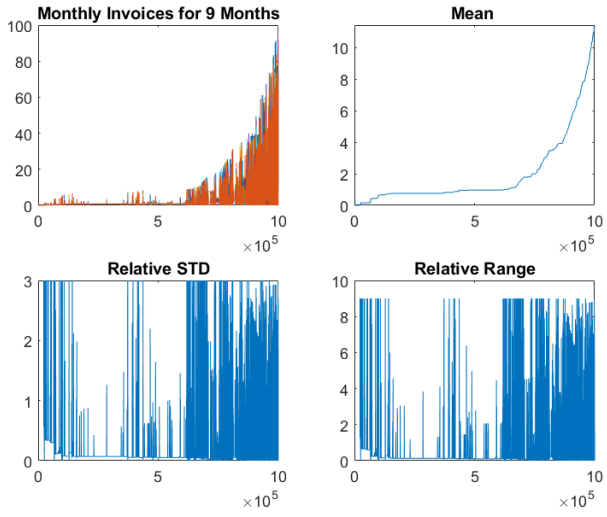

Figure 3: Monthly invoices for institutional customers; Average, standard deviation and the difference between maximum and minimum of monthly invoices for 1 million customers with the lowest average invoice.

We will hierarchically classify the features, each feature being subdivided into 4 categories. This will lead to 4^3=64 groups. We use the following terminology to represent each group by a 3-digit number. Let X, Y, and Z, ranging from 1 to 4 be the group numbers for the mean, the normalized standard deviation and the normalized range. For example, if we have a customer with mean invoice m=20, normalized standard deviation s=2.3, and normalized range 1.7, then according to Table 1, this customer belongs to the second group with respect to the mean, to the third group according to the normalized standard deviation and to the second group according to the normalized range. Therefore X=2, Y=3 and Z=2. We represent the group of this customer by the 3-digit number 232.

Table 2: The number of elements in each of the digitized bins. Upper values are the codes of the lower values are the number of samples in each bin.

111 581399	112 62897	113 1678	114 0		211 339572	212 120039	213 19308	214 0
121 0	122 0	123 2724	124 362		221 0	222 28	223 13535	224 5990
131 0	132 0	133 0	134 198		231 0	232 0	233 0	234 1328
141 0	142 0	143 0	144 794		241 0	242 0	243 0	244 689

311 227291	312 205959	313 19515	314 0		411 5544	412 7627	413 640	414 0
321 0	322 28	323 8687	324 3432		421 0	422 36	423 698	424 251
331 0	332 0	333 0	334 448		431 0	432 0	433 0	434 91
341 0	342 0	343 0	344 124		441 0	442 0	443 0	444 62

The 3-digit numbers for each group are represented as XYZ, each digit ranging from 1 to 4, and the numbers of samples in each group are given in Table 2. Note that some of the groups are empty, mainly because if there is a sudden high jump in the time series, the standard deviation cannot be too low.

5. Discussion of the Results

In this section we present the data for selected groups. Figure 4 displays data for X=1, all values of Y and Z. This group which includes invoices with very low average values, corresponds to customers who are not expected to make any complaint. In Figures 5 and 6, we present two examples from the group X=2, with moderate average bills. Customers in the group with Y=1, Z=1 are not expected to make any complaint while customers in the group Y=2, Z=4 might be associated with retail purchases.

Then in Figures 7 and 8, we repeat the same presentation for the group X=4, with high average bills. Customers in the group with Y=1, Z=1 are not expected to make any complaints, while customers in the group Y=2, Z=4 might be associated with retail purchases. Finally, representative time series for each group are presented in Figure 9.

5.1. The Group X=1

Customers with low average bills in group X=1 are presented as a single entity. The graph of the data in the bins with X=1 is below. We see that despite high values of the standard deviation and the range, monthly invoices are below 1.5 TL. No objections to these invoices are expected.

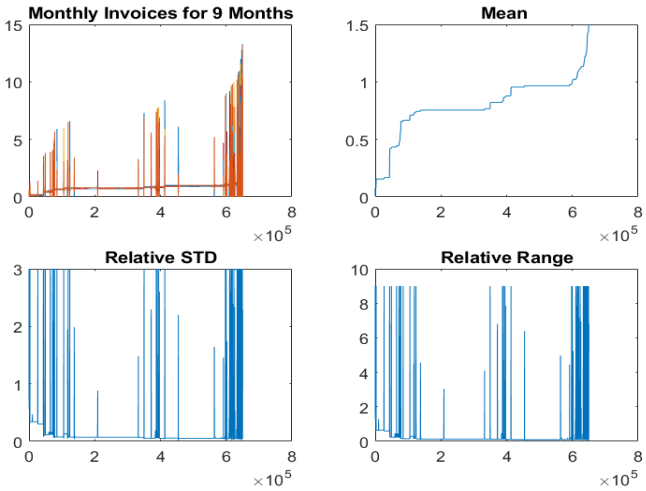

Figure 4: Invoices with a mean less than 1.5 TL. No objection is expected.

