Stadlthanner, Dominik; Steinkellner, Harald; Landschützer, Christian; Kaever, Domenik

Article

A hierarchical density-based clustering method applied to mixed-mail in Austria

Logistics Research

Provided in Cooperation with:
Bundesvereinigung Logistik (BVL) e.V., Bremen

Suggested Citation: Stadlthanner, Dominik; Steinkellner, Harald; Landschützer, Christian; Kaever, Domenik (2024) : A hierarchical density-based clustering method applied to mixed-mail in Austria, Logistics Research, ISSN 1865-0368, Bundesvereinigung Logistik (BVL), Bremen, Vol. 17, Iss. 1, pp. 1-17,
https://doi.org/10.23773/2024_4

This Version is available at:
https://hdl.handle.net/10419/333431

Logistics Research (2024) 17:4
DOI_10.23773/2024_4

A Hierarchical Density-Based Clustering Method Applied to Mixed-Mail in Austria

D. Stadlthanner[1], H. Steinkellner[1], C. Landschützer[1], D. Kaever (geb. Prims)[1]

Received: 03 July 2023 / Accepted: 16 April 2024 / Published online: 2 July 2024

ABSTRACT

As the *courier, express and parcel (CEP)* market has grown rapidly in recent years, shipment packaging has also shifted from classic cuboid cardboards to mixed-mail, typically with flexible plastic or paper packaging. Despite being cost-effective and space-efficient, the physical characteristics of mixed-mail items vary greatly, resulting in substantial difficulties when utilizing existing automated material handling technology in logistics distribution centers. Developing new material handling technologies that meet the requirements of mixed-mail is challenging due to the heterogeneity of the physical properties of mixed-mail, making it difficult to find suitable specimens for testing. To address this issue, this study categorizes mixed-mail based on common combinations of physical characteristics using density-based cluster analysis. The physical characteristics of >400 mixed-mail items were recorded at an Austrian distribution center. The resulting dataset is of the mixed-variable type, meaning that it features both numerical and categorical variables. To homogenize the data for clustering, different methods are available. We compared four homogenization approaches using a benchmark study featuring simulated mixed-variable datasets with varying properties. The approach based on *Gower's distance* in combination with the clustering algorithm *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) showed the best results over a wide range of different dataset properties. We then use this approach to cluster the Mixed-Mail dataset, resulting in two different clustering solutions based on different hyperparameter settings, with a total of six and eight clusters, respectively.

✉ Dominik Stadlthanner[1]
Harald Steinkellner[1]
Christian Landschützer[1]
Domenik Kaever[1]

[1] Institute of Logistics Engineering, Graz University of Technology, Austria

KEYWORDS: CEP · clustering · mixed-mail · small consignment · polybag

1. INTRODUCTION

The *courier, express and parcel (CEP)* industry is one of the fastest-growing industries worldwide and can be considered a key driver of modern consumer behavior and e-commerce business models. Between 2000 and 2021, the CEP volume in Germany increased by approximately 167%. In recent years, the COVID-19 pandemic has further accelerated this growth, leading to a 24% rise in CEP volume from 2019 to 2021 alone [1, 2]. Globally, this trend is even more pronounced, with parcel shipment numbers reaching 159 billion in 2021 from 103 billion in 2019, a remarkable 54% increase [3, 4].[1] The growth of this industry is expected to continue, albeit slightly slower than in previous years mainly due to the effects of the Russo-Ukrainian war on the industry and high inflation rates throughout Europe [2]. Fig. 1 provides an overview of the historical CEP shipment numbers in Germany as well as the forecasted numbers until 2026 [1, 2]. These enormous growth rates mean that logistics distribution centers quickly reach their capacity limits and that the planning of new distribution centers and the planning and adaptation of a CEP service provider's entire logistics network do not always meet market demands fast enough. Additionally, there has been a significant shift in market share from business-to-business (B2B) shipments towards business-to-consumer (B2C) shipments [2, p. 19], leading to the number of potential delivery addresses increasing disproportionately. This poses major challenges for distribution center operators, making fast and reliable automated sorting even more critical [5].

Furthermore, the rise of e-commerce has led to a regional change in the CEP sector, with an increase

[1] The shipment numbers in [3] and [4] are based on data for the following 13 countries: Australia, Brazil, Canada, China, France, Germany, India, Italy, Japan, Norway, Sweden, the United Kingdom and the United States.

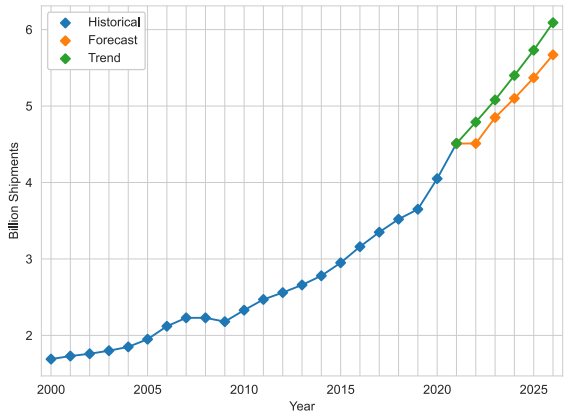


Figure 1: Historical and forecasted CEP shipment numbers in Germany until 2026 [1, p. 11, 2, p. 13].

in cross-border deliveries from China due to the low costs and lack of local availability. According to a survey conducted by the International Post Corporation in 2022, 30% of worldwide participants reported China as the source country for their most recent online crossborder purchase [6, p. 10]. Among other factors, the growing importance of Asian countries in e-commerce has led to a shift in the CEP sector's shipment spectrum, with traditional rectangular cardboards being replaced by small consignments often in the form of flexible polybags (see Fig. 2). The main benefits of polybags are low production costs and efficient use of space compared with traditional cardboard packaging.[2] Another advantage of polybags is that they can be handled more efficiently and safely as bulk materials than cardboard boxes. This advantage is particularly useful for process automation.[3] However, when it comes to the sorting process, small consignments, including polybags, are often too large to be processed by mail sorters and at the same time too small to be efficiently sorted by parcel sorters, and they have other unfavorable characteristics which make efficient handling difficult, such as the great heterogeneity of packaging materials and the associated physical problems (different friction and flexural behaviors), as well as contrast problems in the visual identification of labels. Because of this fact, small consignments are also referred to as mixed-mail [8]. Unlike traditional rectangular cardboards, there are no regulations for small consignments and the only definition to date is that of Schadler et al. [5] which defines small consignments and polybags in particular based on their physical characteristics [5]. Due to the flexural nature of the packaging material of small

[2] According to a study from 2008 conducted by Ballot and Fontane, the volume share of traditional cardboard boxes is only about 10% of the total loading volume of transport vehicles [7].

[3] Despite these benefits, the bulk handling of polybags also poses major challenges when it comes to singulation, due to their tendency to stick together and be difficult to separate.

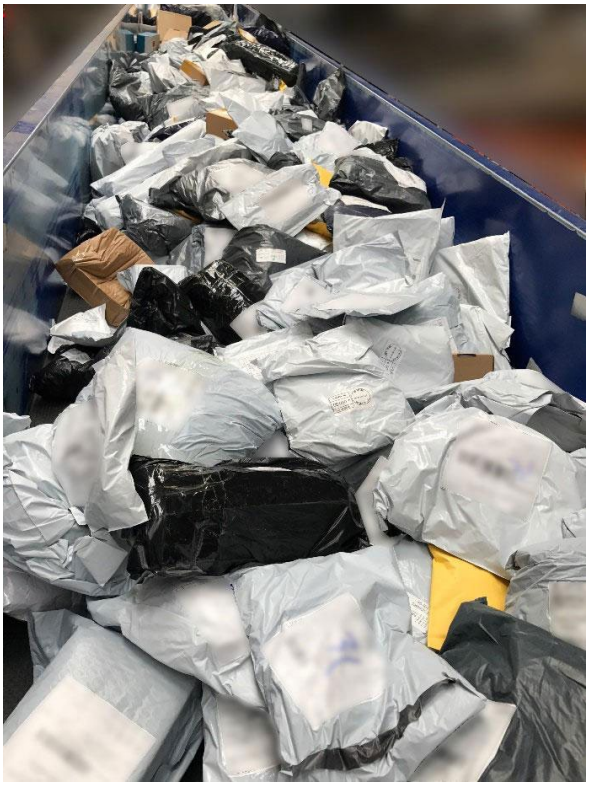


Figure 2: Mixed-mail after bulk unloading from the transportation containers onto a belt conveyor prior to singulation and sortation [5].

consignments, their geometric shape is determined primarily by the consignment's contents rather than the packaging itself. Given the poor dimensional stability of small consignments, it must be ensured that there are no large gaps in the sorting and conveying systems where the consignments can become stuck. This would either lead to damage to the consignment or a stoppage of the system, as the error usually has to be manually corrected. Furthermore, slippage of small consignments on sorting equipment can occur, which often leads to sorting errors [5]. In summary, the widespread use of small consignments has greatly changed the specifications for sorting technology in recent years.

