Investigation of Quality Measures in Cyclists' Dataset Using Dimensionality Reduction Techniques

OLGA SOLOVYEVA , University of Twente, The Netherlands

Multiple external and internal factors could influence the performance of cyclists: heartbeat, age, gender, speed, elevation, wind speed, temperature, and distance, among others. Those factors are essential for planning on how to enhance the overall athlete's performance. However, certain factors could influence the performance more heavily than others. To gain insights into how those factors could intertwine, multidimensional visualization techniques could be useful when exploring visual patterns. In particular, dimensionality reduction techniques may uncover more details on why some athletes perform at a high level, whilst others struggle. With an enormous number of existing dimensionality reduction techniques, this research proposes to find the most qualitative technique in distinct datasets, including a cyclists' dataset. The results show that t-SNE shows outstanding performance in terms of neighborhood and distance preservation and has the potential to be used with clustering algorithms to demonstrate new insights into the cyclists' data. Since dimensionality reduction techniques for cycling data are not well explored by scientific literature, this opens an opportunity for research in this field that could add substantial contributions to those who would be interested in improving cycling behavior.

Additional keywords and phrases: multidimensional visualization, dimensionality reduction techniques, cyclists' dataset.

1. INTRODUCTION

From the moment in time, when a sport becomes more than merely a physical activity, it draws the attention of the audience who enjoys adequate competition and tries to predict the outcome for the sake of entertainment. The enthusiasts explored different ways of collecting the information that could uncover the insights to arrive at the ultimately accurate predictive model. This behavior was adopted by the rest of the people in the industry, who could potentially benefit from it. This would include coaches of the team or an athlete since it is in their interest to transform a mediocre player into a high-performance professional. Furthermore, in the entertainment sector media attempts to give the feeling of involvement to the audience by showing them more behind-the-scenes material. This would keep people on the hook and attract new enthusiasts.

It was proven that a successful analysis of the sports data could drastically increase the team's or athlete's performance [19]. However, the amount of open data for analysis keeps growing, which requires new techniques that could handle the complexity of the analysis. If there's a competition between m number of athletes and n factors that affect their performance, this would result in the m x n matrix. Regularly, data is presented in the form of a table or tabular view [5], where a column represents an attribute, and a row represents the observations on this attribute. Assuming that the numbers for m and n are large, this would result in a time-consuming process of visualizing and analyzing how the attributes in the table are related. An introduced solution is to apply multidimensional visualization, which became a basic tool for handling multidimensional data. It is capable of building a meaningful layout preserving the original data space along with the neighborhood of a data point that could help a user to uncover some insights.

Typically, if there are less than three dimensions, the problem becomes pretty straightforward. The human brain can easily perceive two- or three-dimensional spaces. If the number of dimensions exceeds three, then perception has difficulties imagining and understanding the similarity between points in the projected space. Several techniques could deal with this inconvenience by presenting high-dimensional data in the visual space, which may ease the perception. It includes parallel coordinate plots, glyphs, table lenses, etc.[6] The other approach is to reduce the data to 2D or 3D with the Dimensionality Reduction (DR) techniques [5]. The advantage of DR is its ability to scale much better compared to other techniques in terms of a number of attributes and their observations [6]. In the last couple of years, many new DR techniques were explored and proposed [3,6,17,18,19,24], which leads the user to the question of which one would be the most effective and suitable for a specified dataset. Even though the comparative studies provided a good idea of which techniques are suitable for certain situations and domains, cycling data is not clearly analyzed in the scientific world.

1.1 Objective

The focus of this research paper is to explore and compare possible ways to visualize cycling data, that could discover new insights into how certain attributes are related. It also aims at finding the most effective technique for cluster identification with minimum distortions. The measurement of effectiveness is going to be computed with several quality metrics.

1.2 Research Question

This objective leads to the following research question:

Which dimensionality reduction technique is the most suitable in visualizing distinct patterns that could uncover new insights into the cycling activities?

To give a bit of the organization and a clear understanding of how the research question can be answered, the following sub-questions are used to provide two main steps for the analysis:

- Which dimensionality reduction techniques are suitable to apply to the cycling dataset based on its characteristics?

TScIT 37, July 8, 2022, Enschede, The Netherlands

- What are the quality metrics that could assist in comparing results from applying chosen dimensionality reduction techniques?

Furthermore, Figure 1 provides a step-by-step explanation for answering the research question. The step will have a separate section for detailed clarification. Steps 1 and 2 are going to be discussed in Section 3, where all chosen datasets will be described and their characteristics will be extracted. Steps 3 and 4 will correspond to Section 4, which will provide the explanation for the choices of techniques. A final step is demonstrated in Section 5, where the quality metrics are determined.

2. RELATED WORK

The research was started by exploring the recent developments in sports data visualization. Perin *et al*. [19] discussed various visualization techniques, that are currently being adopted in the field of sports data analysis. This paper made an exceptional contribution by pointing out critical research gaps, those open multiple opportunities for the analysis and exploration. The gaps are related to the taxonomies of sports visualization since it is receiving more attention due to the adaptation of the new sensor technologies.

