On Experimental Study of Hotel Clustering

Ömer Arifoğulları and Günce Keziban Orman

***Abstract*—As in most sectors, the development of an intelligent recommendation system in tourism becomes an important issue. Tourism agencies are putting maximum effort into suggesting the best and most valuable hotels for their customers. With the help of B2B relations between agencies and hotels, tourism agencies hold large hotel feature datasets. Summarizing or interpretation of high-quality data requires the implementation of data analysis methodologies. Tourism data is unique in terms of geography and culture. Thus, every new data set requires a dedicated analytical process. Furthermore, because raw data is in the form of a sparse binary matrix of hotel features, it poses a technical challenge to any analytical process. This paper presents a comparison of different clustering and dimension reduction methodologies for real-world hotel data of this nature. The data set represents 61% of the hotels in Turkey.**

***Index Terms*—data dimension reduction, clustering, hotel clustering, dimension reduction, tourism recommender systems.**

I. Introduction

After the long lockdown periods, the populations of some specific countries (e.g., India, Qatar and Saudi Arabia) increased their expenditures in the tourism sector compared to the pre-pandemic era. According to UNWTO (The World Tourism Organization), some countries, including Turkey, have already reached the same level of income compared between pre-pandemic and post-pandemic. The value of tourism has renewed itself and is still a big player in the countries' economies [1]. Customer tendencies about search or buying have been becoming to change with e-commerce and travel guides that published online. Therefore number of online reservation are passing number of physical store orders year by year. Via online systems, customer behaviours and order features easily collected by computer systems. This collected big-data are used for recommendation systems by travel agencies to compete each other. Travel agencies' service quality should be quick, especially during peak seasons and when competition becomes an arena duel for companies. Time and options are limited, and customers easily access different sources to compare the best-matching hotels according to their needs.

Over time, different papers have been published for the tourism sector. However, it is possible to find tourism recommendation systems or hotel clustering research in other sectors that have a high volatile customer tendency and are highly competitive, such as retailing, streaming, and e-commerce platforms. Sánchez-Pérez, M. et. al. examine the effects of vertical and horizontal differentiation to explain hotel pricing decisions, considering the moderating roles of competition and location [2].

Clustering can be a preliminary step of hotel recommendation systems [3]. At present, hotels provide many features to their customers, and the importance of the features can be a problem for the enterprises and their recommendation algorithms. Location-specific therefore, dataset specific research has been published in recent years. One of the related papers, published by Rodrguez-Victoria, O.E. et al. [4], investigated hotel clustering methodologies with Colombia hotels. Another paper has been published by Dağ, O. et. al. which focuses on Antalya, Turkey hotels in their paper [5]. Natural, cultural, and political differences between tourism-economy territories cause different options and features regarding hotels and their customers.

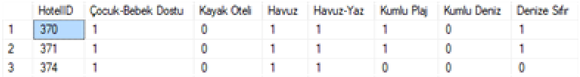
In this paper, we propose an experimental setup for determining the best hotel clustering structure for use in recommendations in the tourism industry. We have understood that previous research for hotel recommendation systems are not focused analytical explanation of hotel features and most of papers have been focused to propose a new recommendation system. On the contrary, like as e-commerce data, hotel and customer data do not tell its own tale. The information that revealed by detailed analysis, can be more useful for hotel recommendation systems. Our work is done on the hotel feature data set provided by Seturtech A.Ş. The registered touristic facility number is 4198 in Turkey [6] and our dataset covers more than 61% of those facilities. Therefore, our intact dataset may provide unexampled insight about hotel groups located in Turkey, and could be an example for the Mediterranean region. Because this is the first step in the process, rather than focusing on the interpretation of final clusters, we focus on solving analytical problems such as determining the best cluster numbers, identifying the most well-separated clusters, or simply selecting the best algorithm for hotel clustering.

Our work can be seen as an example of the data analysis process for finding the most proper clustering structure for the hotels. The challenge in this work is the nature of the raw data set. The hotel features that we work with are all binary variables. While there are plenty of metrics, algorithms, and techniques dedicated to discovering knowledge from numerical variables, the methods for processing binary ones are limited. Thus, our main contribution is to propose an experimental methodology for discovering the best clusters for the hotels, which are explained with binary features. In this methodology, we transform the sparse binary data set into numerical ones by using different dimension reduction techniques. Then, well-known clustering algorithms are applied and evaluated by various metrics.

