JOURNAL OF MODERN SCIENCE
NUMER SPECJALNY
TOM 5/54/2023
www.jomswsge.com

DOI: doi.org/10.13166/jms/176171

JAN FRANCISZEK LASKOWSKI
Lublin University of Technology, Poland
ORCID iD: 0000-0002-4951-8674

PAWEŁ TOMIŁO
Lublin University of Technology, Poland
ORCID iD: 0000-0003-4461-3194

A NEW AI-BASED METHOD FOR CLUSTERING SURVEY RESPONSES

Abstract

Aim: Many research projects, particularly in social science research, depend on clustering survey responses. When analyzing survey data, traditional clustering algorithms have several drawbacks. The ability to analyze survey data more effectively has been made possible by recent developments in artificial intelligence (AI) and machine learning (ML). The aim of this article is to present a new, AI-based method of clustering survey responses using a Variational Autoencoder (VAE).

Materials and methods: To determine the effectiveness of grouping, the new VAE clustering method was compared with K-means, PCA and k-means, and Agglomerative Hierarchical Clustering methods by applying the Silhouette score, the Calinski-Harabasz score, and the Davies-Bouldin score metrics.

Results: In the case of the Silhouette Score, the developed VAE method obtained a 69% higher average score than the others. For the Calinski-Harabasz Score and the Davies-Bouldin Score, respectively, the VAE method outperformed the other methods by 164% and 111%, respectively.

Conclusions: The VAE method allowed for the most effective grouping of responses given by respondents. It has made it possible to capture complex relationships and patterns in the data. In addition, the method is suitable for analyzing different types of survey data (continuous, categorical, and mixed data) and is resistant to noise.

Keywords: *Survey data analysis, clustering, artificial intelligence, variational autoencoder (VAE), machine learning, pattern discovery, exploratory data analysis*

Introduction

Research methodology plays an important role in research processes by shaping their formal basis and translating the theoretical assumptions made into the language of empirical procedures. This is especially true of surveys, which originate from the group of social methods and are widely used in the organization and management sciences allowing the identification of the designated opinions of people (respondents) in relation to certain socio-economic phenomena. The survey research method is categorized as an empirical method and focuses on solving the research problem from the experience side by capturing conditions as close to reality as possible. By its nature, it is part of the nomothetic research approach, focused on the search for generalized judgments, laws and rules of the organizational world, which is carried out through an inductive research path, allowing the truth of a phenomenon to be established on the basis of sentences that confirm its existence in some cases only. Thanks to their relative simplicity, speed and low cost of implementation, surveys have been a key data collection tool in the social sciences for many years. Analysis of survey data is also a common component of organizational research, which enables researchers to analyze patterns and trends, facilitate choices and create plans to improve organizational performance.

The data collected from surveys is usually analyzed using traditional statistical methods such as descriptive statistics, inferential statistics, regression analysis, factor analysis, cluster analysis, conjoint analysis, and discriminant analysis. However, when it comes to evaluating survey data, these conventional techniques have some drawbacks. They presume that the data follows a particular distribution, such as a normal distribution, which is one of their key constraints. For survey data, which can have complicated and non-linear correlations, this assumption might not always be valid. These methods also do not account for the great dimensionality and heterogeneity of survey data, which can lead to inaccurate and biased results.

To overcome these limitations, recent advances in artificial intelligence (AI) and machine learning (ML) have opened up new opportunities for analyzing survey data. In particular, deep learning techniques such as variational autoencoder (VAE) have shown promise in clustering survey responses. Variational Autoencoder (VAE) is a deep learning generative model that encodes input data into a lower-dimensional latent space and then decodes it back to the original high-dimensional space in order to learn a compact representation of the data. Currently, VAE-based data analysis methods are being successfully applied in various fields of science and technology, such as image and video analysis, natural language processing, anomaly detection, drug discovery and recommendation systems.

The purpose of this article is to present and evaluate the effectiveness of a new method for grouping survey responses using Variational Autoencoder (VAE).

In order to achieve such a research objective, independent analyses were made of the results of a survey on the value system of employees in the 50+ generation using VAE and three other popular data grouping methods, namely K-means, PCA and k-means and Agglomerative Hierarchical Clustering. To determine the effectiveness of grouping, the methods presented above were compared by applying metrics such as the Silhouette index, the Calinski-Harabasz index, and the Davies-Bouldin index.

Survey research is a prominent methodology in the social sciences, notably in organizational research. The primary goal of survey research is to collect data from a sample of respondents in order to understand more about their attitudes, habits, and opinions on a specific topic. Data must be evaluated after it has been collected in order to derive valuable findings (Fowler, 2013, p. 134-140). Some of the most common methods for analyzing survey data include descriptive statistics (Holcomb, 2016, p. 1-98), inferential statistics (Asadoorian, 2005, p. 2-28), factor analysis (Tucker, 1951, p. 1-35), regression analysis (Kleinbaum, 2013, p. 34-704) and cluster analysis (Punj, 1983, p. 134-148).