Several research papers [3,6,17,25] provided a comparative analysis of different dimensionality reduction techniques. Since there is a high variety of choices for a DR technique, those papers helped to significantly reduce the number by consideration of which techniques are widely used, often met in the scientific literature, and which could be applied to a basic high-dimensional dataset by a general user. Specifically, Espadatto *et al*. [6] and Nonato *et al*. [17] provided a solid foundation for the current research in terms of a detailed qualitative analysis of the multidimensional projections on the datasets that differ in characteristics. Additionally, they listed quality metrics that are widely applied in the field of visualization. Furthermore, Xia *et al*. [26] gave insightful information regarding how certain techniques behave in terms of cluster identification and accuracy. They pointed out the fact that techniques that utilize non-linear functions and preserve the local geometry of the point show more accurate results in cluster separation. Thus, it guides give priority to the techniques with certain taxonomies.

3. DATASETS AND THEIR CHARACTERISTICS

The first step to finding suitable **D**imensionality **R**eduction **T**echniques (DRTs) for cycling data is to sample datasets that are relevant to the current study. This is a highly important step since the results drawn from a single dataset cannot be a trustworthy measure of quality. So, to avoid bias and underfitting, more datasets are needed for the experiments. Relevance is evaluated by how similar the original data is to the sample. Since the goal is to draw accurate conclusions on which DRTs are the most effective in clustering the cycling data, testing datasets must be strongly related to it in terms of their characteristics.

3.1 Cyclists' dataset

To understand which datasets are relevant for the experiments, the original cycling dataset must be analyzed to discover its characteristics. The feature description of the original dataset can be found in Appendix A.

The most noticeable aspect is that there is no categorical data, besides gender. Meaning that the focus is going to be on tabular numerical datasets with at most one or two categories. However, those categories will most likely be removed due to the possibility of creating misleading results. The encoding of the original data can have a significant impact on the results of the final mapping [8]. The one-Hot encoder creates a new dimension for each category, which directly affects the computational time of the mapping [8,14]. On the other hand, the ordinal encoding transforms the categories into numbers, creating some sort of relation between the variables, which in reality is misleading for the model [14]. Since there is only a single category present in the original dataset, it is not worth applying an encoder, bearing in mind that it can make an impact on computational time or arrive at an ambiguous conclusion.

Besides the type, datasets can vary in the number of dimensions. Looking back at the feature description, it can be noticed that at most 17 features are present in the dataset. This shows that the focus should be on the samples with a low number of dimensions. However, the size for the samples are going to be smaller for the testing purposes, since the size of the original data exceeds 100,000 entries.

3.2 Choosing datasets

Deciding on the type and number of dimensions already assists with the choice of data samples, since the characteristics of those highly affect the performance of the projection [7]. The experimental datasets were found through the other research papers, that were checked for reliability in advance. This helps to avoid incomplete and irrelevant data. Multiple studies [6,11,24,25], which conducted the research for the evaluation of DRTs, provided information about the datasets, that they used for their experiments. To prevent computational burden, the size of the experimental datasets was restricted to 5000 entries maximum. Based on that, Table 1 presents the chosen samples for the current research.

Table 1. Datasets used for the comparison DRTs

Dataset	Size	Dimensions	Source
Original	138700	17	
Segmentation	2100	19	[23]
US Counties	3028	14	[1]
Dermatology	336	34	[23]
Wine	178	13	[23]
WDBC	569	32	[23]

There are two more factors that are influential to the quality of the projection: intrinsic dimensionality and sparsity [6]. Former demonstrates a number of needed variables to describe the resulting data from the projection. If the number is high, then it is harder to project the data. The original dataset has an intrinsic dimensionality of 0.76, so samples were chosen accordingly. Further, sparsity points to the missing information in the dataset. The lower value demonstrates a high probability of clusters in the high-dimensional space, making it easier for the projection to preserve those clusters. Typically, tabular data is very dense, which is confirmed with a sparsity of 0.14 for the original dataset.

3.2.1 Segmentation

This dataset contains the classification of pixels gathered from the 7 outdoor pictures. It contains a single categorical variable,