Manuscript received March 08, 2023. This work was supported by organizations SeturTech A.Ş., sahibinden.com and the Galatasaray University Research Fund (BAP) within the scope of project number FBA-2021-1063, and titled "Niteliklendirilmiş çift yönlü ağlarda bağlantı tahmini ile öneri Sistemleri geliştirilmesi".

Ö. Arifoğulları is with the Department of Smart Systems Engineering, Galatasaray University, Istanbul, 34349 TURKEY e-mail: mr.arifogullari@gmail.com.

G. K. Orman is with the Department of Computer Engineering, Galatasaray University, Istanbul, 34349 TURKEY e-mail: korman@gsu.edu.tr.



	HotelID	Çocuk-Bebek Dostu	Kayak Oteli	Havuz	Havuz-Yaz	Kumlu Plaj	Kumlu Deniz	Denize Sıfır
1	370	1	0	1	1	1	0	1
2	371	1	0	1	1	1	0	1
3	374	1	0	1	1	0	0	0

Fig. 1. First 3 Rows and First 7 Features from Hotel Features Table

In the next chapters, we aim to explain which dimension reduction method and clustering models are less biased or show similar results on the same tourism dataset. The second chapter of this paper explains the dataset, reduction methods for features, and clustering models used in this research. In Chapter 3, we described the experimental setup and presented the results of the experiments. The final chapter consists of performance metrics, results, and discussion parts.

II. Methodology

A. Data Set and Preparation

The data set we used in this work includes hotel features as collected by SeturTech A.Ş. Overall, the dataset consists of the hotels that the Setur customers have visited at least once. There are 2561 different hotels with 27 different hotel features. These features are all binary features, showing if the related hotel has the related feature. One primary key, HotelID, is dedicated to distinguishing every single hotel from each other. There are no null values in the HotelID column, and all binary feature columns have at least one ”1” value in any row on the table. Therefore, there is no need to remove any feature columns from the feature table. Also the HotelID column has been checked for uniqueness. In this step, recurring rows have been eliminated from the table. HotelID column has been anonymized in this study to protect B2B customer interests. Nevertheless, anonymized HotelID values can be matched to the real hotel IDs by the data provider using a reverse algorithm. We show a small sample of data set in Fig. 1.

The features on Fig. 1 explain whether the corresponding hotel is child-friendly, a ski hotel, has a pool, a summer pool, a sandy beach, a sandy sea floor, is next to the shore, has a gravel beach, a gravel sea floor, and so on. In the tourism sector, most of the hotels can provide only limited features at once. Therefore, most of the rows are filled with a 0 value, which means there is no option for customers at the hotel in question. Our raw dataset is binary, sparse and consists 2561 rows and 27 columns. Sparse and binary datasets present some difficulties for clustering and Mao Y. et al propose different approach for this type of datasets [7]

B. Dimension Reduction Methods

In this study, we tackle the clustering of a data set with binary values. Since most of the clustering algorithms are dedicated to segmenting the data sets having numerical features, we first transform this data set into a numerical one. Indeed, there are several different methods for addressing this issue. In this work, we concentrate on two specific feature selection methods, Sparse Principal Component Analysis (SPCA) [8] and Non-Negative Matrix Factorization (NMF) [9] since they are not only assisting to transform a binary dataset into a numerical one but also reducing the dimension if it is necessary.

1) Sparse Principal Component Analysis: Principal component analysis (PCA) is described by Jolliffe et. al. [10] as a simple reduction of the dimension of a dataset while preserving as much statistical variability as possible. This means finding new variables that are linear functions of those in the original dataset, that successively maximize variance, and that are uncorrelated with each other. In our data set, all features consist of binary data. But, as explained in Section 2.1, our dataset has a sparse characteristic. For this reason, we have chosen a more specialized version of PCA, which is sparse PCA (SPCA). In traditional PCA, the principal components are linear combinations of all the original features, whereas in SPCA, the principal components are linear combinations of a subset of the original features. The goal of SPCA is to identify a small set of features that explain the most variance in the data, while ignoring the noise or irrelevant information. This can be particularly useful in high-dimensional datasets, where many of the original features may be redundant or uninformative.