The packaging materials used for small consignments are diverse. Polybags made of *polypropylene* (PP), *polyethylene* (PE), *polyethylene terephthalate* (PET), and bio-degradable polymers are frequently used [5]. Some polybags have air cushioning to protect their contents. Another widespread flexible packaging type is kraft paper. Kraft paper packaging sometimes has an inner layer of air-cushioned plastic. As a result of the different materials used in small consignments, the physical properties such as flexibility and friction values can vary immensely between different small consignments. In addition to the variety of different packaging types, the physical characteristics of the contents can be even more diverse. From a material

handling standpoint, the contents can vary in terms of size, mass, shape, number (single part or multi-part), flexural behavior, fill level, and looseness (fixed or freemoving), all of which lead to a wide range of different physical behaviors [5].

For the reasons mentioned above, there are currently only limited machine solutions for processing small consignments. As a result, many additional manual activities are required to process mixed-mail items within distribution centers, which in turn leads to increased economic costs. Furthermore, the minimum gap between items on a sorter impacts the sorter's throughput. For mixed-mail items, the minimum spacing, especially when roller conveyors are used, often needs to be greater than for traditional parcels to avoid missorting, which reduces throughput and therefore increases costs [5].

While private couriers can refuse shipments with unfavorable characteristics, designated postal operators are generally required by their government to deliver all types of mail and packages, regardless of their characteristics, as long as they meet certain basic requirements, such as being properly packed and labeled. Therefore, in order to make mixed-mail processing more profitable, new automated material handling equipment, tailored to the requirements of mixed-mail, is needed in the medium term. Ideally, new equipment should be modular to allow for adjustments when market demands or packaging regulations change. The diverse nature of small consignments makes the design of new equipment a challenging task, however. A central problem in this respect is selecting suitable test specimens to test new solutions. Live mail is typically not available for this task, and in the rare instances where live mail can be used, these consignments are only available for a very short time (usually only one or two days) and must be handled with extreme care to ensure the postal operator's quality of service. This means that manufacturers of material handling equipment have to resort to test mail created specifically for this purpose, and the number of different types of test consignments is often limited. While it is relatively easy to find suitable test specimens for cardboard boxes, as these parcels are fairly similar to one another, the task becomes much more difficult for small consignments with flexible packaging. As discussed before, small consignments exhibit a wide range of physical properties, but their characteristics and combinations are not evenly distributed. Therefore, it is essential to select representative test specimens that capture the most common mixed-mail item types.

In addition to physical testing, the selection of representative test specimens is also critical for virtual prototyping as well as digital twin simulations. Many companies would like to test the usability of their equipment in a virtual environment (digital twin) in case of short-term changes in the shipment spectrum, e.g., after changes in legislation. Similar to physical testing, it is necessary to cover the most common types of mixed-mail for this purpose, which makes a methodical approach to selecting representative test specimens all the more important.

This study aims to investigate the effectiveness of using cluster analysis to select representative test specimens that capture the most common mixed-mail item types. Specifically, the study aims to collect data from live mail and use cluster analysis to identify clusters of similar mixed-mail items that can serve as a template for test specimens. While previous research by Schadler et al. [5] has explored the characteristics of mixed-mail items, there is currently no methodical and statistical classification of mixed-mail that accounts for the various possible combinations of small consignment characteristics.

One of the challenges of using cluster analysis on this type of data is that some features such as dimensions and mass are continuous while other features such as packaging type are categorical. The difficulty lies in finding a suitable distance metric or model capable of handling both types of data simultaneously. While specific clustering algorithms for mixed-variable data exist, homogenizing mixed-variable data before clustering can greatly expand the available options for an appropriate clustering algorithm. There are several ways to homogenize mixed-variable data. One popular choice is to use a distance metric such as *Gower's distance* [9]. Another approach is through the dimensionality reduction method *Uniform manifold approximation and projection* (UMAP) [10], which has been successfully used as a pre-processing step to cluster high-dimensional numerical data and has been shown to outperform clustering methods that did not involve dimensionality reduction [11, 12, 13]. However, to our knowledge, UMAP has not been used in any published study to homogenize mixed-variable data. Choosing a density-based clustering algorithm over a distance-based one for clustering mixed-variable data after homogenization provides several advantages in this study. First, density-based algorithms do not make any assumptions about the shapes of clusters and can detect outliers in regions of low densities. The latter is particularly important for excluding small consignments with unusual characteristics from the final result. Additionally, density-based clustering algorithms do not require the number of clusters, which is not usually known, to be specified a priori. Among the various density-based clustering algorithms available, *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) is one of the best performing and has been shown to outperform other state-of-the-art density-based clustering algorithms [14].

In the first part of this study, different methods of homogenization in conjunction with HDBSCAN will be assessed in a benchmark study. Since real-world datasets for clustering mixed-variable data with known ground truth are not widely available and are often very problem-specific, a simulation approach was chosen.

Benchmark studies are a popular approach to compare the performance of different clustering algorithms [15]. The main advantage of using simulated datasets as opposed to real-world datasets is the ability to control the characteristics of each dataset, thus enabling a systematic parameter study. Most benchmark studies of clustering algorithms in the literature that use simulated datasets are conducted with either continuous or categorical data, whereas mixed-type datasets are less common. Examples of the latter include works by Foss et al. [16], Jimeno, Roy, and Tortora [17], Preud'homme et al. [18], D'Urso, De Giovanni, and Vitale [19] and Costa, Papatsouma, and Markos [15]. In the second part of this study, the most appropriate clustering approach identified in the benchmark study will be employed to cluster the Mixed-Mail dataset, and the results will be analyzed. The clustering approach chosen should be capable of efficiently clustering the collected live mail data with high accuracy while also detecting and handling outliers effectively. By using the chosen method, it should be possible to identify distinct groups of mail items with similar characteristics and gain insights into patterns that may not be apparent through manual inspection.

2 THEORETICAL BACKGROUND

2.1 HDBSCAN

HDBSCAN is a density-based clustering algorithm developed by Campello, Moulavi, and Sander [14] which can be viewed as an extension of the widely-used *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) algorithm. The main difference compared with DBSCAN is that instead of generating a single flat clustering solution, HDBSCAN uses hierarchical clustering to obtain a hierarchy of different clustering solutions and subsequently extracts a flat clustering solution based on cluster stability. This comes at the cost of computational time.

Compared with other clustering algorithms HDBSCAN has three main advantages. First, the number of clusters is not an input parameter but instead is determined by the algorithm. This is especially useful for clustering high-dimensional datasets where the number of clusters is not known a priori and cannot be estimated easily. As will later be shown, HDBSCAN's hyperparameters are intuitive and the algorithm generally produces sensible clusters even without hyperparameter tuning. Second, HDBSCAN makes no assumptions about the shapes of the clusters. This is in contrast to distance-based algorithms such as k-Means which assume convex-shaped (hyper-spherical or hyper-elliptical) clusters [20]. Third, HDBSCAN can detect outliers and doesn't assign these points to a cluster.

The algorithm starts by estimating the local density of each data point in a dataset, which is defined as the inverse of the core distance. The core distance of a point is the distance of the point to its m_{pts}-th nearest neighbor, where m_{pts} is a hyperparameter. When done for all points in the dataset this gives an estimate for the *probability density function* (PDF) of the dataset.[4] From the core distances, the mutual reachability distance between two points x_p and x_q is obtained as the maximum value of the core distances of x_p and x_q as well as the distance between x_p and x_q. The mutual reachability distances between all pairs of points in a dataset are the edge weights of the mutual reachability graph with the data points as vertices. Next, the algorithm computes the *minimum spanning tree* (MST) from the mutual reachability graph. The MST is a subset of the edges of the complete mutual reachability graph such that all vertices are connected, the total edge weights are minimized, and there are no cycles. Fig. 3 provides a graphical example of a MST for a dataset consisting of points in a Euclidean plane. The MST is then extended by adding a self-loop, i.e., an edge that connects a vertex to itself, to every vertex, resulting in the *extended minimum spanning tree* (MST_{ext}) with the vertex's core distance as its weight. In the next step, a dendrogram is computed to capture the HDBSCAN hierarchy, as depicted in Fig. 4a. This is done in an iterative fashion starting with all points having the same label, which signifies that all points belong to the same cluster. All of the edges are then removed iteratively from the MST_{ext} in decreasing order of their weights. The weight of the edge to be removed denotes the hierarchical level of the dendrogram. In case two edges have the same weight they are removed simultaneously. After each removal, cluster labels are assigned to the connected part(s) that contain(s) the end vertex/vertices of the removed edge(s). If the resulting connected part(s) contain(s) fewer than m_{clSize} vertices the corresponding points are assigned the label "noise." m_{clSize} is a hyperparameter that serves as a lower limit for the number of points in a cluster. In the last step of the HDBSCAN algorithm, a flat clustering solution is obtained by analyzing the dendrogram. A simple way would be to select a level from the hierarchy (a flat line in the dendrogram) and use the corresponding clusters. This is essentially what DBSCAN does[5] but the density threshold for the cut is an unintuitive hyperparameter. Another problem with this approach is that a single density threshold doesn't allow for variable density clusters as this corresponds to removing all edges from the MST_{ext} whose weights are above a certain threshold. The solution to this problem is to cut the dendrogram at different hierarchy levels. To

[4] DBSCAN in contrast uses the non-intuitive and datasetdependent hyperparameter ε to specify a radius around a data point and then counts the number of points that fall inside that radius for every point in the dataset