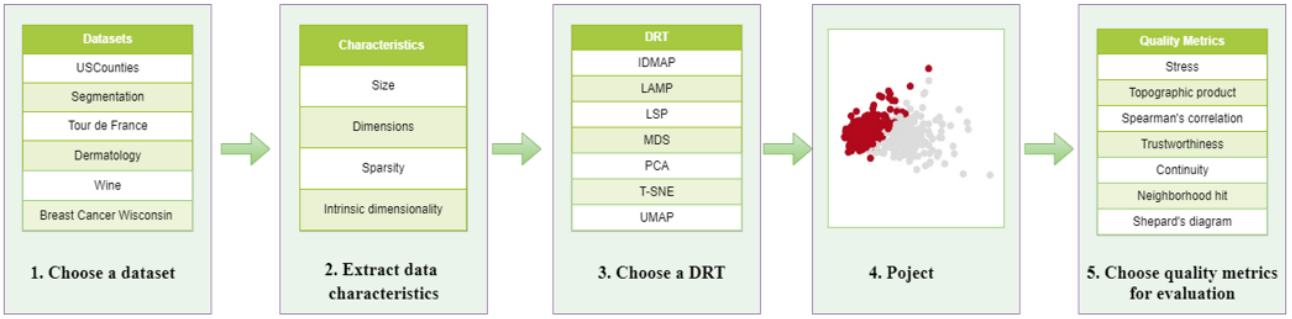


Figure 1: Flowchart of the process

that represents to which picture the pixel belongs. Since the encoding is not used, the feature was removed from the dataset. However, the feature will be used to color the points on the resulting mapping.

3.2.2 US Counties

A sample represents the information about the US counties. It contains a single categorical variable, the name of the county, which was removed, since it would have to be encoded and does not present relevant information to the projections. Also, the ID of the county was eliminated due to the possibility of making a false impression that some counties are more similar than they are.

3.2.3 WDBC

Wisconsin **D**ata of **B**reast **C**ancer contains features that were computed from an image of the fine needle aspirate (FNA) of a breast mass. They provide a numerical description of the cells. A sample contains a single class representing the diagnosis. It was removed as a preprocessing step but will be used to classify the data on the projected space.

3.2.4 Wine & Dermatology

Those datasets are solely numerical. Wine data was used to compare various classifiers, whilst the other contains features that could help to evaluate dermatological diseases.

4. DIMENSIONALITY REDUCTION TECHNIQUES

Due to an enormous number of existing **D**imensionality **R**eduction **T**echniques (DRTs), it is a challenge to choose the ones that will present a qualitative result on a certain dataset. So, several studies [6,7,17,22] compared multiple DRTs and draw insightful conclusions regarding their performance. The DRTs for the current study are going to be chosen based on the academic papers to ensure qualitative results and to avoid a time-consuming trial and error approach. In the following subsections, some taxonomies of the projections are going to be discussed and analyzed.

4.1 Projections' taxonomies

Since applying different DRTs to a single dataset can yield different visual patterns, the choice must be made towards the ones that truly preserve the original information contained by each point such as neighborhood or distance to other points. Among numerous preservation traits, linearity and locality caught the most attention from researchers [25].

4.1.1 Linear vs Non-linear

Mathematically speaking linear transformation satisfies the equation $\Phi(au + bv) = a\Phi(u) + b\Phi(v)$ [22]. It has the advantage of being computationally fast and is believed to outperform in preserving the density of the clusters [4]. It is easier to understand, especially for a not experienced user, since the axes are a linear combination of original dimensions, creating a straightforward and clear relation between high- and low-dimensional spaces. The most well-known [6,25] examples of linear projections are LMNN [6], LPP[10], PCA[12], and FA [12]. Unfortunately, linear techniques cannot handle complex structures without generating distortions, which introduces a major issue in the analysis of the resulting visual space. So, the results would turn out to be poor, if the data actually lies in the non-linear space, meaning that some information would not be visible for linear techniques. Non-linear methods excel in preserving the local structure of the data and, afterward, present it in two-dimensional space. They are advantageous and often used for cluster analysis [26]. There exist cases when classes cannot be linearly separated in the original space, so linear methods would fail at projecting those into well-separated clusters in the low-dimensional visualization. The existing non-linear methods include: UMAP[15], T-SNE [4], IDMAP [6], LAMP [11], Isomap [22], MDS [13], LLE [20], LSP[18] and many more. However, the ones that are mentioned are the most common in the visualization field because of their exceptional performance [6,7,17,24,25]. Most of the methods preserve the class information: making the distance between the same class variables closer and pushing away the ones that belong to different classes.

4.1.2 Local vs Global

Coming from its name, the local approach tends to preserve the local geometry of each point during the mapping process. It depends on two elements: neighborhood and positions of the subset of samples, which were already placed in the visual space. Local methods can be advantageous in terms of computational time since they solely perform sparse matrix computation which is relatively faster, and various studies showed that they performed generally better in terms of cluster identification [25,27]. Specifically, this would be the case for UMAP and T-SNE which performed significantly better than other global techniques with the linearity type [25]. Also, LPP was mentioned in various papers in improving the clustering model [10,27]. Other techniques that also use local embedding and were mentioned in this paper are IDMAP, LAMP, LLE, and LMNN, LSP. The global approach tends to look at the bigger picture, maintaining pairwise distance among all data points. The main advantage of this method is the more trustworthy representation of the original structure to the visual space,