2) Non-Negative Matrix Factorization: Non-negative matrix factorization (NMF), like PCA, is a dimension reduction technique. In contrast to PCA, NMF models are easy to understand and interpret. However, NMF can not be applied to every dataset. It is required that the sample features be ”non-negative”, so greater than or equal to 0. Non-negative matrix factorization (NMF) is a linear algebra method used for matrix decomposition and data analysis. Given a non-negative matrix V, NMF seeks to factorize it into two non-negative matrices, W and H, such that $V \approx WH$. The columns of W represent a set of basis vectors, or patterns, that are used to linearly combine the columns of H to approximate the original matrix V. The columns of H represent the weights that are assigned to each basis vector for each column of V. In other words, the matrix W defines a set of features or patterns that are common to the data, and the matrix H defines how much of each feature is present in each data point.

C. Clustering Algorithms

In this chapter, we will examine four different clustering algorithms. The interior logic of algorithms is explained to the reader in the following paragraphs. Further experimental details are shared in Section II-D.

1) K-Means: K-means is an iterative partition algorithm. Every cluster is represented by its centroid. Around these centroid points, according to neighborhood function, each cluster forms globular, non-overlapping shapes. The designer of the system should specify the number of clusters (K). The fundamental steps of the algorithm are as follows: first, K different centroids are chosen randomly from the data set as the representatives of K clusters. Second, each single data point is assigned to a cluster according to its closeness to cluster centroids. Third, the cluster centroids are reassigned. Fourth, the algorithm converges if the cluster centroids do not change between two consecutive iterations; otherwise, the second and third steps are repeated. Different K-means versions can use different distance metrics (e.g., Euclidean, Chebychev) and different algorithms (e.g., Lloyd, Elkan) for expectation maximization (EM).

2) Hierarchical Agglomerative Clustering: Hierarchical agglomerative clustering (HAC)is also an iterative algorithm. It arranges samples into a hierarchy of clusters. The number of clusters begins with the same number of rows as there are hotels, and each cluster contains one instance. Clusters merge iteratively until all nodes are labeled by a cluster. These merging processes use different distance (linkage) metrics. The most commonly used linkage metrics are single, complete, average, and ward measures. Each algorithm has its pros and cons. In a nutshell, these linkages are:

- 'Single' uses the minimum of the distances between all observations in the two sets.
- 'Complete' or 'maximum' linkage uses the maximum distances between all observations in the two sets.
- 'Average' uses the average of the distances of each observation in the two sets.
- 'Ward' minimizes the variance of the clusters being merged.

3) Density-based Spatial Clustering of Applications with Noise: Density-based Spatial Clustering of Applications with Noise (DBSCAN) is a well-known clustering algorithm. But there's no free launch and relying on DBSCAN to find the right number of clusters completely on its own. The system designer should be aware of two parameters of DBSCAN. Firstly, the epsilon parameter (ε) defines the maximum distance between points within the same cluster. Second parameter is the minimum sample value for clusters, or "how many points can be called a cluster at a minimum?". There can be some points with a wider distance than the epsilon value and lesser size than the minimum sample setting, which are called noise points.

4) Ordering points to identify the clustering structure: Ordering points to identify the clustering structure (OPTICS), which was founded in June 1999 by Ankerst, Breunig, Kriegel, and Sander, is another density-based algorithm [11]. OPTICS is a more advanced version of the DBSCAN algorithm that finds the best Epsilon value by ordering points based on their spatial values. Therefore, further algorithm details have not been included in this paper.

D. Experimental Setup

Main objective of experiments was setting up the most well-separated clustering of hotel data, which was difficult to cluster in its raw form. For this reason, we have applied two different dimension reduction techniques and four clustering algorithms that explained on previous chapters. We have focused on two important criteria of the final cluster structure when choosing both the most proper dimension reduction and the most effective clustering algorithm. These criteria are as follows: first, the clustering structure with the most cohesive clusters with high intra-similarity, and second, the clustering structure with the most well-separated clusters with high inter-distance. Calculation of inter distance and intra similarity methods and different distance metrics have been explained by Solen J. et. al. [12].