[5] without building a hierarchy first

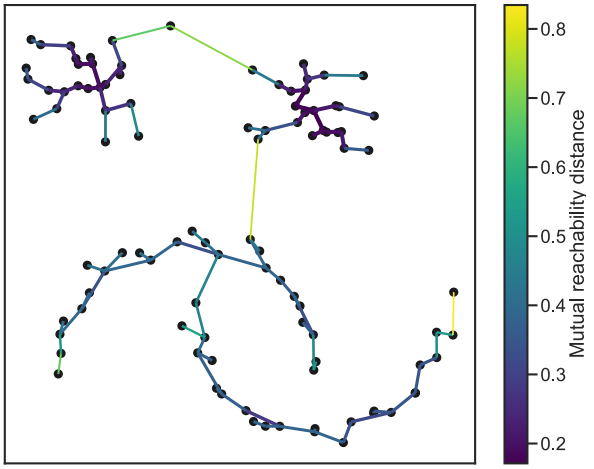


Figure 3: Minimum spanning tree of points in a Euclidean plane. The edge weights are colored by the mutual reachability distance values between the corresponding points [21].

achieve this, the cluster hierarchy has to be condensed. Condensing is done by removing hierarchy levels that do not result in a cluster split but rather only shrink clusters by splitting off individual noise points (or groups with fewer than m_{clSize} points). That way, only the hierarchy level corresponding to "true" splits, i.e., splits where two or more clusters at least as large as m_{clSize} emerge from a single cluster, remain.

The result is a condensed dendrogram, as shown in Fig. 4b. The goal now is to select cut points in a way that the resulting clusters persist the longest in the condensed dendrogram, meaning that for every cluster a decision has to be made whether or not to split the cluster further into sub-clusters. For this, the cluster persistence measure λ is introduced, which is defined as the inverse of the mutual reachability distance. λ_{birth} for a given cluster denotes the λ-value when the cluster is formed by splitting off from a parent cluster while λ_{death} is the λ-value when a cluster splits into smaller clusters. λ_p, where $\lambda_{\text{birth}} \leq \lambda_p \leq \lambda_{\text{death}}$, is defined as the λ-value where point p of a cluster drops out of the cluster, either by being labeled as "noise" or by a cluster split. The stability of a given cluster is given by

$$\sum_{p \in \text{cluster}} (\lambda_p - \lambda_{\text{birth}})$$

Using cluster stability, a flat clustering solution is obtained by selecting the clusters with the highest stabilities from the condensed dendrogram, subject to the constraint that when a cluster is selected none of its descendants can be selected. A graphical example of a flat clustering solution in a condensed dendrogram is provided in Fig. 4b, where the resulting clusters are marked by ellipses. The two most important hyperparameters for HDBSCAN are the minimum cluster size m_{clSize}, which is self-explanatory, and the number of nearest neighbors m_{pts} used for the estimation of the PDF. The latter parameter is a measure of how conservative a clustering solution will be. Larger values lead to fewer points being assigned to a cluster, i.e., more points being labeled as noise/outliers, and there tend to be fewer clusters in total. Small numbers of m_{pts}, in contrast, lead to fewer outliers and a large number of smaller clusters [22]. The authors propose a way to simplify the hyperparameters even further by setting $m_{\text{pts}} = m_{\text{clSize}}$, which effectively turns HDBSCAN into a clustering algorithm with a single intuitive hyperparameter [20].
Comprehensive descriptions of the algorithm featuring graphical examples can be found in [20, 23, 21].

2.2 UMAP

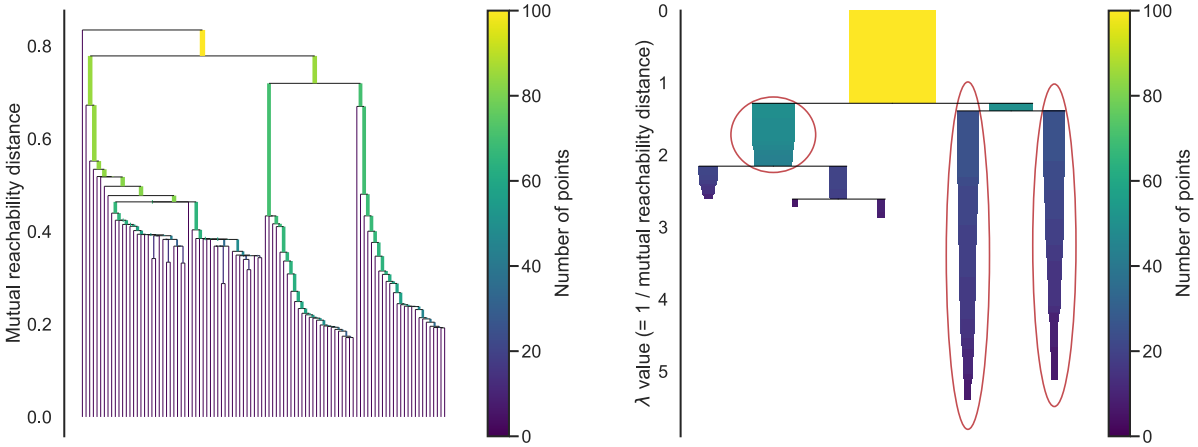
Uniform manifold approximation and projection (UMAP) [10] is a nonlinear dimensionality reduction method based on manifold learning and aims to preserve both the local and global structure of a dataset in some lower dimension. It is similar to *t-distributed Stochastic Neighbor Embedding* (t-SNE) [24] but is faster than t-SNE and thus scales better with large datasets. It is also better at preserving global structure [10]. UMAP relies on the following three assumptions:

1. The data is distributed uniformly on a Riemannian manifold.
2. The manifold is locally connected.
3. The manifold is locally constant.

Using these assumptions, UMAP's main idea is to approximate the manifold of some high-dimensional data with a fuzzy simplicial set representation retaining all relevant topological information. This fuzzy topological representation is then used to construct a low-dimensional fuzzy simplicial set with similar properties. The high-dimensional simplicial set representation is an undirected weighted graph G where the edge weights are the similarities of pairs of points. Before G can be constructed, a directed graph $G^* = (V,E,\omega)$ has to be found, where V is the set of vertices, i.e., the data points of a dataset, E is the set of directed edges, and ω is the weights of the edges. The edge weight $w((x_i, x_{ij}))$ between vertex x_i and its j-th nearest neighbor x_{ij} is given by

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right),$$

where $d(x_i, x_{ij})$ is the distance between those points and ρ_i is the distance between point x_i and its nearest neighbor. σ_i is a normalization factor that is set for every point x_i with respect to its k nearest neighbors by the following equation.



(a) Dendrogram

(b) Condensed dendrogram with clusters marked by ellipses.

Figure 4: Example of a dendrogram and the corresponding condensed dendrogram [21]. In this example, there are only four "true" cluster splits, i.e., splits where two (or more) sub-clusters emerge from a single parent cluster.

using attractive and repulsive forces which are derived from the gradients of the cross-entropy between the weights of G and H. Through this process, the low-dimensional graph H is optimized to closely conform to the topology of the high-dimensional data represented by G.

The most important hyperparameters of UMAP are the number of nearest neighbors k, the embedding dimensionality, and the minimum distance of two points in the low-dimensional representation. For large values of k more of the global structure of the data is preserved, whereas small values of k lead to a focus on local structure. Further details on the algorithm itself and the implementation can be found in [10].

Making use of the combinatorial nature of simplicial sets used in UMAP, it is possible to combine two or more different fuzzy simplicial set representations of the same underlying data using intersections or unions. This property is useful for finding UMAP embeddings of mixed-variable data. For example, a dataset may be split into numerical and categorical variables with distance matrices based on Euclidean and Dice distance, respectively. UMAP can be used to find high-dimensional fuzzy topological representations for both the numerical and the categorical part of the data, which can then be combined to obtain a composite representation. This composite representation can then be embedded into low-dimensional space as described above. This approach effectively homogenizes mixed-variable data for clustering purposes [25].

2.3 Gower's Distance

Gower's distance $D_{\text{Gower}}(x, y)$, initially proposed by Gower [9], is a distance measure used to calculate the dissimilarity between two records x and y with mixed numeric and categorical data. It is defined as

$$D_{\text{Gower}}(x, y) = 1 - \frac{1}{m} \sum_{j=1}^{m} s_j(x, y),$$

where m is the total number of variables and $s_j(x, y)$ is a similarity function with respect to variable j. $s_j(x, y)$ depends on the variable type. Its definition is

$$s_j(x, y) = \begin{cases} 1 & \text{if } j \text{ is categorical } \wedge\, x_j = y_j \\ 0 & \text{if } j \text{ is categorical } \wedge\, x_j \neq y_j \\ 1 - \frac{|x_j - y_j|}{R_j} & \text{if } j \text{ is numerical} \end{cases}$$

where R_j is the difference between the largest and the smallest value of j [9].

Gower's distance can take values in the interval [0,1], where 0 means that two records are identical, and higher numbers indicate greater dissimilarity between records.

2.4 Cluster Validation

The clustering solutions obtained from a clustering algorithm can be validated[6] using various metrics,

[6] It should be noted that validation in an engineering context requires objectivity, while clustering is usually a subjective process. However, the term validation is still commonly used in the context of clustering.

which can be divided into external and internal metrics. External validation metrics are used to compare a clustering solution to a known ground truth. Since the ground truth for real-world applications is not known a priori, external validation metrics are mainly used in benchmark studies [26]. Internal validation metrics, on the other hand, depend only on the data at hand and can be used for hyperparameter tuning by comparing different clustering solutions.