Table 2: Dimensionality Reduction Techniques used in this research

Projection acronym	*Projection full-name*	*Linearity*	*Locality*	*Complexity*	*Implementation*
IDMAP	Interactive document map	Non-linear	Local	$O(N^2)$	Rdimtools
LAMP	Local affine multidimensional projection	Non-linear	Local	$O(Nn)$	MP
LSP	Least Squares Projection	Non-linear	Local	$O(N^3)$	MP
MDS	Multidimensional scaling	Linear	Global	$O(N^3)$	Scikit-learn
PCA	Principal component analysis	Linear	Global	$O(n^3)$	Scikit-learn
T-SNE	t-distributed stochastic neighbor embedding	Non-linear	Local	$O(iN^2)$	Scikit-learn
UMAP	Uniform manifold approximation and projection	Non-linear	Local	$O(iN^2)$	Scikit-learn

which would be beneficial for cluster identification. For example, it was proven that PCA implicitly behaves as a K-means clustering algorithm, since its principal components appear to be similar to the cluster membership indicators [25]. Furthermore, Isomap performs great on the number of tasks where it is required to identify the number of clusters [7]. The rest of the global techniques include MDS and FA.

4.2 Choosing DRTs

After analyzing projections' taxonomies, it is seen that different techniques can provide great results depending on the tasks a user desires. This research is focused on finding techniques that would provide well-separated clusters with minimum distortions. It is expected that local techniques would outperform global if they have the same linearity type. Furthermore, PCA is supposed to show high performance, even though it does not fall into the category of local methods. In regard to linearity, the methods that stand out the most in terms of cluster separation are UMAP and TSNE.

However, before choosing those methods, it is beneficial to look at how they would perform on the tabular dataset. For instance, Espadoto *et al*. [6] provided a benchmark for choosing the DRTs based on the results of the qualitative analysis. By extracting the information about the average results from the benchmark, it is seen that IDMAP, T-SNE, and UMAP have the highest scores out of 44 DRTs used in the research. Nevertheless, the focus should be solely on the tabular dataset, since Espadoto *et al.* experimented with three different types of data: tables, images, and text. By shifting attention to the results of the quality evaluation only for the tabular datasets, the above-mentioned techniques still slightly outperform the rest. The techniques that also show results higher than average are LAMP, MDS, PCA, LSP, Isomap, and FA. The rest of the DRTs that were previously mentioned in the paper, such as LMNN, LPP, and LLE failed to show worthwhile results. Furthermore, Xia *et al*. [25] also mentioned in their experiments that LLE performs significantly poorer in comparison to other techniques used in the research, especially in identifying to which cluster a point belongs. However, they identified that LPP is the 4th preferred technique by the participants of the experiments, where the goal was to correctly identify clusters. Even though LPP also showed great results for precision and recall, it was identified that nonlinear techniques perform better in terms of cluster identification than linear ones if they have the same locality type, so LPP, being local and linear, is outperformed by local and non-linear techniques such as IDMAP and UMAP. Regarding LMNN and FA, the techniques were barely mentioned in the various academic papers used as the foundation of this research, so this would signify that it is not as influential to the field as the rest of the abovementioned DRTs. After gathering those arguments, it was decided not to choose LMNN, LPP, LLE, and FA for the current research.

It is currently clear that IDMAP, T-SNE, and UMAP are strongly preferred for the task of cluster identification. To add more to the list of chosen techniques, Joia *et al.* [11] help to discover the perks of LAMP and LSP methods. Both techniques were applied to the tabular datasets, where they showed improvements in accuracy, computational time, and preserving the distances between the groups. Since PCA and MDS showed high average results and were quite influential in the field of data visualization, they will also be added to the list of chosen techniques. Additionally, PCA is believed to implicitly behave like a clustering algorithm, whilst MDS and Isomap are expected to perform greater than all of the local techniques in the task of identifying clusters [25]. By mentioning Isomap above, the method is expected to show some worthy results when applied to a tabular dataset, so it will be chosen for this research. Table 2 shows the resulting list of techniques that are going to be used in this paper. It provides information regarding their linearity, locality, complexity, and package used for implementation.

4.3 Implementation

Python was chosen as a primary programming language for the experiments since it provides flexibility and eases visualization. Half of the methods were implemented with the library scikit-learn. However, this library contains only the most essential and well-known DRTs. After some research, two packages in the R language were found, that contained the rest of the techniques: Rdimtools [26] and MP [9]. Both packages were integrated into the project using the RPY2(3.5.0) Python library, which presents the interface of using the R language in Python.

5. QUALITY METRICS

Once datasets and DRTs are chosen, a quality assessment can be carried out to find the most suitable technique for the cyclists' data. Several studies provided detailed reviews on existing quality measurements [6,16,17], which formed the basis for the current evaluation. The metrics are split into scalar and pair-point types, which will be introduced in the following subsections.

5.1 Scalar metrics

Scalar metrics are easy to evaluate since they yield a numerical result. Those metrics were chosen based on two factors: the emergence in various scientific papers, showing their high value in the field of visualization, and the complexity of implementation. They are subdivided into two categories: pairwise distance comparison between the original and lower-dimensional spaces; neighborhood analysis of a point in the original and lower-dimensional spaces. Obviously, other scalar metrics exist, however, the research cannot cover every aspect of quality measurements, so only the most common metrics were taken into account.