There are several methods for clustering survey response data, including:

- K-Means Clustering: This approach of grouping data is well-liked and straightforward. Based on the average distance of the data points from the centroid of each cluster, it divides the data into K clusters (Bock, 2007, p. 5–28; Likas, 2003, p. 1-27).
- Hierarchical Clustering: This approach creates a hierarchy of clusters by first treating each data point as its own cluster, then grouping similar clusters into larger clusters until every data point is a member of a single cluster (Day, 1984, p. 7-24; Murtagh, 2012, p. 86-97).
- Density-Based Clustering: This technique is used to locate data clusters with a high point density but a potentially ill-defined boundary. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most popular density-based clustering algorithm (Campello, 2013, p. 160-172; Kriegel, 2011, p. 231-240).
- Model-Based Clustering: The distribution of data in each cluster is described using statistical models in this manner. Latent class analysis and Gaussian mixture models are two popular model-based clustering techniques (Fraley, 1998, p. 578-587; Kriegel, 2011, p. 231-240).

- Affinity Propagation: The foundation of this approach is the idea of "message transmission" between data points. By comparing the similarity of different data points, it discovers clusters (Wang, 2007, p. 1242-1246).
- Spectral Clustering: This technique divides the data into clusters using the eigenvalues and eigenvectors of a similarity matrix. It is frequently applied to non-linear clustering issues (Ng, 2001, 14-19).
- Multiple measures can be used to assess a clustering method's performance. Several of the frequently used metrics include:
- Silhouette score: By taking into account both intra-cluster and inter-cluster similarity, this evaluates the caliber of a clustering solution. A high silhouette score means that the data points have been successfully divided into discrete clusters using the clustering process (Shahapure, 2020, p. 124-131; Shutaywi, 2021, p. 759).
- Calinski-Harabasz score: This measures a ratio of the sum of between-cluster dispersion and within-cluster dispersion (Lima, 2020, p. 97-106).
- Davies-Bouldin score: This measures the average similarity of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances (Arturo, 2018, p. 1-8; Petrovic, 2006, p. 1-12).

Materials and methods

Data characteristic

The CAVI questionnaire was the primary research tool used to collect data. The survey was anonymous, with a sample size of 600 people (377 women and 223 men). The survey was designed to investigate the value systems of women and men representing the "silver" generation of employees and it included respondents aged 50 and up who are professionally active (Laskowska, 2022, p. 194-224). More than half (52%) of those polled had a secondary education. The majority of respondents (23%) lived in large cities and worked in commerce (27%) as well as industry and construction (15%). The attempt was deliberate. The research was carried out in the first quarter of 2022.

To investigate the value system that members of the "silver" generation live by, respondents were asked to rate characteristics chosen based on the assumptions of Shalom H. Schwartz's theory of basic human values (Schwartz, 2012, p. 663-688). The questionnaire contained 16 questions with a semantic differential scale based on Charles E. Osgood's theory of semantic differences (Osgood, 1964, 171-200; Themistocleous, 2019, p. 394-407). The scales used have values ranging from 1 to 10, with 1 being the least significant and 10 being the most significant. The intervals between successive scale values were designed to be equal, resulting in interval scales. The internal consistency of the survey questionnaire was examined using Cronbach's Alpha (α) and McDonald's omega (ω) test (α = 0.72-0.91 and ω = 0.81-0.90).

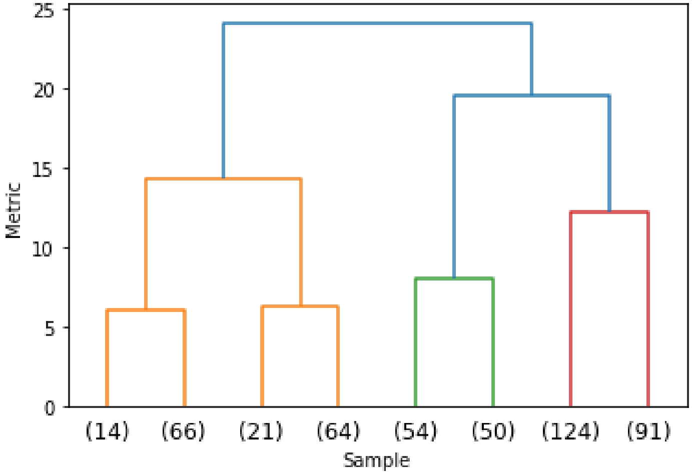
Data clustering methods

A neural network model was created expressly for the task of clustering survey data responses. The Variational Autoencoder (VAE) model serves as the foundation for the neural network's structure. To compare the results from various clustering strategies, three distinct methodologies will be used:

- K-means;
- PCA and k-means;
- Agglomerative Hierarchical Clustering;

Due to the algorithms used, the total number of groups was set in advance. The initial number of clusters for our data has been set by dendrogram. Figure 1 indicates that three clusters should be obtained from our data. This number of clusters has poor variance between classes because it has been split into all most positive, mean, and all most negative sets. The deviation from the whole attitude mean value for each group is shown in Table 1. So, the next proposed number of clusters with lower dissimilarity was 4. A total of four clusters produced satisfactory results.