2.4.1 External Cluster Validation

Arguably, the two most common external cluster validation metrics used in benchmark studies to compare clustering solutions with a known ground truth are the *Adjusted Rand Index* (ARI) [27] and the *Adjusted Mutual Information* (AMI) [28]. The former is based on pair counting, while the latter is based on information theory. Both measures are extended versions of the *Rand Index* (RI) and *Mutual Information* (MI), respectively, in which a correction for chance is applied. They take the value 0 when the compared partitions are random and 1 when the partitions are identical. For both measures, negative values are also possible if the compared partitions have a worse-than-random agreement, although this has little practical significance. While ARI and AMI are often used simultaneously in benchmark studies, Romano et al. [29] recommend using ARI when clusters are approximately equal in size and AMI when the cluster size distribution is unbalanced. In the study at hand, the cluster sizes are not evenly distributed, so AMI is used for validation. The theory behind AMI is described in detail in [28].

2.4.2 Density-Based Clustering Validation

Internal cluster validation metrics usually assess the similarity of points belonging to the same cluster and the dissimilarity of points belonging to different clusters. Since most internal validation metrics are distance-based they may fail with density-based clustering algorithms like HDBSCAN. This problem is addressed with the *Density-Based Clustering Validation* (DBCV) index [26], which is compatible with density-based clustering algorithms. DBCV uses density sparseness $DSPC_i$ of a cluster C_i and density separation $DSPC_{i,j}$ of a pair of clusters C_i and C_j . The validity V_{Ci} in respect to cluster C_i in a set of l clusters is then defined as

$$V_{C_i} = \frac{\min\limits_{0 \le j \le l, j \neq i} (DSPC_{i,j}) - DSPC_i}{\max\left(\min\limits_{0 \le j \le l, j \neq i} (DSPC_{i,j}), DSPC_i\right)}.$$

The DBCV index $DBCV(C)$ with respect to a clustering solution C is given by

$$DBCV(C) = \sum_{i=1}^{l} \frac{|C_i|}{|O|} V_{C_i}$$

where $|C_i|$ is the size of cluster C_i and $|O|$ is the total number of points in the dataset used for clustering, including noise.
A detailed description of the algorithm including the calculation of *DSC* and *DSPC* can be found in [26].

3 METHODS

3.1 Benchmark Study

We carried out a benchmark study in order to select the method most suitable for clustering the Mixed-Mail dataset presented in section 3.2 using a simulation approach which is inspired by the simulation framework of Preud'homme et al. [18] but differs from it in some aspects.
The clustering performances of four clustering methods were assessed in the benchmark study using simulated mixed-variable datasets. The solutions obtained from each method were compared with the ground truth using the AMI as a performance metric. Section 3.1.1 contains brief descriptions of all four clustering methods featured in this study while section 3.1.2 provides a detailed description of the study framework.

3.1.1 Clustering Methods

Each of the four clustering methods presented in the following paragraphs uses HDBSCAN for clustering in conjunction with a different data homogenization approach. Unless otherwise specified, the default hyperparameters are used as per the documentation of the software used.

Gower & HDBSCAN This clustering method uses *Gower's distance* (see section 2.3) in order to calculate a distance matrix for a mixed-variable dataset. HDBSCAN is then used directly on the distance matrix to find clusters. Python implementations of both *Gower's distance* [30] and HDBSCAN [31] were used for this. We set the minimum cluster size m_{clSize} to 5% of the dataset size and used the default value of $m_{\text{pts}} = m_{\text{clSize}}$ to set the number of nearest neighbors as per the authors' recommendation to reduce complexity.

Gower & UMAP & HDBSCAN This clustering method is similar to the previous one except that it uses UMAP as an intermediary step. We used a Python implementation of UMAP by McInnes, Healy, and Melville [32], authors of the algorithm [10]. The UMAP hyperparameters used in the benchmark study were:

- Number of nearest neighbors $k = 15$ (default value)
- Embedding dimensionality: 2 (default value)
- Minimum distance of two points in the lowdimensional representation: 0.0001

The same HDBSCAN-hyperparameters were used as in "Gower & HDBSCAN."

UMAP Union & HDBSCAN This method splits the dataset into a numerical and a categorical part and calculates the distance matrix for each part using Euclidean distance and Dice distance, respectively. Next, a UMAP model is generated for each distance matrix. While the model for the numeric variables uses the default value for the number of nearest neighbors ($k = 15$), a higher value ($k = 100$) is used for the categorical variable model. This is a necessary shift of focus toward global structure since with categorical variables many points will have the same level settings and thus not a lot of useful information can be gained from the local structure. The two models are then combined using the union operator and embedded into two-dimensional space, as explained in section 2.2. The minimum distance hyperparameter is set to 0.0001.

UMAP Intersection & HDBSCAN This method is identical to "UMAP Union & HDBSCAN" except that the intersection operator is used instead of the union operator to combine the two models.

3.1.2 Study Framework

The datasets used for the simulation study were generated automatically. We used a method developed by Qiu and Joe [33] which is based on Milligan's method [34] to generate multi-dimensional continuous variable clusters with a specified degree of separation as well as noise points (points that do not belong to any cluster). The algorithm is available through the R-package clusterGeneration [35]. We then augmented the datasets generated in this manner with categorical variables for a total of 12 variables (numerical and categorical). First, we randomly chose the number of levels per categorical variable in the [2, 5] interval. Next, we randomly assigned a unique level setting across all categorical variables to each cluster, while we assigned random level settings to the noise points. Finally, we added "local noise" by randomly changing some of the categorical level settings of a certain percentage of non-noise points.

We carried out a parameter study to assess the performance of the different clustering methods for different scenarios. Because of computational constraints, a factorial design was infeasible. Instead, we varied the parameters one at a time. The consequence of this approach was that no parameter interactions could be observed. The benchmark study investigated four dataset parameters at three levels each, resulting in nine experiments, which are listed in table 1. For every parameter setting, we randomly generated a total of 1,000 datasets. The number of points per cluster was randomly selected from a uniform distribution in different intervals depending on the number of clusters in each scenario. While the total number of non-noise data points varies between datasets, the expected total number is constant at 450 for every scenario and is therefore independent of the number of clusters. The following paragraphs detail the parameters used in the benchmark study.

Number of Clusters The number of clusters was tested at three different levels:

- l: 2
- m: 6 (default)
- h: 10

As mentioned above, the number of clusters influences the cluster sizes. The intervals used for the sampling of the cluster sizes were [90, 360], [30, 120], and [18, 72] for the levels l, m, and h, respectively.

Proportion of Categorical Variables This parameter changes the proportion of the categorical variables relative to the total number of variables at three different levels:

- l: 25%
- m: 50% (default)
- h: 75%

The number of numerical variables was chosen accordingly to keep the total number of variables at 12 for all scenarios.

Global Noise This parameter is used to set the number of outlier points (noise) per dataset. Outlier points were added after all non-outlier data points had been created and therefore increase the total number of data points by a certain percentage of the number of nonoutlier points per dataset, depending on the parameter setting. The three different levels used in the study are:

- l: 10%
- m: 30% (default)
- h: 50%

Local Noise This parameter quantifies how similar the points in each cluster are to one another and how well

the clusters are separated from each other. It consists of two parts. The first part is the separation index from Qiu and Joe's method [33] used for the numerical variables. The separation index I_{sep} falls somewhere in the] - 2, +2[interval, with bigger numbers signifying a higher degree of separation between clusters.
The second part of this parameter is the amount of categorical noise Q_{cat}. Depending on the parameter setting, Q_{cat}% of all non-outlier points is selected at random. For each selected point, a random selection of Q_{cat}% of its categorical variables is changed to a random different setting. The parameter levels and the corresponding settings for I_{sep} and Q_{cat} are:

- l: $I_{sep} = 0.3$, $Q_{cat} = 0$
- m: $I_{sep} = 0.0$, $Q_{cat} = 15$ (default)
- h: $I_{sep} = -0.3$, $Q_{cat} = 30$

3.2 Mixed-Mail Data Collection

The Mixed-Mail dataset was recorded at the logistics centre Salzburg of Austrian Post during nighttime hours in spring 2022. This location handles the destination distribution of Austrian Post and receives pre-sorted shipments from various federal states, which are sorted overnight. This provides a relatively high degree of mixing and is therefore fairly representative of the volume of shipments throughout Austria. In total, more than 600 consignments were randomly selected and analyzed non-destructively, and the following properties were recorded:

- Size
- Mass
- Packaging
- Shape
- Flexibility
- Movability of contents
- Fill level
- Multi-part contents

Due to the reasons given below, we only used 406 of the consignments for clustering. The following paragraphs offer short descriptions of the recorded properties.

Size The size of each consignment was measured using the measuring device illustrated in Fig. 5, allowing all necessary dimensions to be read in one step. The measurement was made with a resolution of 1 cm, since recording in the millimeter range is not practical, as deviations occur due to the deformability of the objects.

For use in clustering, we combined the dimensions into a single variable, "size," defined as the sum of length, width, and height.