5.1.1 Pairwise distance comparison

The quality is accessed by analyzing how well the distances between each pair of points are preserved during the mapping from the original to a lower-dimensional space. The input is the original and reduced dissimilarity matrices, which are compared based on the chosen method. In this paper, three metrics are adopted: stress [3], topographic product [2], and spearman's correlation [21].

Stress (Ms): the range of the result is [0,1] with 0 being the best. The equation can be found in [3] and it represents the ratio of Reiman sums measuring the dissimilarity of two points in the original and visual spaces.

Topographic product (Mtp): the range of the result is [0,1] with 0 being the best. The metric relies on dissimilarities between pairs of points in both spaces, however, it computes the product of dissimilarity ratios. Since the topographic product in [2] can be negative, to ensure fair comparison and better overall assessment the final result will be squared to eliminate minus sign. The formula can be found in [2].

Spearman's correlation (Msc): the range of the results is [-1,1] with 1 being the best. The metric measures the strength of the monotonic relationship between two variables. The closer it is to 1, the stronger the correlation between dissimilarities of two variables in the original and visual spaces.

5.1.2 Neighborhood analysis

The quality is accessed by comparing the neighborhood of each point between the original and lower-dimensional spaces. The evaluation is performed by comparing two sets of K nearest neighbors, where K=7, which is chosen in line with [6,16]. Three measurements from [6] are chosen to evaluate this type of quality: trustworthiness, continuity, and neighborhood hit.

Trustworthiness (Mt): the range of the results is [0,1] with 1 being the best. It measures the number of false neighbors of a projected point, meaning the points that appeared to be in the set of K nearest neighbors of the visual space, even though they are not in the set of K nearest neighbors of the original space.

Continuity (Mc): the range of the results is [0,1] with 1 being the best. It measures the number of missing neighbors of a projected point, meaning the points that did not appear in the set of K nearest neighbors of the visual space, even though they are in the set of K nearest neighbors of the visual space.

	US Counties	Segmentation	Dermatology	Wine	BWC	Average
PCA	0.58	0.87	0.62	0.63	0.81	0.71
UMAP	0.61	0.77	0.67	0.71	0.76	0.71
TSNE	0.64	0.82	0.69	0.75	0.82	0.74
LAMP	0.61	0.83	0.63	0.64	0.77	0.69
LSP	0.33	0.49	0.37	0.37	0.46	0.41
MDS	0.63	0.84	0.63	0.68	0.81	0.72
IDMAP	0.41	0.47	0.42	0.41	0.57	0.46
Average	0.54	0.72	0.57	0.61	0.71	

Table 3: Average results for the quality analysis for each dataset and each DRT

Neighborhood hit (Mnh): the range of the results is [0,1] with 1 being the best. It measures points in the set of K nearest neighbors that have the same label as the actual point in the original space. Hence, the method assesses how well the labeled data remain separated after its projection to the lower-dimensional space.

5.2 Point-pair metrics

This measurement helps to notice the details that the scalar metrics cannot capture. For example, two projections can have similar results for the analysis with only scalar assessment, however in reality one may preserve small distances better than the other, making it more effective for the cluster analysis. The technique that will help to uncover those insights is called the **Shepard diagram** [3]. It builds a scatterplot of pairwise distances between all points in the visual and the original spaces. The closer the resulting scatter points to the diagonal the better the preservation of the distances.

5.3 Quality measurement

A formula to measure the quality will be the summation of all scalar metrics multiplied by their weight of importance, which is a ratio of one over six. Since the best value of topographic product and stress is 0, it will be converted to have the best value of 1 by subtracting the result from 1. The equation:

$$\mu = \frac{1}{6}(Mnh + Mc + Mt + Msc + (1 - Mtp) + (1 - Ms)$$

6. RESULTS

From the average results of Table 3, it is seen that LSP and IDMAP perform utterly poorly in comparison to the rest of the techniques that showed quite acceptable results. It is also worth noticing that for some datasets (US Counties, Dermatology, and Wine), all DRTs demonstrated low scores. The characteristic that makes those datasets different is a relatively high intrinsic dimensionality in comparison to the rest of the samples. The same observation was noted by Espadato *et al*. [6], mentioning the high correlation between intrinsic dimensionality and the optimal quality values.