Fig. 1. *Top part of dendrogram*



K-means

K-means is an algorithm that allows us to separate samples into groups of equal variances by minimizing the known criterion (equation 1). This algorithm requires a predetermined number of k clusters. The set of N samples X is divided into K clusters C, where each cluster is described by the mean u_j of the samples in this cluster (Arthur, 2007, p. 1-9).

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2 \tag{1}$$

PCA and k-means

The K-means algorithm for multidimensional spaces suffers from the so-called "course of dimensionality," because Euclidean distances tend to become inflated. Therefore, it is a good idea to use a dimensionality reduction algorithm to mitigate the problem and speed up calculations. In order to reduce the dimensions, the Principal Component Analysis (PCA) algorithm was chosen. PCA is a technique that reduces the dimensionality of data sets. The use of this technique increases the interpretability of data and minimize the loss of information. PCA creates new variables that are uncorrelated and maximizes variance. This method is an adaptive data analysis technique, due to the fact that the search for principal components is based on the available dataset, and the solution is the eigenvalue problem (Jollife, 2016, p. 374-382).

Agglomerative Hierarchical Clustering

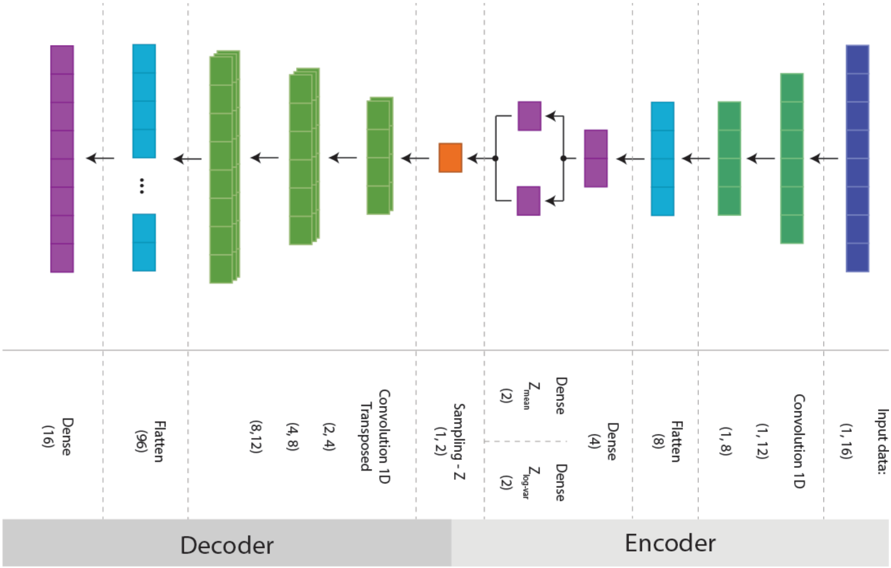
Agglomerative Hierarchical Clustering (AHC) is a family of different methods that are related to each other at the computational level. All of the methods in this family establish structured relationships between data rather than assuming a priori data structure. AHC creates hierarchically ordered clusters that represent the proximity structure of the analyzed data. The data is not presented as a spatial cluster, but as a dendrogram or constituency tree (Campello, 2013, p. 160-172). Agglomerative clustering performs a hierarchical clustering operation using a bottom-up approach. Each observation is initially treated as a separate cluster, and then the observations are combined with each other by the linkage criteria.

Variational Autoencoder (VAE)

Variational Autoencoder was used to reduce the dimension of the input data and map it to 2D space. VAE is an Artificial Neural Network (ANN) architecture. This architecture belongs to the generative modeling field in machine learning. The main goal of this method is to capture dependencies between each input vector and maximize the probability for each X in the dataset so the model could generate data very similar to the input data. Let X be out datapoints in some high-dimens`ional space $\mathcal{X}$ and let $P(X)$ be distribution that is defined over X, let Z be a latent variable vector in a high-dimensional space $\mathcal{Z}$, and assume that it is possible to sample in accordance with the probability density function that is defined over $\mathcal{Z}$. $\mu(Z;\theta)$ is a deterministic function family, and it is parametrized by a vector θ in space Θ, where f: $Z \times \Theta \rightarrow \mathcal{X}$, if we assume that Z is random, and θ is fixed variable, then $f(Z;\theta)$ is random variable in space $\mathcal{X}$. Finally, the function (2) that is maximized in the training process is (Doersch, 2016, p. 1-21):