Mass The mass of each consignment was measured using a scale with an accuracy of 1 g.

Figure 5: Experimental setup showing the dimension measurement.

Packaging The classification of each consignment's packaging into one of five categories (cardboard, polybag, kraft paper, kraft paper w/ bubble wrap, and bubble wrap) was based on previous research by Schadler et al. [5] and Schadler et al. [36] and was refined to provide finer distinctions between packaging types. The definition of polybag encompasses all plastic packaging, excluding bubble wraps, while the term kraft paper refers to pure kraft paper packaging and not to a combination of an inner layer of bubble wrap and an outer layer of kraft paper, which is referred to as kraft paper w/ bubble wrap.

During the analysis, we assigned each sample to one of the aforementioned categories. However, we excluded consignments with cardboard packaging from the cluster analysis as the properties *multi-part contents* and *fill level* could not be determined for them, resulting in a total of 406 items for clustering.

Shape We categorized the consignments into four groups based on their physical form. Consignments with a box-like appearance were classified as "cubic," those with a cylindrical shape were designated as "cylindrical," those with a bulky, ill-defined shape were categorized as "convex/concave," and the remaining consignments were classified as "flat." The height of flat consignments tends to be lower in comparison with their length and width, however, no specific proportion threshold was established. The four different shapes considered in this study are depicted in Fig. 6.

Flexibility Both its packaging and contents influence the flexural characteristics of a consignment. The only rigid packaging we observed was cardboard, but as mentioned earlier, consignments with cardboard packaging were excluded from the cluster analysis. To evaluate the flexibility of a consignment, we employed the procedure shown in Fig. 7. This involved clamping approximately 30% of each consignment length-wise while allowing the opposite end to freely hang. We then recorded the resulting bending angle α and classified each consignment into three levels of flexibility based on the following criteria:

(a) *Cubic*

(b) *Cylindrical*

(c) *Convex/concave*

(d) *Flat*

Figure 6: The different consignment shapes considered in this study.

- Level 1: $\alpha \leq 5$
- Level 2: $5 < \alpha \leq 45$
- Level 3: $\alpha > 45$

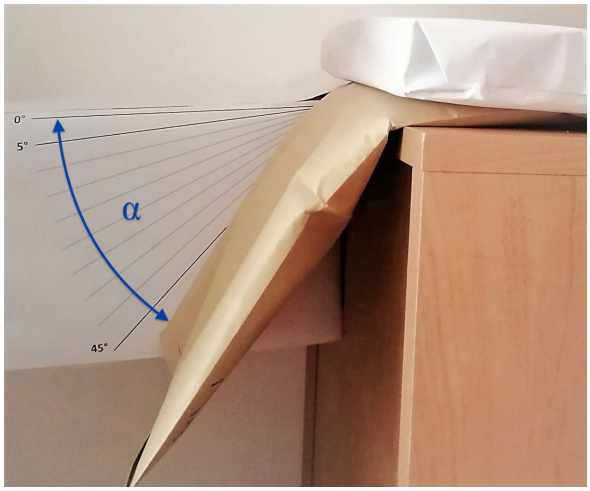


Figure 7: Experimental setup showing the flexibility measurement. In this case, the bending angle α exceeds 45°, indicating a flexibility level of 3.

Movability of Contents The movability of contents was *true* for a consignment where the contents could move freely inside the packaging and *false* if the contents were secured.

Fill Level We determined four distinct fill levels through manual tactile evaluation. We based the classification on the following filling degree ranges:

- Level 1: below 25%
- Level 2: from 25% to 50%
- Level 3: from 51% to 75%
- Level 4: above 75%

An exact determination of the degree of filling is not possible by simple means. This rough classification was deemed adequate for the present study and did not lead to ambiguities in data collection.

Multi-part Contents As with fill level, this property was also determined through tactile evaluation. We classified consignments containing two or more distinct items as *true* and consignments containing only a single item as *false* with respect to the property multi-part contents.

3.3 Cluster Stability

In the absence of ground truth, the validity of a clustering solution can be assessed through a stability measure that quantifies the reproducibility of the solution. This is typically achieved by generating perturbed versions of the original dataset, which are then clustered using a clustering algorithm. Subsequently, the resulting clustering solutions are compared pairwise using a similarity measure such as ARI or AMI. The mean similarity value across all pairs of clustering solutions serves as a metric of the stability of the clustering algorithm with respect to the original dataset [37].

In this study, we utilized this methodology to evaluate the clustering solution of the Mixed-Mail dataset. This was accomplished by creating 1,000 variations of the original dataset through a process of random subsampling without replacement, with each subsample representing 80% of the original dataset's size. We then clustered each variation and constructed a similarity matrix through pairwise comparisons between all clustering solutions using AMI which was limited to the set of points that were present in both variations in a pairwise comparison. Finally, we took the average of all entries from the similarity matrix, excluding the main diagonal.

4 RESULTS

4.1 Benchmark Study

The results of the benchmark study for each clustering method along with the dataset parameters are summarized in Table 1. The results are given as the mean and standard deviation of the AMI score over all 1,000 randomly generated datasets for each experiment and clustering method. The table also shows means and standard deviations of the proportion of data points assigned to a cluster, i.e., not points labeled as "noise," for each method and scenario. A different view of the benchmark results is provided in Fig. 8, which shows boxplots of the performance of each clustering method

Table 1: Parameter levels and clustering method performance for each simulation experiment. For each experiment, the non-default variable level as well as the AMI mean and standard deviation of the best-performing clustering method are highlighted.

Experiment	Variable level				AMI mean ± standard deviation Cluster proportion mean ± standard deviation			
	NC	PC	GN	LN	G/H	G/U/H	UU/H	UI/H
1	6	50%	m	m	**0.96 ± 0.03** 75.6% ± 2.3%	0.79 ± 0.03 95.9% ± 3.0%	0.91 ± 0.03 99.9% ± 0.4%	0.76 ± 0.14 94.4% ± 4.3%
2	**2**	50%	m	m	**0.95 ± 0.05** 77.3% ± 2.2%	0.71 ± 0.07 95.7% ± 3.9%	0.90 ± 0.04 100% ± 0.0%	0.76 ± 0.11 93.0% ± 7.2%
3	**10**	50%	m	m	**0.90 ± 0.04** 67.4% ± 4.4%	0.79 ± 0.02 92.7% ± 4.2%	0.89 ± 0.03 98.5% ± 2.4%	0.75 ± 0.15 92.6% ± 5.7%
4	6	**25%**	m	m	**0.79 ± 0.03** 65.2% ± 2.5%	0.71 ± 0.02 95.7% ± 3.1%	0.62 ± 0.11 95.1% ± 4.3%	0.64 ± 0.13 85.2% ± 9.3%
5	6	**75%**	m	m	**0.97 ± 0.02** 77.5% ± 1.4%	0.82 ± 0.06 97.9% ± 2.4%	0.94 ± 0.02 100% ± 0.4%	0.74 ± 0.17 97.2% ± 3.3%
6	6	50%	**l**	m	**0.97 ± 0.03** 89.5% ± 2.2%	0.88 ± 0.01 99.9% ± 0.4%	0.90 ± 0.03 99.5% ± 1.1%	0.88 ± 0.02 98.9% ± 1.1%
7	6	50%	**h**	m	**0.95 ± 0.04** 65.0% ± 2.6%	0.75 ± 0.05 92.1% ± 4.8%	0.90 ± 0.03 99.8% ± 0.7%	0.71 ± 0.18 93.8% ± 5.4%
8	6	50%	m	**l**	**0.95 ± 0.02** 78.6% ± 1.2%	0.83 ± 0.04 98.2% ± 2.2%	0.90 ± 0.04 99.9% ± 0.3%	0.79 ± 0.11 95.9% ± 3.2%
9	6	50%	m	**h**	0.65 ± 0.04 54.5% ± 3.1%	0.71 ± 0.04 92.7% ± 4.4%	**0.72 ± 0.07** 97.9% ± 2.4%	0.69 ± 0.14 90.3% ± 6.7%

AMI: Adjusted Mutual Information
NC: Number of clusters; PC: Proportion of categorical variables; GN: Global noise; LN: Local noise
G/H: Gower & HDBSCAN; G/U/H: Gower & UMAP & HDBSCAN; UU/H: UMAP Union & HDBSCAN;
UI/H: UMAP Intersection & HDBSCAN

in terms of AMI scores for the various parameters. Each row looks at a different dataset parameter while the columns represent the different clustering methods. The parameter levels l, m, and h are color-coded blue, orange, and green, respectively.

It is immediately apparent that the clustering methods "Gower & HDBSCAN," "Gower & UMAP & HDBSCAN" and "UMAP Union & HDBSCAN" generally perform quite well, while "UMAP Intersection & HDBSCAN" performs significantly worse and is also the most inconsistent method, as indicated by the height of the boxes (Fig. 8) as well as the higher standard deviations compared with the other methods for almost every experiment (Table 1). The poor performance of "UMAP Intersection & HDBSCAN" can most likely be attributed to the fact that the combination of the two underlying fuzzy simplicial sets (see section 2.2) reduces the connectivity of the resulting graph. This is because, in contrast to the union operation, the intersection operation leads to high edge weights only for point pairs with a strong connection in both data views, which results in a loss of information. The union operation on the other hand increases the connectivity of the resulting graph.