The two dimension reduction methods, SPCA and NMF are applied with the feature number from 2 to 26. The minimum dimension count has been chosen as 2, since with less than 2 dimensions, the variance of the data will be lower than 0.7 which means a high loss from the variability information of the data set. The maximum value must be less than all feature numbers.

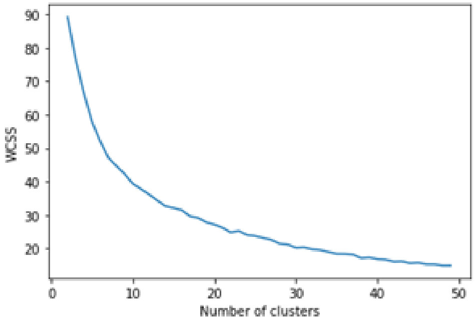


Fig. 2. WCSS Score - Cluster Numbers

Cluster numbers, K or n, have been selected between 2 and 50 for K-Means and HAC Single/Average/Complete/Ward methods according to the empirical experiments. The best cluster number is chosen by applying well-known elbow method. Figure 2 is an example of 15-dimensioned kmeans clustering with WCSS scores. This figure shows that WCSS (Within-Cluster Sum of Square) values are becoming flat for $n > 50$ values. Regarding to these experiments, k-means cluster numbers chosen between 1 and 50. As explained in DBSCAN section, epsilon (ε) variable defines the distance between nodes in the same cluster. In our work, we have determined $\varepsilon_{min} = 0.02$, $\varepsilon_{max} = 0.3$ and $\Delta = 0.001$.

E. Used Performance Metrics

All of the clustering methods and both dimension reduction methods that used in this paper have been aimed at maximizing the silhouette score. A couple of performance metrics, which are explained in the next sub-sections, have been used to compare these clustering and dimension reduction methods. Some of these performance metrics (i.e., Rand Index Score, Adjusted Rand Index Score, Mutual Information Score, Adjusted Mutual Information Score) requires ground truth labels which the labels obtained from field specialists. And the rest (i.e., Silhouette Score, Calinski Harabasz Score, Davies Bouldin Score) are internal performance metrics of the clustering model, which does not require ground truth labels.

1) Silhouette Score: With simple terms, silhouette score is a measurement that compare the point distance to neighbour cluster's points. Rousseeuw, Peter explains as: "Each cluster is represented by so-called silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster and which ones are merely somewhere in between clusters." [13]. The result of the nth iteration of the clustering model that achieves the maximum silhouette score has been recorded.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, if|C_I| > 1 \qquad (1)$$

where $|C_I|$ is the number of points belonging to cluster i.

2) Calinski-Harabasz Score / Index: This internal metric was introduced by Calinski and Harabasz in 1974. The C-H index is an evaluation method based on the degree of dispersion between clusters and within-cluster dispersion. A higher score on the index means better clustering dispersion. The C-H index for K number of clusters on a dataset D,

$$C-H=[\frac{\sum_{k=1^K} n_k||c_k-c||^2}{K-1}]/[\frac{\sum_{k=1^K}\sum_{i=1} n_k||d_i-c_k||^2}{N-K}] \tag{2}$$

where, n_k and c_k are the number of points and centroid of the k^{th} cluster respectively, c is the global centroid, N is the total number of data points.

3) Davies Bouldin Score / Index: Another internal metric that we have used in our evaluation method is the Davies Bouldin Score. The scoring algorithm is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters that are farther apart than the others and less dispersed will result in a better score. On the D-B score, Lower scoring shows a higher quality of clustering. Similarity is defined as a measure R_{ij} that trades off:

- s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.
- d_{ij}, the distance between cluster centroids i and j.

$$R_{ij}=\frac{s_i+s_j}{d_{ij}} \tag{3}$$

Then the Davies-Bouldin index is defined as:

$$D-B=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j} R_{ij} \tag{4}$$

where k is number of cluster.