$$P(X) = \int P(X|Z;\theta)P(Z)dz \tag{2}$$

Fig. 2. *Structure of VAE*



A Variational Autoencoder is made up of two basic components: encoder and decoder. The role of the encoder is to reduce the dimensionality of the input data; the decoder needs to recreate the input data from the output of the encoder so that the loss function is minimized. In our case, VAE is probabilistic, which means that we use latent distributions to sample latent space points. In the encoder part, the input data is fed to two convolution layers; later, it is flattened and encoded as a distribution over the latent space. 2D point coordinates in latent space are sampled from the latent distribution. Encoded distributions are specified to be normal, which allows our encoder to return the mean and the covariance matrix. This helps in regularizing the latent space so that the returned distributions are close to the standard normal distribution. The output of the encoder part is input to the decoder. The role of the decoder is to reconstruct the input data from latent spatial coordinates. The data is fed to three transposed convolution layers, then flattened to size 96 and fed to a dense layer with 16 neurons. The structure of our VAE is shown in figure 2.

In the described model, the sampled values are Z_{mean} – the mean value, and $Z_{log\text{-}var}$ – the log variance. Both values are respectively represented by equations (3) and (4).

$$Z_{mean} = \mu_e = f_{w_\mu}(x) \tag{3}$$

$$Z_{log-var} = \log(\sigma_e^2) = f_{w_\sigma}(x) \tag{4}$$

where:

μ_e – mean (encoder);
σ_e – standard deviation (encoder);
fw_μ – neural network with weights w_μ;
fw_σ – neural network with weights w_σ;
$\mathcal{X}$ – observations from the previous layer.

Let $q(z|x,\ W)$ be a function that is used to approximate the true posteriori and that, based on variable x produces a distribution over the latent variable z. The parameters W correspond to the distribution q.

The reparameterization trick described by (Kingma, 2019, p. 307-392) was used for sampling, which consists in introducing a random variable ϵ with a known distribution $p(\epsilon)$. Sample is obtained from this distribution, and then let be a deterministic, differentiable function – equation (5).

$$Z = g\ (x,\ \epsilon,\ W) \tag{5}$$

By applying this trick, we can use the Monte Carlo method to estimate the expectation and differentiation of the equation (6).

$$\nabla_{\theta,W}\mathcal{L}(q,\theta,W) \approx \nabla_{\theta,W}\frac{1}{L}\sum_{i=1}^{L}\log p\left(x\middle|Z^{(i)},\theta\right) \tag{6}$$

where:

L – number of samples;
θ – generative parameters (decoder).

We use the sum of the mean square error as the loss function for reconstruction and the Kullback-Leibler divergence for distribution – equation (7) (Kingma, 2019, p. 307-392):

$$\mathcal{L}_{kl} = \frac{1}{2}\sum_{j=1}^{J}\left(-1 - \log\left(\sigma_{Z,j}^{(l)}\right)^2 + \left(\mu_{Z,j}^{(l)}\right)^2 + \left(\sigma_{Z,j}^{(l)}\right)^2\right) \tag{7}$$

The second step is to cluster the data. For two-point clustering, we have used Ward's agglomerative clustering algorithm. This method is calculating the Euclidean distance between all the points as a way to find a pair with the smallest possible dissimilarity. In Ward's method, the initial cluster distance is the squared Euclidean distance between points, so our metric is defined as equation (8) (Ward, 1963, p. 236-244):

$$d_{ij} = d\left(\{X_i\},\{X_j\}\right) = \left\|X_i - X_j\right\|^2 \tag{8}$$

Methods comparison

To determine the effectiveness of grouping, the methods presented above were compared by applying the following metrics:

- Silhouette Score;
- Calinski-Harabasz Score;
- Davies-Bouldin Score.

The Silhouette Score is determined by equation (9) (Rousseeuw, 1987, p. 53-65):

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{9}$$

where:

i – sample;

a – i – sample's average distance from every other point in its class;

b – average distance between sample and every other point in the follow ing cluster.

This score is used to evaluate how effective a clustering method is. The value varies from – 1 to 1, where 1 indicates clusters that are clearly distinct and spaced widely apart. Zero means that clusters are overlapping.

The Calinski-Harabasz Score (Varian2ce Ratio Criterion) is defined by equation (10) (Caliński, 1974, p. 1-27):

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{10}$$

where:
tr(W$_k$) – trace of the within cluster dispersion matrix;
W_k – defined by equation (11);
tr(B$_k$) – trace of between group dispersion matrix;
B_k – defined by equation (12);
k – number of clusters;
n_E – size of the data.