Among the remaining methods "Gower & HDBSCAN" performs best for most scenarios (bold entries in Table 1) with "UMAP Union & HDBSCAN" being second and "Gower & UMAP & HDBSCAN" coming third. However, there were two notable exceptions. The first exception is the poor performance of "UMAP Union & HDBSCAN" when the number of categorical variables is low. In fact, each method performed significantly better with a higher proportion of categorical variables. This behavior is to be expected since the differentiation gained from categorical variables is usually much higher than for numerical variables.[7] Thus, a smaller

[7] Categorical variables between two points can either match or mismatch, while there are granular distinctions for numerical variables.

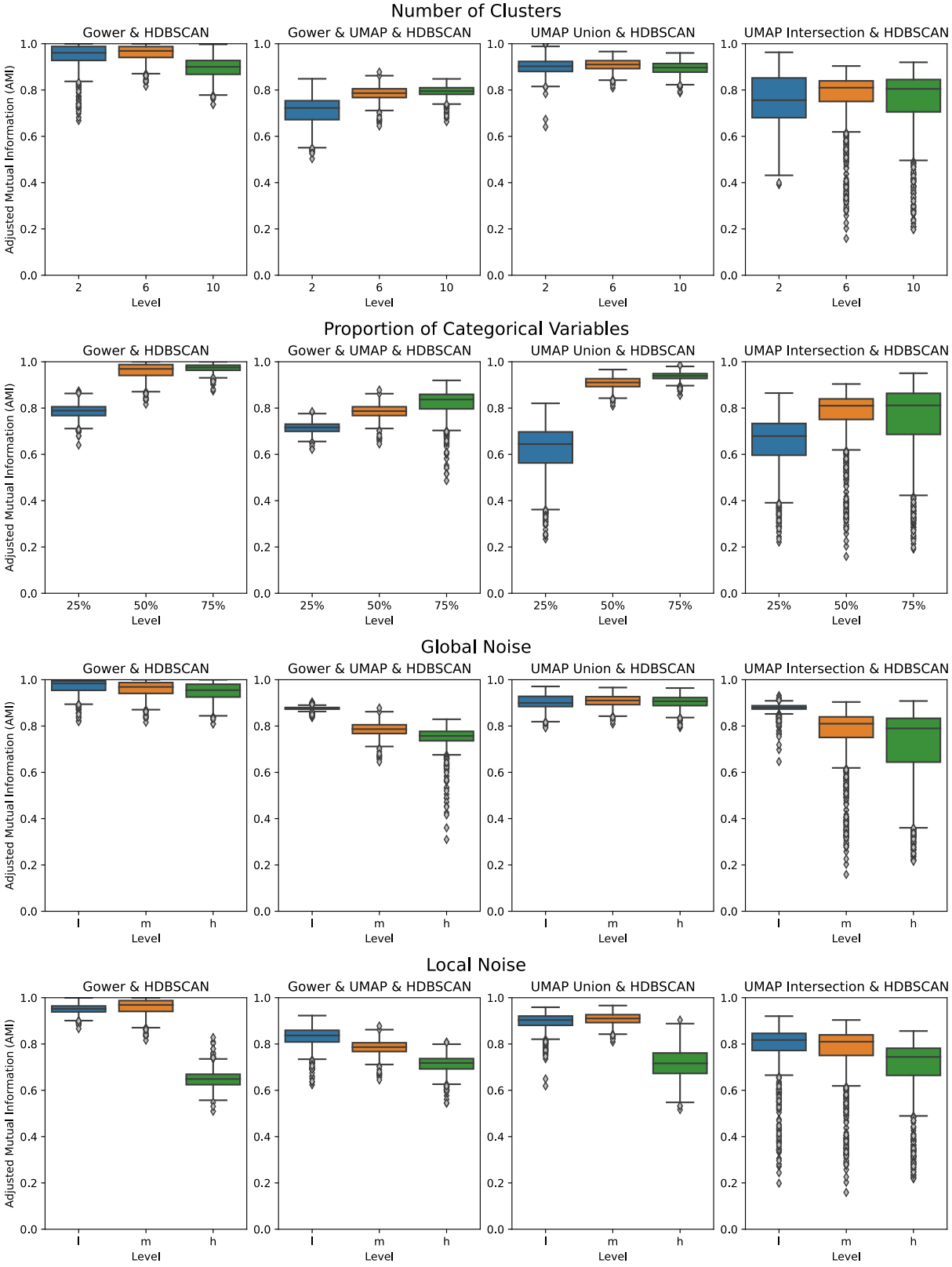


Figure 8: Simulation benchmark results. Each row shows the influence of a different parameter on the clustering performance of each clustering method in the corresponding column.

number of categorical variables results in a lower amount of information that can be extracted from the dataset and used for clustering. For "UMAP Union & HDBSCAN" and "UMAP Intersection & HDBSCAN," this circumstance is even more problematic because a mismatch in the number of categorical and numerical variables is not accounted for in the composition, i.e., both sub-models used for the composition have the same weight, regardless of the number of variables in each sub-model. The second exception to the above performance trend is the apparent poor performance of "Gower & HDBSCAN" in the presence of high local noise. Based on the default hyperparameters used in this study, "Gower & HDBSCAN" displays highly conservative behavior in label assignment and readily classifies data points as outliers, as indicated by the low cluster proportion for the corresponding scenario in Table 1. "Gower & HDBSCAN" generally yielded lower cluster proportions compared with the other methods. Depending on the use case, this behavior can be problematic or advantageous. For instance, if it is known a priori that a dataset does not contain outliers then the goal should be to cluster as many points as possible, making "Gower & HDBSCAN" a suboptimal choice. However, if the goal is to have relatively clean clusters, as is the case with the Mixed-Mail Dataset, "Gower & HDBSCAN" may be the preferred choice. It should also be noted that the hyperparameter m_{pts} used in HDBSCAN can be modified to adjust the conservatism of a clustering solution. Hence, by appropriately tuning the hyperparameters, the benchmark results could significantly differ from the ones presented in this study, where fixed hyperparameter values were used.

Based on the benchmark results and subsequent discussion, "Gower & HDBSCAN" was deemed the most suitable method for clustering the Mixed-Mail dataset, albeit not by a large margin.

4.2 Mixed-Mail Clustering

4.2.1 Mixed-Mail Dataset

A total of 406 mixed-mail items were recorded according to the methodology described in section 3.2. Table 2 provides an overview of the various variable distributions of the Mixed-Mail dataset. For continuous variables, the observed ranges along with the mean and standard deviation are given. Nominal, Boolean, and ordinal variables are presented in terms of their frequency counts.

Fig. 9 shows the correlation strengths between all pairs of variables in the Mixed-Mail dataset.[8] In this

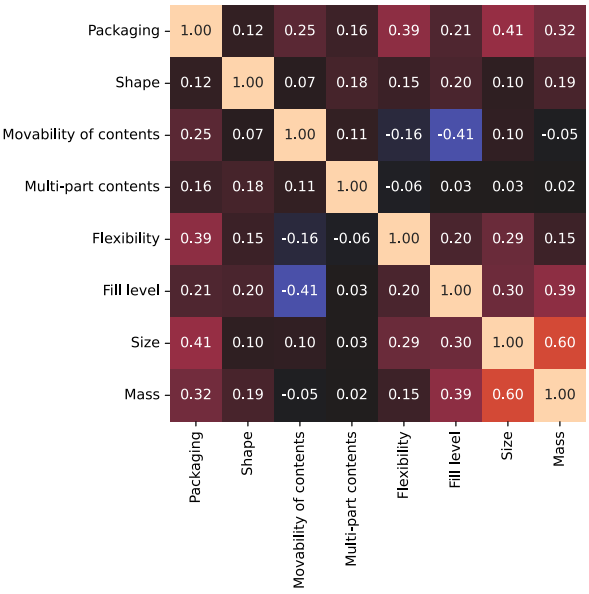


Figure 9: Correlation matrix between the different variables of the Mixed-Mail dataset. Pearson's R is used for numerical-numerical cases, Cramer's V for categorical-categorical cases, and Correlation Ratio for numerical-categorical cases.

view, the combined variable "size" is used instead of the separate dimension variables length, width, and height (see section 3.2). Most of the correlation strengths are not surprising given the nature of the dataset. These include the relatively strong correlations between flexibility and packaging, size and mass, fill level and mass as well as the negative correlation between fill level and the movability of contents, indicating that the contents of tightly packed consignments tend to be less movable.

An interesting observation, however, is the relatively high correlation between packaging and size and to a lesser extent mass. This is caused by the fact that kraft paper tends to be used for larger consignments, whereas bubble wrap is associated with smaller consignments. Finally, the correlation between fill level and size can be explained by the fact that packaging for very small items is often oversized relative to the size of the contents, while packaging for larger items is usually more adequately sized.

4.2.2 Mixed-Mail Clustering Results

As indicated in section 2.1, the HDBSCAN algorithm has two primary hyperparameters, namely, m_{clSize} and m_{pts}. The minimum number of points a cluster must contain to be considered a cluster, denoted by m_{clSize}, is fixed at 10, which corresponds to 2.5% of the dataset size and is deemed a reasonable cutoff point. The parameter m_{pts} is set by comparing the DBCV scores of the clustering solutions resulting from different m_{pts} values as shown in Fig. 10. The maximum DBCV score of 0.2005 is attained for an m_{pts} value of 15. The values

[8] The correlation strengths were calculated using Pearson's R for numerical-numerical cases, Cramer's V for categorical-categorical cases, and Correlation Ratio for numerical-categorical cases.