4) (Adjusted) Rand Index Score: The Rand index score determines the degree of similarity between calculated and ground truth cluster labels. Since our dataset does not contain the true label as a column, we can not compare the true clusters and predicted clusters. Therefore, we have taken SPCA cluster labels as true labels and compared them with NMF cluster labels. Adjusted Rand Index (ARI) improved algorithm from Rand Index. It's proposed to avoid higher cluster numbers that cause a higher index score trap. Higher scoring means identical cluster sequences in the compared arrays for the adjusted rand index score. If C is a ground truth class assignment and K the clustering, let us define a , b as:

- a, the number of pairs of elements that are in the same set in C and in the same set in K
- b, the number of pairs of elements that are in different sets in C and in different sets in K

The unadjusted Rand index is then:

$$RI=\frac{a+b}{C_2^{n_{samples}}} \tag{5}$$

where $C_2^{n_{samples}}$ is the total number of possible pairs in the dataset.

$$ARI=\frac{RI-E_{xpected}[RI]}{max(RI)-E_{xpected}[RI]} \tag{6}$$

5) (Adjusted) Mutual Information Score: Mutual information score is a measurement of the similarity between two different labels of the same data. Even though this performance metric requires ground-truth labels, it is also symmetric, which means that switching predicted labels and true labels results in the same score. This feature allows us to use labels as a second label array that is gathered with SPCA, instead of ground truth labels. Like adjusted rand index scoring, an adjusted version of the mutual information score (AMI) helps us avoid the same trap caused by a higher number of clustering labels. Higher scoring means a higher chance of mutual information within same-labeled clusters. For two clusterings U and V, the mutual information is given as:

$$MI(U,V)=\sum_{i=1}^{|U|}\sum_{j=1}^{|V|}P(i,j)log\frac{P(i,j)}{P(i)P'(j)} \tag{7}$$

$$AMI(U,V)=\frac{[MI(U,V)-E_{xpected}(MI(U,V))]}{[avg(H(U),H(V))-E_{xpected}(MI(U,V))]} \tag{8}$$

III. Results

Within the maximum silhouette scores of different clustering models, different cluster numbers have been recorded for NMF and SPCA dimension reduction methods. Figure 3 shows the maximum silhouette scores and cluster numbers obtained.

The complexity of the experimental setup increases the complexity of the result sets. We represent all the result scores related to our experiments in Table I. The data owner does not have any approved clustering labels except salesman inquiries such as 'most liked, easiest to sell, most chosen...' to bring to light the hotel groups (i.e., clusters) we had to take one dimension reduction method as true labels for rand index, mutual information and their adjusted versions. Following paragraphs explain clustering methods versus dimension reduction methods and overall high-performing clustering decisions. The dilemma of "performance scores vs. cluster numbers" has been left for the discussions section.

K-means and HAC (Hierarchical Agglomerative Clustering) require the cluster numbers in advance before the analysis. Akyol, Mert [3] worked with a similar dataset (hotel feature set from Turkey) and the researcher used 20 clusters for their K-means model. Furthermore, we have calculated WCSS values without applying any dimension reduction method to our dataset. With help of previous work and our analysis, we took cluster values between 2 and 50. Despite of reached highest silhouette scores at 2-dimensional space, cluster numbers are quite far on ranking. Therefore, SPCA reduced space has a higher C-H score and a lower D-B score. Since we have different cluster numbers on each reduction model, we have only compared ARI and AMI scores. Since these metrics are comparisons of two different clustering label arrays, we have only one score for both dimension reduction methods that is calculated as: "NMF Cluster Arrays/Members" divided by "SPCA Cluster Arrays/Members". ARI and AMI score almost as high as the same values on HAC models but lower than the OPTICS and DBSCAN models.