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{11}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{12}$$

where:
C_q – set of points in cluster q;
c_e – center of cluster q;
n_q – amount of data in cluster q;
c_E – center of whole data.

Clustered data with more clearly defined clusters correlates with a higher Calinski-Harabasz score. Lower values of the Davies-Bouldin Score indicate that data is better separated between clusters. This method represents the average of a metric that contrasts the size of the clusters with the distance between clusters. The discussed method was presented using the equation (13) (Davies, 1979, p. 224-227):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{dij} \tag{13}$$

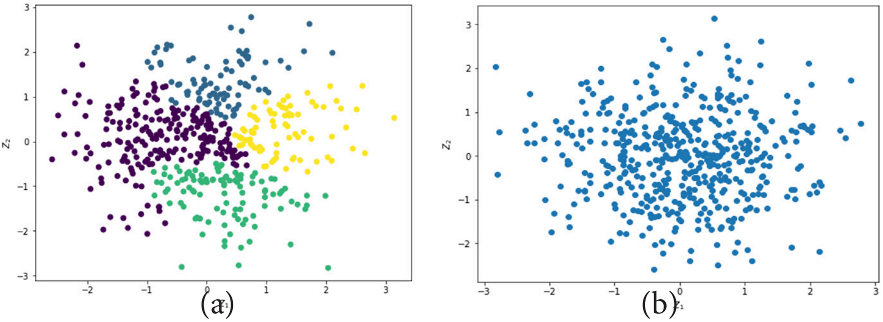
where:

k – number of clusters;
i – cluster;
j – most similar cluster to $_i$;
s – cluster diameter;
dij – the distance between $_i$ and $_j$ (cluster centroids).

Results and discussion

In order to acquire a 2D representation of our data, we used only the encoder's output. To visualize variables from latent space: z_1 and z_2 were used. The mapped data in 2D space is show in figure 3a. The next step is to cluster the 2D data. In order to do this, the hierarchical agglomerative clustering algorithm with minimum variance clustering was used. The criterion of the merge in this method is a function of all individual distances from the centroid (Manning, 2009, p. 235). The results of this algorithm for our 3D representations are shown in Figure 3b. The deviation from the attitude mean value for each group is shown in Table 3.

Fig. 3. *Survey data mapped to 2D space (a) and clustered data*



(a)

(b)

Table 3. *Deviation from the attitude mean value for each group*

Attitude	**Cluster**			
	1	**2**	**3**	**4**
Happiness	8,2701	8,9444	8,1548	7,7231
Family	8,9368	9,4028	9,0595	8,9231
Love	8,4483	8,5694	8,2143	8,1077
Work and career	7,3276	7,1111	7,1905	7,0615
Prosperity and wealth	6,2759	6,5	6,4643	6,1692
Friendship	7,6724	7,4167	7,369	7,6154
Honesty	8,7644	8,9028	8,7619	9
Knowledge	8,2069	8,1389	8,0119	8,0769
Personal development	7,977	7,8889	7,6786	7,5692
Security	8,8793	9,2917	8,8214	9,0769
Complacency	8,6379	8,9444	8,3929	8,5538
Stabilization	8,5747	8,8333	8,4048	8,6154
Passion and hobby	7,4483	7,4722	7,5952	7,3231
Health	8,8161	9,1944	8,8571	9,1231
Professional status	6,6322	6,2778	6,6786	6,0615
Admiration and respect	5,7414	5,7361	5,75	5,6615
Group size	**125**	**50**	**72**	**43**

As we can see, at this stage, none of the groups formed have visible features that could provide a logical link between them. Therefore, for further analysis, the data need to be filtered. For this purpose, the significance interval has been calculated by equation (14). Table 4 shows deviation from the attitude mean value for each group after filtering.

$$F_r\left(x_{m_r}\right) = \frac{\max\left(x_{m_r}\right) + \min\left(x_{m_r}\right)}{N_c}\beta \qquad (14)$$

where:

X_{m_r} – set of values from -th row;

N_c – number of clusters;

β – size coefficient.

Table 4. *Deviation from the attitude mean value for each group with filter – VAE*

Attitude	Cluster			
	Self-enhancement	Social focus	Personal focus	Nihilists
Happiness	0,00	0,67	-0,12	-0,55
Family	-0,14	0,32	-0,02	-0,16
Love	0,11	0,23	-0,12	-0,23
Work and career	0,15	-0,06	0,02	-0,11
Prosperity and wealth	-0,08	0,15	0,11	-0,18
Friendship	0,15	-0,10	-0,15	0,10
Honesty	-0,09	0,05	-0,10	0,14
Knowledge	0,10	0,03	-0,10	-0,03
Personal development	0,20	0,11	-0,10	-0,21
Security	-0,14	0,27	-0,20	0,06
Complacency	0,01	0,31	-0,24	-0,08
Stabilization	-0,03	0,23	-0,20	0,01
Passion and hobby	-0,01	0,01	0,14	-0,14
Health	-0,18	0,20	-0,14	0,13
Professional status	0,22	-0,13	0,27	-0,35
Admiration and respect	0,02	0,01	0,03	-0,06
Group size	**215**	**85**	**104**	**80**

The use of the Variational Auto Encoder allowed the data to be grouped into 4 clusters, for which, by determining the average value for each answer and applying the formula (4), the results presented in Table 4 were obtained. As we can see, the groups of responses obtained were characterized by dominant sets of features, from which it is possible to determine the characteristic pattern (attitude) of each group.