5.2. The Group X=2, Y=1, Z=1

Another group for which no objection is expected is the group with a relatively low average invoice (X=2) and low relative standard deviation (Y=1), and low relative range (Z=1).

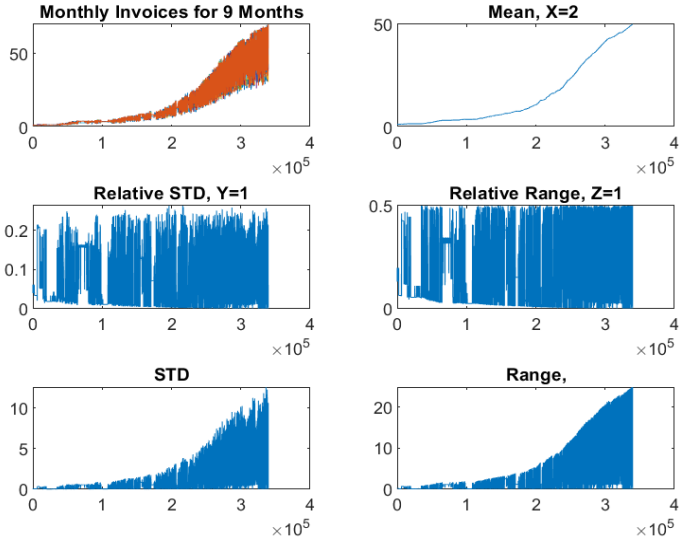

Figure 5: Invoices with a mean greater than 1.5 and less than 50, relative standard deviation less than 1, and relative range less than 0.5. No objection is expected.

5.3. The Group X=2, Y=2, Z=4

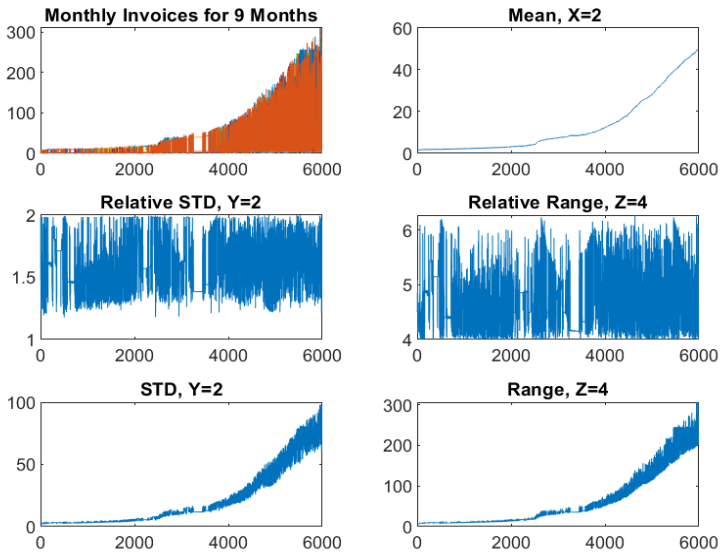

Figure 6: Invoices with a mean greater than 1.5 and less than 50, relative standard deviation between 1 and 2, and relative range greater than 4. Data is associated with retail expenditure.