As we mentioned in Chapter II-C2, HAC has four different linkage methods. According to our silhouette scores, we can

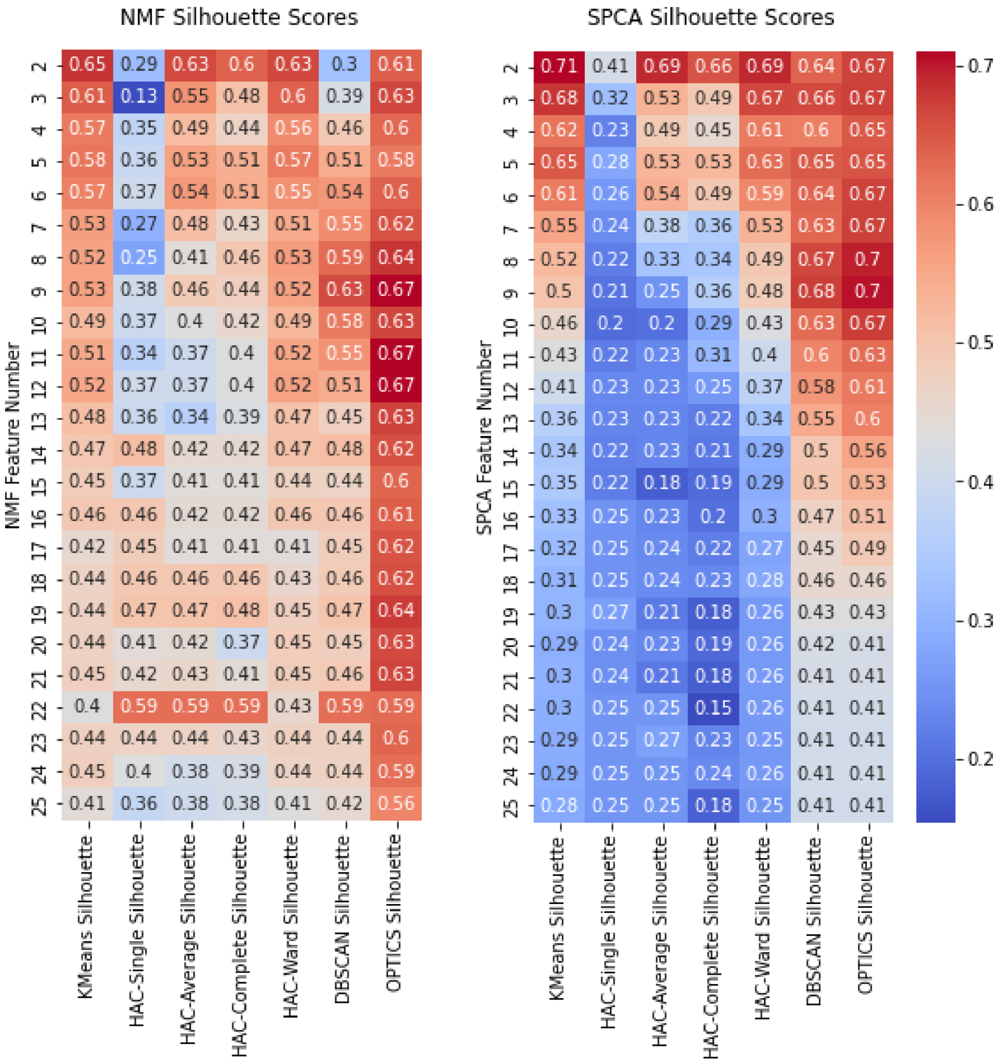


Fig. 3. NMF and SPCA Silhouette Scores

order the HAC linkage types as (weakest to strongest): single $<$ complete $<$ average = ward for NMF and SPCA reductions. C-H and D-B scores follow the same order as the given order. Single linkage has average RI and lowest AMI scores. Despite having the same cluster numbers, NMF generates ten times the dimension space as the SPCA method. Randomness and non-mutuality could have been caused by dimension difference. On the contrary, complete linked models have fewer space dimensions (only 2) at maximum silhouette scores. Almost the same ARI scores were obtained with single linkage's. Higher AMI scores have been obtained with complete linked HAC. Average and ward linked HAC models show the same patterns between NMF and SPCA methods. Overall evaluation of C-H, ARI and AMI scores; shows that ward linkage has higher clustering quality over dimension reduction methods except D-B scoring. For that exception, the difference is only 0,009.

DBSCAN provides the second highest value of silhouette score on summary table. Despite ranking second in silhouette scoring, the DBSCAN model has the highest value in RI, ARI, MI, AMI scores. This proves that 9th dimensional space and 112 different clusters have the same order and contain the same nodes. Even though the calculated epsilon values (0.047 on NMF, 0.282 on SPCA) are quite different, the noise counts are the same for both methods.