Before diving into the detailed analysis of each technique, it is important to notice the information in Table 4 and Table 5. The former carries the quality values based solely on pairwise distance comparison (values from stress, topographic product,

	US Counties	Segmentation	Dermatology	Wine	BWC	Average
PCA	0.51	0.91	0.58	0.59	0.72	0.66
UMAP	0.49	0.57	0.64	0.72	0.58	0.61
TSNE	0.56	0.66	0.67	0.81	0.69	0.68
LAMP	0.55	0.78	0.59	0.62	0.61	0.63
LSP	0.09	0.13	0.16	0.21	0.16	0.15
MDS	0.61	0.79	0.59	0.69	0.68	0.67
IDMAP	0.26	0.39	0.32	0.31	0.37	0.33
Average	0.44	0.61	0.51	0.56	0.54	

Table 4: Results solely for the pairwise distance comparison over all datasets and all DRTs

	US Counties	Segmentation	Dermatology	Wine	BCW	Average
PCA	0.65	0.83	0.67	0.67	0.89	0.74
UMAP	0.72	0.97	0.72	0.69	0.95	0.81
TSNE	0.72	0.97	0.71	0.71	0.96	0.82
LAMP	0.67	0.88	0.67	0.66	0.94	0.76
LSP	0.57	0.85	0.58	0.53	0.76	0.66
MDS	0.66	0.91	0.69	0.67	0.93	0.77
IDMAP	0.54	0.55	0.52	0.49	0.76	0.58
Average	0.65	0.85	0.65	0.63	0.88	

Table 5: Results solely for the neighborhood analysis over all datasets and all DRTs

and spearman's correlation). The latter contains values of the neighborhood analysis (trustworthiness, continuity, and neighborhood hit). Both tables will assist with finding the reason why certain techniques showed better or poorer results. Shepard's diagrams and projections can be found in Appendix A.

6.1 LSP

As it can be seen from Table 3, LSP demonstrated the lowest average result, which is due to unacceptably high values of stress and topographic product. However, it should be noted that it is common for stress results to be incompatible with visual layouts [18]. This means that it is possible for a projection with a higher stress value to better preserve a cluster separation rather than a projection with a lower stress value. This would apply to LSP since it shows one of the highest results for neighborhood hit over all of the datasets. LSP takes as an input control points, which are supposed to represent significant information about the dataset and to be projected by MDS. After the control points are projected, LSP interpolates the rest of the points around their neighbors, keeping local geometry. Shepard diagram precisely shows how MDS projects the control points by demonstrating some distance points to appear close to the main diagonal, while points that are located in the straight line are the ones that were projected using LSP. It was noticed that LSP indeed shows decent cluster separation, however only for datasets with more than 300 entries. The reason is connected to a poor choice of control points: either selecting the points that contain bias or missing an important representative. So, LSP would work best as a projection for a high-precision approximation of large datasets. Computationally costly techniques could represent solely control points and LSP will accurately project the rest of the sample around the dataset's representatives.

6.2 IDMAP

From looking at all three Tables (3,4,5), it is clear that IDMAP showed low results among all sectors. Furthermore, Shepard diagrams indicate the distortion of the distances after projection.IDMAP is a technique for generating maps of documents aiming at placing similar ones in the same neighborhood. Since the data, presented in the current research is not textual, it was shown that IDMAP could not cope with the tabular data. A technique showed a higher performance only for the BWC sample. It managed to demonstrate a mild cluster separation; however, it can be explained by a low intrinsic dimensionality of the sample (lowest in comparison to the rest of the samples). So, if there are no more than 2 dimensions needed for presenting the data, then IDMAP can mildly preserve the distances and cluster between classes. However, it would be advisable to use this technique on more complex tabular samples.

6.3 UMAP

UMAP is the fifth-best or the third-worst technique based on the average result. Since it is similar to TSNE, it was expected to achieve as high results. However, it can be seen from Table 3, that UMAP is worse at preserving the pairwise distance between points. Shepard diagram is evidence of this statement. For example, by looking at Shepard diagrams for Wine and Dermatology samples, a user can predict how many clusters are expected by counting the number of groups that are horizontally separated on the diagram. UMAP excels at preserving the topology of the initial dataset, which can be proved by noticing the second-highest result in Table 4. However, it should be noted that if the goal is metric structure preservation, then another technique should be chosen for the specified intention since UMAP does not prioritize the global structure. For example, if the original data contained a dense structure in one part and a loose in another, then UMAP would attempt to put these two local parts on an even footing. So, UMAP seeks a manifold on which the data is distributed. From observing the projected samples, it is seen that the clusters are accurately separated, however, the metric structure is not preserved.

6.4 PCA

PCA demonstrated adequate results among all specters: average performance, pairwise distance preservation, and neighborhood analysis. The main advantages of PCA over other techniques is a computational speed and easy implementation. So, for the user, who is looking for a fast and accurate approximation of projected data, PCA would be a perfect technique. However, the linearity of PCA cannot assure a truthful representation of the data. For example, PCA seemed to experience an information loss for datasets USCounties, Dermatology and Wine. Those samples have high intrinsic dimensionality of 0.63, 0.47, and 0.77 with the number of dimensions being 16,34, and 13

respectively. This means that each of the samples would need at least 10 dimensions for an accurate representation. PCA did not manage to preserve the pairwise distance well for those samples. The number of principal components must be carefully chosen as well as which ones are more significant. It is an essential task since PCA maximizes the information in the first two components. It can be the case that during the projection of, for example, a Wine sample, the data did not lay in the linear space, so chosen principal components could not yield a clustered projection.