To verify the effectiveness of the VAE method, independent analyses were made of the results of a survey on the value system of employees in the 50+ generation using three other popular data grouping methods, namely K-means, PCA and K-means and Agglomerative Hierarchical ClusteringThe results obtained by these algorithms are presented in tables 5, 6, and 7, respectively.

Table 5. *Deviation from the attitude mean value for each group with filter – K-means*

Attitude	Cluster			
	1	2	3	4
Happiness	1,13	1,97	-1,64	-1,46
Family	1,60	2,15	-3,11	-0,64
Love	1,11	2,05	-2,64	-0,53
Work and career	-0,43	1,46	0,56	-1,59
Prosperity and wealth	-0,32	0,79	1,11	-1,59
Friendship	0,22	1,97	-0,50	-1,69
Honesty	0,95	1,59	-0,56	-1,98
Knowledge	-0,07	1,50	0,22	-1,65
Personal development	-0,51	1,51	0,57	-1,56
Security	0,94	1,81	-0,55	-2,19
Complacency	0,85	1,67	-0,54	-1,98
Stabilization	0,95	1,74	-0,18	-2,51
Passion and hobby	-0,12	1,22	0,78	-1,88
Health	1,22	1,75	-0,62	-2,35
Professional status	-0,81	1,06	1,49	-1,74
Admiration and respect	-1,39	0,89	1,70	-1,21
Group size	**151**	**260**	**27**	**46**

Table. 6. *Deviation from the attitude mean value for each group with filter – K-means + PCA*

Attitude	**Cluster**			
	1	**2**	**3**	**4**
Happiness	1,13	1,97	-1,77	-1,34
Family	1,63	2,20	-3,44	-0,39
Love	1,26	2,22	-3,13	-0,34
Work and career	-0,38	1,46	0,58	-1,66
Prosperity and wealth	-0,39	0,70	1,40	-1,71
Friendship	0,35	2,05	-0,79	-1,61
Honesty	0,88	1,56	-0,56	-1,88
Knowledge	-0,07	1,49	0,19	-1,62
Personal development	-0,52	1,47	0,64	-1,59
Security	0,97	1,84	-0,78	-2,03
Complacency	0,91	1,70	-0,64	-1,98
Stabilization	1,00	1,76	-0,37	-2,39
Passion and hobby	-0,13	1,18	0,80	-1,84
Health	1,14	1,70	-0,59	-2,25
Professional status	-0,84	1,01	1,53	-1,70
Admiration and respect	-1,34	0,89	1,66	-1,22
Group size	**152**	**261**	**22**	**49**

Table.7. Deviation *from the attitude mean value for each group with filter – Agglomerative* clustering algorithm

Attitude	Cluster			
	1	2	3	4
Happiness	-2,84	-0,11	1,20	1,74
Family	-4,34	0,76	1,61	1,96
Love	-3,08	0,10	1,22	1,76
Work and career	-0,93	-0,17	-0,40	1,51
Prosperity and wealth	-0,22	-0,52	-0,06	0,79
Friendship	-1,42	-0,31	0,03	1,71
Honesty	-2,24	-0,22	0,85	1,62
Knowledge	-1,36	-0,16	-0,05	1,57
Personal development	-0,59	-0,51	-0,43	1,52
Security	-1,94	-0,82	1,05	1,71
Complacency	-1,61	-1,31	1,21	1,70
Stabilization	-2,06	-1,05	1,33	1,78
Passion and hobby	-0,77	-0,23	-0,19	1,19
Health	-2,84	-0,23	1,33	1,74
Professional status	-0,09	-0,45	-0,54	1,08
Admiration and respect	0,44	-0,17	-1,50	1,24
Group size	**30**	**79**	**138**	**237**

As we can see, only the use of the VAE model made it possible to group the answers from the survey into 4 clusters, which, compared to other methods, are characterized by features that can be distinguished. Additionally, the groups obtained in this way have a more even quantitative distribution than the other algorithms discussed, which also proved the superiority of the VAE method. The comparison of clusters' group sizes is presented in Table 8.