On the overall models explained thus far, the OPTICS model has the highest silhouette scores over both dimension reduction methods. We have obtained different numbers of dimension space and cluster numbers, C-H and D-B scores differ too. ARI score getting close to k-means and HAC ward/average models' ARI scores. OPTICS models have been calculated to have lower noise points (in orderly 226, 244 at maximum silhouette scores) than the DBSCAN models (400 for both methods at maximum silhouette scores).

As a final statement, both of dimension reduction methods can be applied on our dataset and OPTICS model will provide high quality clusters.

A. Further Discussions

In this study, we focused on the Toursim data set and aimed at revealing an experimental procedure for finding the best cluster structure by using binary hotel features. More specifically, we also aimed to understand how different dimension reduction methods affect the clustering process. These results have sparked off discussion about the importance of dimension reduction methods and clustering

TABLE I
SUMMARY TABLE OF NMF AND SPCA DIMENSION REDUCTION METHODS OVER CLUSTERING METHODS

Feature	**NMF**	**SPCA**
Reduced Dimension Number on K-means	2	2
K-means Cluster Number	28	43
K-means Highest Silhouette Score	0,65	0,71
K-means Calinski Harabasz Score	8111,64	15980,60
K-means Davies Bouldin Score	0,52	0,47
K-means Rand Index Score	0,97	-
K-means Adjusted Rand Index Score	0,57	-
K-means Mutual Information Score	2,71	-
K-means Adjusted Mutual Information Score	0,77	-
Reduced Dimension Number on HAC Single	22	2
HAC Single Cluster Number	2	2
HAC Single Highest Silhouette Score	0,59	0,41
HAC Single Calinski Harabasz Score	212,42	1734,61
HAC Single Davies Bouldin Score	0,71	1,07
HAC Single Rand Index Score	0,50	-
HAC Single Adjusted Rand Index Score	0,00	-
HAC Single Mutual Information Score	0,00	-
HAC Single Adjusted Mutual Information Score	0,01	-
Reduced Dimension Number on HAC Average	2	2
HAC Average Cluster Number	29	38
HAC Average Highest Silhouette Score	0,63	0,69
HAC Average Calinski Harabasz Score	7062,89	12576,80
HAC Average Davies Bouldin Score	0,51	0,44
HAC Average Rand Index Score	0,96	-
HAC Average Adjusted Rand Index Score	0,51	-
HAC Average Mutual Information Score	2,52	-
HAC Average Adjusted Mutual Information Score	0,75	-
Reduced Dimension Number on HAC Complete	2	2
HAC Complete Cluster Number	27	48
HAC Complete Highest Silhouette Score	0,60	0,66
HAC Complete Calinski Harabasz Score	6168,66	13795,40
HAC Complete Davies Bouldin Score	0,57	0,53
HAC Complete Rand Index Score	0,96	-
HAC Complete Adjusted Rand Index Score	0,51	-
HAC Complete Mutual Information Score	2,57	-
HAC Complete Adjusted Mutual Information Score	0,75	-
Reduced Dimension Number on HAC Ward	2	2
HAC Ward Cluster Number	29	42
HAC Ward Highest Silhouette Score	0,63	0,69
HAC Ward Calinski Harabasz Score	7670,26	15087,40
HAC Ward Davies Bouldin Score	0,53	0,45
HAC Ward Rand Index Score	0,97	-
HAC Ward Adjusted Rand Index Score	0,57	-
HAC Ward Mutual Information Score	2,74	-
HAC Ward Adjusted Mutual Information Score	0,79	-
Reduced Dimension Number on DBSCAN	9	9
DBSCAN Cluster Number	112	112
DBSCAN Highest Silhouette Score	0,63	0,68
DBSCAN Calinski Harabasz Score	105,79	106,13
DBSCAN Davies Bouldin Score	1,03	1,04
DBSCAN Rand Index Score	1,00	-
DBSCAN Adjusted Rand Index Score	1,00	-
DBSCAN Mutual Information Score	4,06	-
DBSCAN Adjusted Mutual Information Score	1,00	-
Reduced Dimension Number on OPTICS	12	9
OPTICS Cluster Number	182	178
OPTICS Highest Silhouette Score	0,67	0,70
OPTICS Calinski Harabasz Score	76,27	93,03
OPTICS Davies Bouldin Score	1,17	1,12
OPTICS Rand Index Score	0,98	-
OPTICS Adjusted Rand Index Score	0,54	-
OPTICS Mutual Information Score	4,12	-
OPTICS Adjusted Mutual Information Score	0,79	-

algorithms, which lie behind hotel recommendation system designs.