6.5 LAMP

LAMP is, technically, the third-best technique in the current research, however, it showed almost identical results to PCA. Nevertheless, it dealt slightly better with the datasets that had a high intrinsic dimensionality, which is due to the non-linear essence. It demonstrated the second-highest results in spearman's correlation metric, which shows that there is a high correlation between the pairwise distances in the original and visual spaces. So, if the distance between pair of points is increasing in the original space, then the same tendency will take place in the projection. Furthermore, it excels at minimizing stress values. LAMP could have been the number one technique for preserving pairwise distances, however, the topographic product shows that a local neighborhood of each point changed internally. So, clusters in the original space were kept during mapping, but the internal structure was slightly changed, which reduced the score for the topographic product. However, LAMP did not demonstrate a distinct cluster separation for most of the samples. This could be fixed by choosing better representative control points. So, it follows, that LAMP is a remarkable technique for preserving a local geometry of points, but to show a distinct cluster separation, a user must identify accurate control points, that could represent a piece of valuable information in the dataset.

6.6 MDS

MDS is a second-best technique solely based on average results, and the first best on preserving pairwise distance between points in original and visual spaces. Shepard diagrams demonstrate an almost 45-degree straight line, confirming its greatness at preservation of distances. The biggest burden is its computational complexity. For some large datasets (TourDeFrance, USCounties, and Segmentation), it could take twice or three times as much as for the rest of the techniques. So, to speed up the process MDS can be used in combination with the other techniques. For example, MDS can project a smaller subset of the sample, which can as input of control points for LAMP or LSP. On top of that, it also assures a high accuracy for techniques previously-mentioned techniques. However, MDS did not show distinct cluster separation. It preserves the local neighboring of a point but does not demonstrate clusters if they were not there in the first place. MDS can also be used in combination with a clustering algorithm to ensure the appearance of distinct clusters. After MDS projected the data, a clustering algorithm can be used on top to see if there are any groups in the data. Nevertheless, it would seem complex for a user, who wants to quickly see if the data contains any clusters. So, MDS would mostly be used only if accuracy is the number one priority.

Method	Stress	Topographic Product	Spearman	Trustworthiness	Continuity	Neighborhood hit	Result
LAMP	0.8001	-0.181964	0.27135	0.7271	0.84336	0.441971	0.55
T-SNE	0.23842	-1.21815	0.4167	0.99284	0.98935	0.8227857	0.664
PCA	0.145098	-2.52808	0.68222	0.81343164	0.97363	0.35267	0.613

Table 6: Metrics results for original dataset

6.7 T-SNE

T-SNE demonstrated the highest average result. It is the best at neighborhood analysis and the second-best in pairwise distance preservation. The strongest aspect of T-SNE is the fast computation of projected points and cluster identification. However, from the deeper analysis, it is seen, that the correlation between original and projected points is quite low, which can be noticed from Shepard's diagrams. It is because T-SNE does not exactly preserve distances but estimates a probability distribution. To achieve a higher correlation, tuning of parameters is required. Nevertheless, it still shows higher results for distance preservation in comparison to the rest of the techniques, besides MDS. A feature that makes T-SNE stand out from MDS is a distinct cluster separation. Overall, it shows outstanding results for all the datasets in comparison to the rest of the techniques.

7. CASE STUDY: t-SNE ON CYCLISTS' DATASET

As the results above have shown, T-SNE is the most suitable technique for the tabular dataset. In this section, T-SNE will be applied to the original cycling dataset with the purpose of investigating new insights into the data. Furthermore, two more techniques will be applied as well for a comparison: LAMP and PCA. Even though MDS is the second most suitable technique, the size of the original dataset is large, which presents enormous computational complexity in terms of memory allocation. As it was mentioned earlier, MDS can be used in conjunction with a different technique. However, it is outside of the scope of this research. So, the experiments were made with only three techniques that showed the best results during the analysis: T-SNE, LAMP, and PCA. The resulting projections can be found in Appendix A. Each technique has three projections with different color mapping based on the attributes from the data: height, weight, and age. The idea was to identify which feature represents a certain cluster.

Table 6 presents the results of quality analysis from applying three dimensionality reduction techniques to the cycling data. The color of the cell indicates how great the score is in comparison to the rest of the scores. So the green color presents an excellent score, whilst red shows poor performance. It can be seen that t-SNE had quite average results for the first three metrics and outstanding performance for neighborhood preservation. However, by looking at the projection itself, it is quite difficult to identify clusters. It seems that t-SNE produced a significant number of clusters, which collided into one colossal group. On the other hand, LAMP and PCA demonstrated a relatively better cluster segmentation, for example showing that middle-aged and tall athletes show similar performance. However, the reliability of this result is questionable, since both techniques show low results in some of the metrics. So bearing in mind the reliability of t-SNE, applying a clustering algorithm on top of the resulting projection might help with finding new insights into the cycling dataset.