Table 2: Overview of the range of properties of small consignments in the Mixed-Mail dataset.

Variable name	Variable type	Unit	Range (mean ± standard deviation) / Values (frequency)
Packaging	nominal	-	kraft paper w/ bubble wrap (184), polybag (98), kraft paper (89), bubble wrap (35)
Shape	nominal	-	flat (296), cubic (71), convex/concave (35), cylindrical (4)
Movability of contents	boolean	-	true (212), false (194)
Multi-part contents	boolean	-	true (369), false (37)
Flexibility	ordinal	-	1 (332), 2 (51), 3 (23)
Fill level	ordinal	-	1 (44), 2 (59), 3 (93), 4 (210)
Length	continuous	cm	11–45 (26 ± 7.1)
Width	continuous	cm	8–35 (19 ± 5.6)
Height	continuous	cm	1–14 (3 ± 2.6)
Mass	continuous	g	10–1720 (287 ± 349)

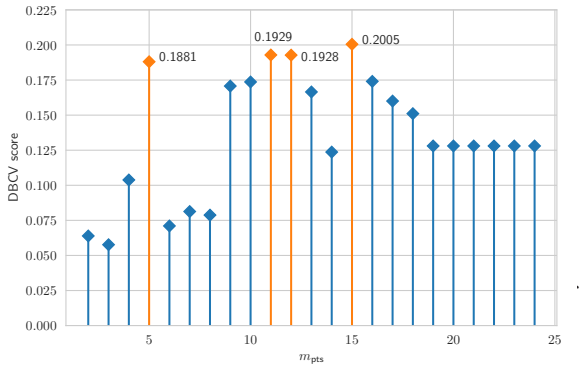


Figure 10: DBCV scores at various m_{pts} settings. The top four settings are marked in orange and the corresponding numerical values are displayed.

numerical (flexibility, fill level, length, width, height, and mass) and categorical, including Boolean variables (packaging, shape, movability of contents, multi-part contents), respectively.

There is a total of six clusters of varying sizes. The non-noise data points contained in these clusters account for 203 instances, representing exactly 50% of the dataset. The clustering outcomes reveal the prevalence of consignments featuring flat shapes and singlepart contents, as well as a high proportion of consignments with low flexibility. However, cluster 3 stands out as an exception, displaying a moderate degree of flexibility. This observation can be explained by the fact that polybag packaging is commonly used for clothing items. Furthermore, the flexibility of the packaging type can significantly influence the overall flexibility of the consignment as well and polybags tend to be the most flexible packaging type of the types considered in this study.

The correlations discussed in section 4.2.1 are also evident in the clustering solution. Clusters 4 and 5, which both have kraft paper packaging, exhibit the largest dimensions among all clusters. On the other hand, cluster 6, which has bubble wrap packaging, displays the smallest dimensions. In addition, the high correlation between fill level and movability of contents is also reflected in the centroids of the clustering solution.

Table 4 shows the corresponding cluster centroids of clustering solution for $m_{\text{pts}} = 11$, which has the second highest DBCV score. This solution has two additional clusters compared with the solution for $m_{\text{pts}} = 15$ and comprises 258 non-noise data points, representing 64% of the dataset. This is not surprising, given that the clustering solution for $m_{\text{pts}} = 11$ is less conservative, resulting in slightly higher numbers of data points for each cluster and affecting the median values of these variables. The first six cluster centroids are almost identical between both solutions, with some variations in the length, width, and mass variables. The differences in the centroid variables between both solutions are highlighted in Table 4. Cluster centroids 7 and 8 in Table 4 both exhibit a cubic shape along with the largest height values among all cluster centroids. These two clusters do not appear in the solution for $m_{\text{pts}} = 15$ due to the cutoff limit posed by $m_{\text{clSize}} = 10$, which is not met in that case.

In summary, both clustering solutions obtained from different m_{pts} values provide very similar results, with the solution for $m_{\text{pts}} = 11$ containing some additional information in the form of two additional clusters. When setting m_{pts} to even lower values, fewer points are classified as noise and thus solutions with even higher numbers of clusters emerge. In the present study, those additional clusters are deemed insignificant and are therefore not investigated.

To evaluate the stability of the clustering solutions, AMI was employed using the methodology described in section 3.3. The resulting mean AMI values for the hyperparameter values $m_{\text{pts}} = 15$ and $m_{\text{pts}} = 11$ were 0.85 and 0.83, respectively, indicating a high degree of stability for both cases.

5 CONCLUSION AND OUTLOOK

This study aimed to use cluster analysis to gain a better understanding of the physical characteristics of mixed-mail in Austria. First, we carried out a benchmark study comparing different approaches for clustering mixed-variable data. Preprocessing the data using *Gower's distance* and clustering the

Table 3: Cluster centroids for the clustering solution for m_{pts} = 15. Total number of data points assigned to a cluster: 203 (50% of the dataset size).

Cluster	Cluster size	Packaging	Shape	Movability of contents	Multi-part contents	Flexibility	Fill level	Length [cm]	Width [cm]	Height [cm]	Mass [g]
1	75	Kraft paper w/ bubble wrap	Flat	True	False	1	3	27	19	2	140
2	60	Kraft paper w/ bubble wrap	Flat	False	False	1	4	23	16	2	80
3	24	Polybag	Flat	False	False	2	4	29	22	2	180
4	17	Kraft paper	Flat	False	False	1	4	33	25	3	380
5	16	Kraft paper	Flat	True	False	1	3	32	20	2	140
6	11	Bubble wrap	Flat	False	False	1	4	16	13	2	80

Table 4: Cluster centroids for the clustering solution for m_{pts} = 11 with the differences in the centroid variables compared with the solution for m_{pts} = 15 highlighted. Total number of data points assigned to a cluster: 258 (64% of the dataset size).

Cluster	Cluster size	Packaging	Shape	Movability of contents	Multi-part contents	Flexibility	Fill level	Length [cm]	Width [cm]	Height [cm]	Mass [g]
1	77	Kraft paper w/ bubble wrap	Flat	True	False	1	3	27	20	2	140
2	61	Kraft paper w/ bubble wrap	Flat	False	False	1	4	23	17	2	80
3	34	Polybag	Flat	False	False	2	4	30	23	2	185
4	21	Kraft paper	Flat	False	False	1	4	33	25	3	340
5	23	Kraft paper	Flat	True	False	1	3	32	23	2	140
6	16	Bubble wrap	Flat	False	False	1	4	18	14	2	75
7	14	Kraft paper w/ bubble wrap	Cubic	True	False	1	4	28	22	4	370
8	12	Polybag	Cubic	False	False	1	4	27	15	4	150

homogenized data using the HDBSCAN clustering algorithm showed the best outcomes across a variety of synthetic datasets with different properties. We then applied this approach to cluster the Mixed-Mail dataset obtained from live mail at an Austrian Post logistics center. We presented two different clustering solutions based on two different hyperparameter settings, both of which provided comparable results with the less conservative hyperparameter setting identifying two additional clusters. The clusters represent some of the most common manifestations of mixed-mail and make the complexity of the heterogeneous characteristics of mixed-mail more manageable. The cluster centroids can be used as templates for creating test specimens for testing material handling equipment with regard to its ability to handle mixed-mail. Furthermore, this study establishes a framework that can be extended to other scenarios where a large variety of heterogeneous objects must be reduced to a manageable number while preserving important information. Examples of technical applications include waste management, robotics, e.g. for material handling tasks, and inventory management. In addition, this approach could also be useful in other fields where mixed-variable data is common such as medicine, psychology, marketing, and e-commerce.

Due to budget limitations, the Mixed-Mail dataset used in this study includes only a small subset of the live mail from a single logistics center on a single night. Although similar findings would likely be obtained at other locations in Austria or even Central Europe, the results of our study cannot be directly extrapolated to such scenarios. Therefore, further research encompassing various locations and longer durations is necessary to validate our findings. Given the rapid changes in the CEP market and its goods, it is advisable to conduct such studies periodically to ensure the validity of the findings. Future research should also expand on the benchmark study conducted in this paper by considering additional clustering methods and by utilizing a factorial design approach to obtain a better understanding of the impact of various dataset characteristics on the clustering performance of different algorithms.

AUTHOR CONTRIBUTIONS

For brevity, the authors' initials are used. Conceptualization, D.S., H.S., and D.K.; methodology, D.S. and H.S.; software, D.S.; validation, D.S.; formal analysis, D.S.; investigation, D.S., and H.S.; data curation, D.S., and H.S.; writing – original draft, D.S.; writing – review and editing, D.S., H.S., C.L., and D.K.; visualization, D.S., and H.S.; supervision, C.L.; project administration, D.S., H.S., and C.L.; funding acquisition, D.S., and C.L. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

This research was conducted as part of the Austrian Research Promotion Agency (FFG)-funded project „Entwicklung eines Simulationsansatzes zur Analyse von Kleinsendungen (ISAAK)" (No. 891088) in collaboration with the project partners Körber Supply Chain Logistics GmbH in Konstanz and Austrian Post.