We have applied two dimension reduction techniques; SPCA and NMF and four clustering algorithms; K-means, HAC (with different linkages), DBSCAN and OPTICS. The results reveal that OPTICS can be the best clustering algorithm for this data set especially when the number of features are reduced to between 8 and 12. Moreover, clustering performance metrics of the DBSCAN model prove that; we can have the exact same clusters with the same dimensional space that are reduced by different dimension reduction methods.

An important future step in this work could be to focus on data noise. DBSCAN and OPTICS result in high performance metrics but they find many noisy points. Understanding the dynamics of such noises can give us an idea about the ”unique” hotels, which can be on the other side of the medallion for marketing strategies.

The dilemma of ”higher cluster numbers cause higher silhouette scores” may be another discussion topic about this paper. Silhouette scores may not be the only evaluation score of high-quality clustering process. As we examined in this paper, other performance metrics might be chosen in a bunch of iteration steps to understand hotel clustering.

Our dataset covers more than 61% of registered touristic facilities. Thus, this clustering results could provide an overview to Turkey’s hotels. With these insights, tourism recommendation systems may improve and increase the satisfaction of customers and hotel owners in the further steps.

REFERENCES

[1] W. T. Organization. (accessed: 28.11.2022) World tourism organization. [Online]. Available: https://www.unwto.org/news/international-tourism-back-to-60-of-pre-pandemic-levels-in-january-july-2022

[2] M. D. M.-P. S. Sánchez-Pérez, Manuel; Illescas-Manzano, “You’re the only one, or simply the best. hotels differentiation, competition, agglomeration, and pricing,” 02 2020.

[3] M. Akyol, “Clustering hotels and analyzing the importance of their features by machine learning techniques,” *Journal of Computer Science and Technologies - Mersin Üniversitesi*, pp. 16–23, 2021.

[4] F. G.-L. M. Rodríguez-Victoria, O.E.; Puig, “Clustering, innovation and hotel competitiveness: evidence from the colombia destination,” *International Journal of Contemporary Hospitality Management*, pp. 2785–2806, 11 2017.

[5] K. O, DAĞ; M., “Resort otellerin kümeleme analizi ile incelenmesi: Antalya ili Örneği,” *Journal of Suleyman Demirel University Institute of Social Sciences*, pp. 200–232, 04 2020.

[6] T. S. A. Birligi. (2023, Jan.) Turistik tesis ve İşletmeler. [Online]. Available: http://www.tursab.org.tr/istatistikler/turistik-tesis-isletmeler

[7] M. Ye, P. Zhang, and L. Nie, “Clustering sparse binary data with hierarchical bayesian bernoulli mixture model,” *Computational Statistics Data Analysis*, vol. 123, pp. 32–49, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016794731830029X

[8] I. M. Johnstone and A. Y. Lu, “On consistency and sparsity for principal components analysis in high dimensions,” *J. Am. Stat. Assoc.*, vol. 104, no. 486, pp. 682–693, Jun. 2009.

[9] I. S. Dhillon and S. Sra, “Generalized nonnegative matrix approximations with bregman divergences,” in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, ser. NIPS’05. Cambridge, MA, USA: MIT Press, 2005, p. 283–290.

[10] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016.

[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure,” *SIGMOD Rec.*, vol. 28, no. 2, p. 49–60, jun 1999. [Online]. Available: https://doi.org/10.1145/304181.304187

[12] J. Soler, F. Tencé, L. Gaubert, and C. Buche, “Data clustering and similarity,” in *The Twenty-Sixth International FLAIRS Conference*. Citeseer, 2013.

[13] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.