5.4. The Group X=4, Y=1, Z=1

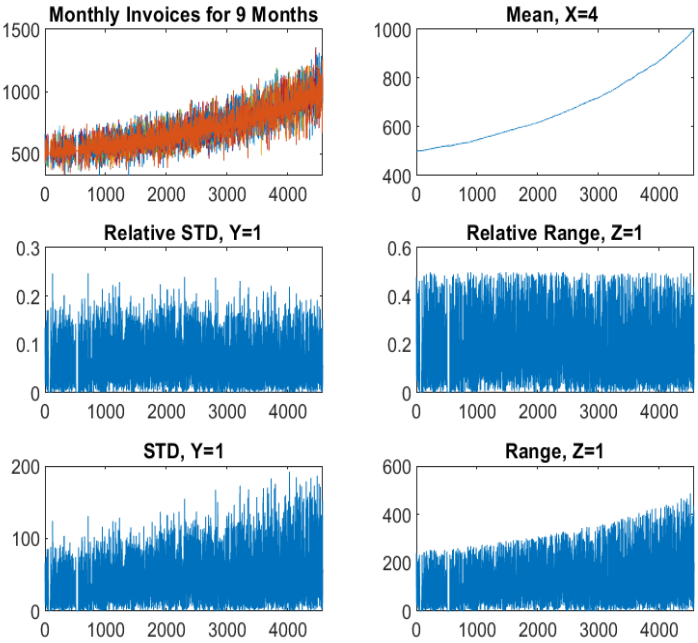

Figure 7: Invoices with mean greater than 500, relative standard deviation less than 1, and relative range less than 0.5. No objection is expected.

5.5. The Group X=4, Y=2, Z=4

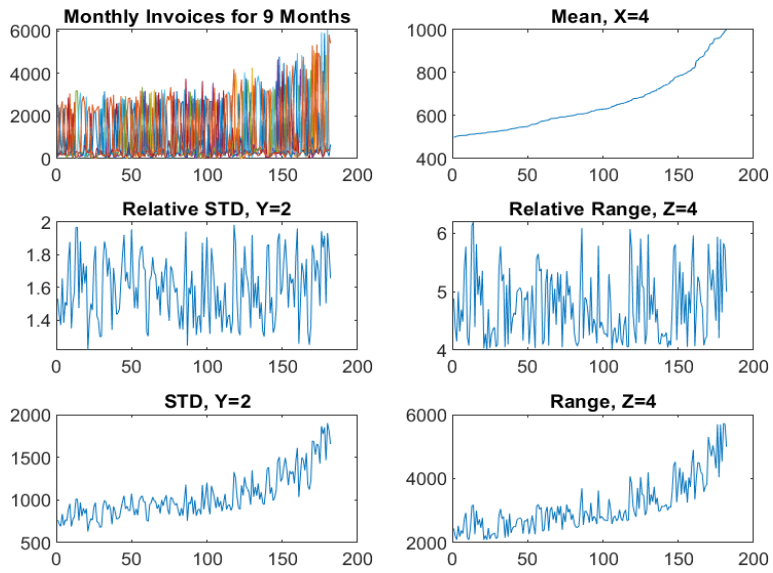

Figure 8: Invoices with a mean greater than 500, relative standard deviation greater than 2, and relative range greater than 4. Data is associated with retail expenditure.

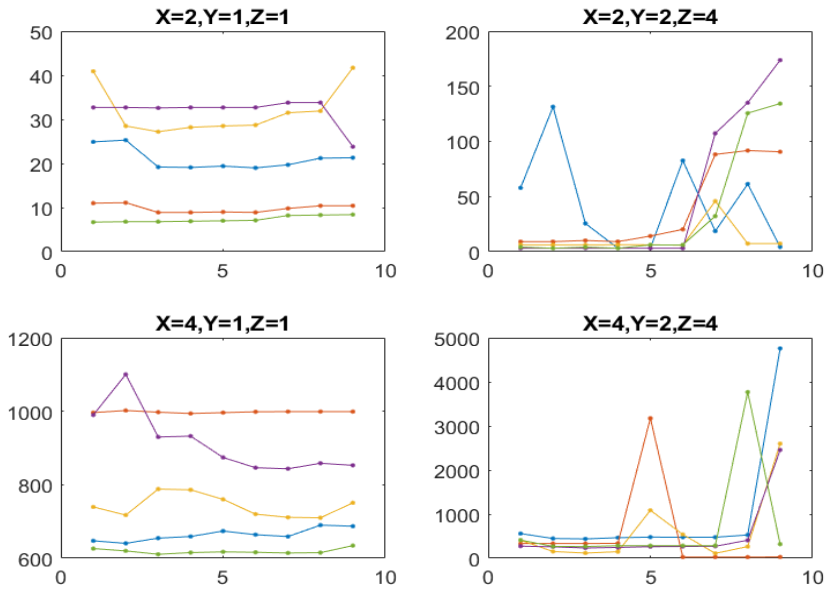

Figure 9: Selected time series of invoices for the types, as indicated.

6. Conclusion

In this work we clustered customers in the telecommunication sector, based on the time series of their monthly bills. We used the 9-month bills of more than 1.7 million corporate customers of the Turkish telecommunication

company Turkcell. The clustering method aimed to detect retail expenditure, characterized by a sudden increase in otherwise stationary time series data. For this purpose, we applied hierarchical clustering by subdividing first with respect to the average bill, then with respect to relative standard deviation and finally with respect to the relative range of monthly bills. Each feature was subdivided into 4 classes, separated by threshold values based on the characteristics of the data. The scheme that we applied resulted in 4^3=64 groups, some of which being empty, in our case. The groups with low standard deviation but high range were identified as reflecting retail expenditure.