We wish to express our gratitude to Austrian Post for enabling us to conduct the data collection required for this research. Furthermore, we would like to thank the former employees of the Institute of Logistics Engineering, Martin Knödl and Bastian Mayer, for their assistance during data collection.

REFERENCES

[1] Esser, K. and Kurte, J. (2019) KEP-Studie 2019 – Analyse des Marktes in Deutschland. Eine Untersuchung im Auftrag des Bundesverbandes Paket und Expresslogistik e. V. (BIEK).

[2] Esser, K. and Kurte, J. (2022) KEP-Studie 2022 – Analyse des Marktes in Deutschland. Eine Untersuchung im Auftrag des Bundesverbandes Paket und Expresslogistik e. V. (BIEK).

[3] Pitney Bowes (2021) Parcel Shipping Index 2021. url: https://www.pitneybowes.com/content/dam/pitneybowes/us/en/shipping-index/22-pbcs-04529-2021-global-parcel-shipping-index-ebook-web-002.pdf (visited on 04/15/2023).

[4] Pitney Bowes (2022) Parcel Shipping Index 2022. url: https://www.pitneybowes.com/content/dam/pitneybowes/us/en/shipping-index/22-pbcs-04529-2021-global-parcel-shipping-index-ebook-web-002.pdf (visited on 04/15/2023).

[5] Schadler, M., Schedler, M., Knödl, M., Prims, D., Landschützer, C., and Katterfeld, A. (2022) Characteristics of 'Polybags' Used for Low-Value Consignments in the Mail, Courier, Express and Parcel Industry. *Logistics Journal*, 25.

[6] International Post Corporation (2023) Cross-Border e-Commerce Shopper Survey 2022. Survey.

[7] Ballot, E. and Fontane, F. (2008) Rendement et Efficience Du Transport: Un Nouvel Indicateur de Performance. *Revue française de gestion industrielle* 27.2, 41–55.

[8] GmbH, S. L. (2021) Creating New Business with Mixed-Mail Automation. Integrating Small Parcel Processing into Mail Processing Centers. url: https://www.siemens-logistics.com/en/news/whitepaper/creating-new-business-with-mixed-mail-automation.

[9] Gower, J. C. (1971) A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27.4, 857. issn: 0006341X. doi: 10.2307/2528823. JSTOR: 2528823. url: https://www.jstor.org/stable/2528823?origin=crossref (visited on 10/24/2022).

[10] McInnes, L., Healy, J., and Melville, J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426 [cs, stat]. url: http://arxiv.org/abs/1802.03426 (visited on 09/22/2022). preprint.

[11] Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020) Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: *International Conference on Image and Signal Processing*. Springer, pp. 317–325.

[12] Pealat, C., Bouleux, G., and Cheutet, V. (2021) Improved Time-Series Clustering with UMAP Dimension Reduction Method. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5658–5665. doi: 10.1109/ICPR48806.2021.9412261.

[13] Hozumi, Y., Wang, R., Yin, C., and Wei, G.-W. (2021) UMAP-Assisted K-means Clustering of Large-Scale SARS-CoV-2 Mutation Datasets. *Computers in Biology and Medicine* 131, 104264. issn: 00104825. doi: 10.1016/j.compbiomed.2021.104264. url: https://linkinghub.elsevier.com/retrieve/pii/S0010482521000585 (visited on 09/22/2022).

[14] Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 160–172. isbn: 978-3-642-37456-2. doi: 10 . 1007 / 978-3-642-37456-2_14.

[15] Costa, E., Papatsouma, I., and Markos, A. (2022) Benchmarking Distance-Based Partitioning Methods for Mixed-Type Data. *Advances in Data Analysis and Classification*. issn: 1862-5347, 1862-5355. doi: 10.1007/s11634-022-00521-7. url: https://link.springer.com/10.1007/s11634-022-00521-7 (visited on 12/02/2022).

[16] Foss, A., Markatou, M., Ray, B., and Heching, A. (2016) A Semiparametric Method for Clustering Mixed Data. Machine Learning 105.3, 419–458. issn: 1573-0565. doi: 10.1007/s10994-016-5575-7. url: https://doi.org/10.1007/s10994-016-5575-7 (visited on 12/02/2022).

[17] Jimeno, J., Roy, M., and Tortora, C. (2021) Clustering Mixed-Type Data: A Benchmark Study on KAMILA and K-Prototypes. In: *Data Analysis and Rationality in a Complex World*. Ed. by T. Chadjipadelis, B. Lausen, A. Markos, T. R. Lee, A. Montanari, and R. Nugent. Studies in Classification, Data Analysis, and Knowledge Organization. Cham: Springer International Publishing, pp. 83–91. isbn: 978-3-030-60104-1. doi: 10.1007/978-3-030-60104-1_10.

[18] Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J. P., Zannad, F., Rossignol, P., and Girerd, N. (2021) Head-to-Head Comparison of Clustering Methods for Heterogeneous Data: A Simulation-Driven Benchmark. *Scientific Reports* 11.1, 4202. issn: 2045-2322. doi: 10.1038/s41598-021-83340-8. url: http://www.nature.com/articles/s41598-021-83340-8 (visited on 12/02/2022).

[19] D'Urso, P., De Giovanni, L., and Vitale, V. (2022) A Robust Method for Clustering Football Players with Mixed Attributes. *Annals of Operations Research*. issn: 0254-5330, 1572-9338. doi: 10.1007/s10479-022-04558-x. url: https://link.springer.com/10.1007/s10479-022-04558-x (visited on 12/02/2022).

[20] Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. (2015) Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data* 10.1, 5:1–5:51. issn: 1556-4681. doi: 10.1145/2733381. url: https://doi.org/10.1145/2733381 (visited on 09/22/2022).

[21] How HDBSCAN Works — Hdbscan 0.8.1 Documentation (2023). url: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html (visited on 01/05/2023).

[22] Parameter Selection for HDBSCAN*— Hdbscan 0.8.1 Documentation (2023). url: https://hdbscan.readthedocs.io/en/latest/parameter_selection.html (visited on 01/05/2023).

[23] Stewart, G. and Al-Khassaweneh, M. (2022) An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences (Switzerland)* 12.5. issn: 2076-3417. doi: 10.3390/app12052405.

[24] Van der Maaten, L. and Hinton, G. (2008) Visualizing Data Using T-SNE. *Journal of machine learning research* 9.11.

[25] Combining Multiple UMAP Models — Umap 0.5 Documentation (2023). url: https://umap-learn.readthedocs.io/en/latest/composing_models.html (visited on 01/04/2023).

[26] Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., and Sander, J. (2014) Density-Based Clustering Validation. In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 839–847. isbn: 978-1-61197-344-0. doi: 10.1137/1.9781611973440.96. url: https://epubs.siam.org/doi/10.1137/1.9781611973440.96 (visited on 10/13/2022).

[27] Hubert, L. and Arabie, P. (1985) Comparing Partitions. *Journal of classification* 2.1, 193–218.

[28] Vinh, N. X., Epps, J., and Bailey, J. (2010) Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *The Journal of Machine Learning Research* 11, 2837–2854. issn: 1532-4435.

[29] Romano, S., Vinh, N. X., Bailey, J., and Verspoor, K. (2016) Adjusting for Chance Clustering Comparison Measures. Journal of Machine Learning Research 17.134, 1–32. issn: 1533-7928. url: http://jmlr.org/papers/v17/15-627.html (visited on 10/24/2022).

[30] Yan, M. (n.d.) Gower. url: https://github.com/wwwjk366/gower.

[31] McInnes, L., Healy, J., and Astels, S. (2017) Hdbscan: Hierarchical Density Based Clustering. *Journal of Open Source Software* 2.11, 205. issn: 2475-9066. doi: 10.21105/joss.00205. url: https://joss.theoj.org/papers/10.21105/joss.00205 (visited on 01/06/2023).

[32] McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018) UMAP: Uniform Manifold Approximation and Projection. The *Journal of Open Source Software* 3.29, 861.

[33] Qiu, W. and Joe, H. (2006) Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification* 23.2, 315–334. issn: 1432-1343. doi: 10.1007/s00357-006-0018-y. url: https://doi.org/10.1007/s00357-006-0018-y (visited on 11/21/2022).

[34] Milligan, G. W. (1985) An Algorithm for Generating Artificial Test Clusters. *Psychometrika* 50.1, 123–127. issn: 1860-0980. doi: 10.1007/BF02294153. url: https://doi.org/10.1007/BF02294153 (visited on 01/06/2023).

[35] Qiu, W. and Joe., H. (2020) clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). manual. url: https://CRAN.R-project.org/package=clusterGeneration.

[36] Schadler, M., Stadlthanner, D., Mayer, B., Schedler, M., and Landschützer, C. (2022) A Method for Pre-Sorting Mixed Mail Using Convolutional Neural Networks and Transfer Learning. In: *24th International Conference on Material Handling, Contructions and Logistics: MHCL 2022*. Faculty of Mechanical Engineering, Belgrade University, pp. 71–80.

[37] Von Luxburg, U. et al. (2010) Clustering Stability: An Overview. *Foundations and Trends® in Machine Learning* 2.3, 235–274.