Table. 8. *Clusters' group size*

Clustering method	Cluster			
	1	2	3	4
VAE	215	104	85	80
K-means	260	151	46	27
K-means + PCA	261	152	49	22
Agglomerative clustering	237	138	79	30

To determine the effectiveness of grouping, the methods presented above were compared by applying metrics such as the Silhouette index, the Calinski-Harabasz index, and the Davies-Bouldin index. The presented VAE method employing agglomerative clustering outperforms the previously presented methods for comparing clusters for data after dimension reduction. In the case of the Silhouette Score, the developed method obtained a 69% higher average score than the others. For the Calinski-Harabasz Score and the Davies-Bouldin Score, respectively, the VAE method outperformed the other methods by 164% and 111%, respectively. The discussed metrics are presented in Table 9.

Table 9. *Clustering method comparison*

	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
VAE	0.2787	232.6745	0.99456
K-means	0.1764	92.6486	2.0047
K-means + PCA	0.1743	92.1125	2.0074
Agglomerative clustering	0.1472	80.1873	2.3090

Conclusion

The research has confirmed that the proposed AI-based VAE clustering method provided the most effective grouping of respondents' responses. Complex relationships, trends, and patterns in the data could be captured using the VAE method, which was not achievable using the other grouping techniques. The study also demonstrated the flexibility and scalability of VAE as a strategy for handling a variety of survey data types, including continuous, categorical, and mixed data. Additionally, VAE can handle missing data and is robust to noise, making it a suitable method for analyzing survey data, which can often be noisy and have missing responses.

Using the VAE method to cluster survey responses has a number of advantages for organizational research. First, compared to conventional clustering approaches, the suggested method is more adept at handling high-dimensional and heterogeneous survey data. Multiple questions and variables are frequently used in surveys to measure various characteristics of the topic being studied. The suggested strategy can uncover the data's underlying structure and spot trends that conventional approaches might miss. Second, the suggested approach is capable of simulating intricate, non-linear interactions between survey responses. Complex interactions between various variables that influence organizational results are possible in organizational research. The suggested approach can record these interactions and offer a more precise and in-depth comprehension of the topic under study. Third, the outcomes produced by the suggested strategy may be easier to understand and more significant. In a lower-dimensional latent space, VAE can learn a compressed representation of the data. As a result, it may be simpler to see and understand the clustering results, which may result in wiser judgments.

However, there are also some challenges to using VAE for clustering survey data. Choosing the proper model architecture and hyperparameters, which can have a big impact on the outcomes, is one of the issues. Additionally, because the latent space representation could not be exactly related to the original data, it can be challenging to comprehend the VAE results. Therefore, additional study is required to confirm the usefulness of the suggested strategy in various scenarios and to evaluate it against other cutting-edge AI-driven methodologies.

References

Arthur, D., Vassilvitskii, S. (2007). K-means. the advantages of careful seeding. Symposium on Discrete Algorithms. Accessed 20.04.2023 at https://forge.agroparistech.fr/p/bcbgmiap/source
/tree/670/biblio/clustering/kMeansPP-soda.pdf

Arturo, A., Scuola, V., Santanna, S., Binaghi, E., Vergani, A. A. (2018). A soft davies--bouldin separation measure. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). https://doi.org/10.1109/FUZZ-IEEE.2018.8491581

Asadoorian, M., Kantarelis, D. (2005). Essentials of inferential statistics. Accessed 23.04.2023 at https://www.google.com/books?hl=pl&lr=&id=MUyrEx2ATwkC&oi=fnd&pg=PA1&dq=Inferential+statistics&ots=06pbqd0z2H&sig=wf0Cps61skrTa7e_9RIEZ0SXF2E

Bock, H. (2007). *Clustering Methods: A History of k-Means Algorithms. Selected Contributions in Data Analysis*. Accessed 12.05.2023 at https://link.springer.com/content/pdf/10.1007/978-3-540-73560-1.pdf#page=167

Caliński, T. (1974). *A dendrite method for cluster analysis*. Taylor & Francis, 1–27. https://doi.org/10.1080/03610927408827101

Campello, R. J. G. B., Moulavi, D., Sander, J. (2013). *Density-based clustering based on hierarchical density estimates*, 7819 LNAI(PART 2), 160–172. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-642-37456-2_14/COVER

Davies, D. L., Bouldin, D. W. (1979). *A Cluster Separation Measure*, *PAMI-1*(2), 224–227. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.1979.4766909

Day, W. H. E., Edelsbrunner, H. (1984). *Efficient algorithms for agglomerative hierarchical clustering methods.*, 1(1), 7–24. Journal of Classification. https://doi.org/10.1007/BF01890115

Doersch, C. (2016). *Tutorial on Variational Autoencoders*. Accessed 20.04.2023 at https://arxiv.org/abs/1606.05908v3

Fowler, F. J. (2013). *Survey research methods*. Taylor & Francis.