In the present work we used 3 criteria that were divided into 4 subgroups to obtain a clustering of the customers. In a general application of the method, the number of subgroups need not be the same for each criterion; one can use as many subgroups as needed, provided that these are separated by suitable threshold values. With the new separation to be created, a more realistic division can be achieved.

The groupings that were obtained constitute a basis for classifying customers into segments with respect to criteria that the company would be interested in. For example, if the company is interested in whether a certain customer will file an objection to a given bill or not, whether the current bill falls in the clusters X=1, corresponding to a very low total bill, is irrelevant. Thus, the hierarchical clusters that we obtained can be used for various classifying schemes.

Finally, we should also note that the main goal of our method used in the study is to identify a target group with minimal information. For example, even if detailed information on specific items of a bill is obtainable, our purpose is to detect and discriminate unusual expenditures from aggregate information, without reference to billing details. In this sense, the method is expected to be applicable to detecting billing anomalies in different sectors.

Acknowledgements

The study is funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) in the scope of the University Industry Cooperation Support Program with grant number 5210098.

References

[1] T.-Y. Ou, W.-L. Tsai, Y.-C. Lee, T.-H. Chang, S.-H. Lee, and F.-F. Huang. “A recommendation model for selling rules in the Telecom Retail Industry,” *Axioms*, vol. 11, no. 6, p. 265, 2022.

[2] A. K. Jain and R. C. Dubes. “Algorithms for clustering data.” *Englewood Cliffs*, NJ: Prentice Hall, 1988.

[3] S. Guha, R. Rastogi, and K. Shim. “Cure,” *ACM SIGMOD Record*, vol. 27, no. 2, pp. 73–84, 1998.

[4] Y. Zhao and G. Karypis. “Empirical and theoretical comparisons of selected criterion functions for document clustering,” *Machine Learning*, vol. 55, no. 3, pp. 311–331, 2004.

[5] S. P. H. A. and R. R. Sokal. “Numerical taxonomy: The principles and practice of numerical classification.” *San Francisco*: Freeman, 1973.

[6] B. King. “Step-wise clustering procedures,” *Journal of the American Statistical Association*, vol. 62,

no. 317, pp. 86–101, 1967.

[7] G. Karypis, Eui-Hong Han, and V. Kumar. “Chameleon: Hierarchical clustering using dynamic modeling,” *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[8] MacQueen, J. "Classification and analysis of multivariate observations." *5th Berkeley Symp. Math. Statist. Probability*. 1967.

[9] Cheeseman, Peter C., and John C. Stutz. "Bayesian classification (AutoClass): theory and results." *Advances in knowledge discovery and data mining* 180 (1996): 153-180.

[10] Zahn, Charles T. "Graph-theoretical methods for detecting and describing gestalt clusters." *IEEE Transactions on computers* 100.1 (1971): 68-86.

[11] Strehl, Alexander, and Joydeep Ghosh. "A scalable approach to balanced, high-dimensional clustering of market-baskets." *International Conference on High-Performance Computing*. Springer, Berlin, Heidelberg, 2000.

[12] Boley, Daniel. "Principal direction divisive partitioning." *Data mining and knowledge discovery* 2.4 (1998): 325-344.

[13] Cutting, Douglass R., et al. “Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections.” *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, Jan. 1992, pp. 318-329–329. EBSCOhost, https://icproxy.khas.edu.tr:2071/10.1145/133160.133214.

[14] Larsen, Bjornar, and Chinatsu Aone. "Fast and effective text mining using linear-time document clustering." *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999.

[15] Aggarwal, Charu C., Stephen C. Gates, and Philip S. Yu. "On the merits of building categorization systems by supervised clustering." *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999.

[16] Karypis, Michael Steinbach George, Vipin Kumar, and Michael Steinbach. "A comparison of document clustering techniques." *TextMining Workshop at KDD2000 (May 2000)*. 2000.

[17] Graham, Ronald L., and Pavol Hell. "On the history of the minimum spanning tree problem." *Annals of the History of Computing* 7.1 (1985): 43-57.

[18] De, Rham, and C. De Rham. "La Classification Hierarchique Ascendante Selon La Methode Des Voisins Reciproques." (1980).