Fraley, C., Raftery, A. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis.* The Computer Journal. Accessed 19.04.2023 at https://academic.oup.com/comjnl/article-abstract/41/8/578/360856

Holcomb, Z. (2016). *Fundamentals of descriptive statistics*. Accessed 22.04.2023 at https://www.google.com/books?hl=pl&lr=&id=X18PDQAAQBAJ&oi=fnd&pg=PP5&dq=Descriptive+statistics&ots=UUFiPpR_V6&sig=NISwEjH7XsHNvSdf-q_dfkxTGvA

Jollife, I. T., Cadima, J. (2016). *Principal component analysis: a review and recent developments*. 374(2065). Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://doi.org/10.1098/RSTA.2015.0202

Kingma, D. P., Welling, M. (2019). *An Introduction to Variational Autoencoders*, *12*(4), 307–392. Foundations and Trends® in Machine Learning. https://doi.org/10.1561/2200000056

Kleinbaum, D., Kupper, L., Nizam, A., Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Kriegel, H. P., Kröger, P., Sander, J., Zimek, A. (2011). *Density-based clustering*, 1(3), 231–240. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. https://doi.org/10.1002/WIDM.30

Laskowska, A., Laskowski, J. F. (2022). *Silver Generation at Work – Implications for Sustainable Human Capital Management in the Industry 5.0 Era,* 15(1), 194. Sustainability. https://doi.org/10.3390/SU15010194

Likas, A., Vlassis, N., Verbeek, J. (2003). *The global k-means clustering algorithm*. Pattern Recognition. Accessed 19.04.2023 at https://www.sciencedirect.com/science/article/pii/S0031320302000602?casa_token

Lima, S., Aplicada, M. C. (2020). *A genetic algorithm using Calinski-Harabasz index for automatic clustering problem,* 12(3), 97–106. Revista Brasileira de Computação. https://doi.org/10.5335/rbca.v12i3.11117

Manning, C. (2009). *An introduction to information retrieval*. Accessed 11.04.2023 at https://ds.amu.edu.et/xmlui/bitstream/handle/123456789/14697/Book%20558%20pages.pdf?sequence=1&isAllowed=y

Murtagh, F., Contreras, P. (2012). *Algorithms for hierarchical clustering: An overview*, 2(1), 86–97. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. https://doi.org/10.1002/WIDM.53

Ng, A., Jordan, M., Weiss, Y. (2001). *On Spectral Clustering: Analysis and an algorithm*, 14. Advances in Neural Information Processing Systems.

Osgood, C. E. (1964). *Semantic Differential Technique in the Comparative Study of Cultures*, 66(3), 171-200. American Anthropologist.

Petrovic, S. (2006). *A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters*. Proceedings of the 11th Nordic Workshop of Secure. Accessed 15.04.2023 at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe 12e97cfdaefbb2fefc253b

Punj, G., Stewart, D. W. (1983). *Cluster Analysis in Marketing Research: Review and Suggestions for Application*, 20(2), 134–148. Journal of Marketing Research. https://doi.org/10.1177/002224378302000204

Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, 20(C), 53–65. Journal of Computational and Applied Mathematics. https://doi.org/10.1016/0377-0427(87)90125-7

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J. E., Demirutku, K., Dirilen-Gumus, O., Konty, M. (2012). *Refining the theory of basic individual values*, 103(4), 663-688. Journal of Personality and Social Psychology. https://doi.org/10.1037/A0029393

Shahapure, K., Nicholas, C. (2020). *Cluster quality analysis using silhouette score*. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Accessed 11.04.2023 at https://ieeexplore.ieee.org/abstract/document/9260048/

Shutaywi, M., Kachouie, N. N., Scarfone, M. (2021). *Silhouette analysis for performance evaluation in machine learning with applications to clustering,* 6(23), 759. Entropy, https://doi.org/10.3390/e23060759

Themistocleous, C., Pagiaslis, A., Smith, A., Wagner, C. (2019). *A comparison of scale attributes between interval-valued and semantic differential scales*, 61(4), 394-407. International Journal of Market Research. https://doi.org/10.1177/1470785319831227

Tucker, L. (1951). *A method for synthesis of factor analysis studies*. ETS Program Report. Accessed 21.04.2023 at https://apps.dtic.mil/sti/pdfs/AD0047524.pdf

Wang, K. J., Zhang, J. Y., Li, D., Zhang, X. N., Guo, T. (2007). *Adaptive affinity propagation clustering*. *33*(12), 1242–1246. Acta Automatica Sinica. https://doi.org/10.1360/aas-007-1242

Ward, J. H. (1963). *Hierarchical Grouping to Optimize an Objective Function,* 58(301), 236–244. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1963.10500845