8. CONCLUSION AND FURTHER WORK

In this paper, an evaluation of seven dimensionality reduction techniques was presented to find the most effective technique for a cyclists' dataset based on the quality metrics that aimed to assess the pairwise distance and neighborhood of a point in the original and visual spaces. From the analysis of the results, t-SNE has the highest quality and would be a preferred technique for the visualization of the cyclists' data. Even though it demonstrates outstanding performance, the visual layout needs a clearer cluster representation. So, it is advisable to apply a clustering algorithm on top of the resulting visualization. This has the potential of explicitly demonstrating new insights into the cycling data. Furthermore, to increase the accuracy of the pairwise distance a tuning of parameters will be necessary. Since t-SNE is a non-deterministic technique, it produces slightly different layouts each time it tries to project data. Further research could be done to explore if there is an approach to making a deterministic t-SNE. Research can also be extended to finding an optimal tuning of parameters and examining which parameters affect layout more.

More interesting insights were noted during the research. MDS achieved almost as high results as t-SNE, however it presents a computational burden, especially if the size of the dataset is quite large (e.g., more than 2000 entries). A proposition could be to use MDS in combination with techniques (e.g., LSP or LAMP), that take control points as parameters. MDS will present the control points and, for example, LSP will project the rest of the data around its neighbors. Even though LSP showed an utterly low result, it is mostly due to stress minimization and topographic product. However, LSP showed one of the highest results for the neighborhood hit, so it is expected that this technique will manage to project the rest of the points around its representatives. Further research can be done to explore the feasibility of this idea. Another concept that could be tested involves the application of clustering algorithms on projected data from dimensionality reduction techniques. There are multiple DRTs (such as PCA and LAMP), that have a high score in preserving the distance between points, but they lack clustering separation. So, a clustering algorithm may uncover some patterns in data.

REFERENCES

[1] *Application Examples of the Hierarchical Clustering Explorer*. Application examples of the Hierarchical Clustering Explorer. (n.d.). Retrieved June 17, 2022, from http://www.cs.umd.edu/hcil/hce/examples/application_examples.html

[2]: Bauer, H.-U., & Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, *3*(4), 570–579. https://doi.org/10.1109/72.143371

[3]: Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, *17*(2), 444–472. https://doi.org/10.1198/106186008x318440

[4]: Chatzimparmpas, A., Martins, R. M., & Kerren, A. (2020). T-visne: Interactive assessment and interpretation of T-Sne Projections. *IEEE Transactions on Visualization and Computer Graphics*, *26*(8), 2696–2714. https://doi.org/10.1109/tvcg.2020.2986996

[5]: dos Santos, S., & Brodlie, K. (2004). Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, *28*(3), 311–325. https://doi.org/10.1016/j.cag.2004.03.013

[6] Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S., & Telea, A. C. (2021). Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, *27*(3), 2153–2173. https://doi.org/10.1109/tvcg.2019.2944182

[7] Etemadpour, R., Motta, R., Paiva, J. G., Minghim, R., de Oliveira, M. C., & Linsen, L. (2015). Perception-based evaluation of projection methods for Multidimensional Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *21*(1), 81–94. https://doi.org/10.1109/tvcg.2014.2330617

[8] Fitkov-Norris, E., Vahid, S., & Hand, C. (2012). Evaluating the impact of categorical data encoding and scaling on neural network classification performance: The case of repeat consumption of identical cultural goods. *Engineering Applications of Neural Networks*, 343–352. https://doi.org/10.1007/978-3-642-32909-8_35

[9]: Francisco M. Fatore, S. G. F. (2019, May 1). *MP: Multidimensional projection techniques in MP: Multidimensional projection techniques*. mp: Multidimensional Projection Techniques in mp: Multidimensional Projection Techniques. Retrieved June 20, 2022, from https://rdrr.io/cran/mp/man/mp.html

[10]: He, X., & Niyogi, P. (2005). In *Locality preserving projections*.

[11] Joia, P., Paulovich, F. V., Coimbra, D., Cuminato, J. A., & Nonato, L. G. (2011). Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2563–2571. https://doi.org/10.1109/tvcg.2011.220

[12]: Jolliffe, I. T. (1986). Principal component analysis and Factor Analysis. *Principal Component Analysis*, 115–128. https://doi.org/10.1007/978-1-4757-1904-8_7

[13]: Kruskal, J., & Wish, M. (1978). Multidimensional scaling, *11*. https://doi.org/10.4135/9781412985130

[14]: Kunanbayev, K., Temirbek, I., & Zollanvari, A. (2021). Complex encoding. *2021 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn52387.2021.9534094

[15]: McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

[16]: Mokbel, B., Lueks, W., Gisbrecht, A., & Hammer, B. (2013). Visualizing the quality of Dimensionality Reduction. *Neurocomputing*, *112*, 109–123. https://doi.org/10.1016/j.neucom.2012.11.046

[17] Nonato, L. G., & Aupetit, M. (2019). Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, *25*(8), 2650–2673. https://doi.org/10.1109/tvcg.2018.2846735

[18]: Paulovich, F. V., Nonato, L. G., Minghim, R., & Levkowitz, H. (2008). Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, *14*(3), 564–575. https://doi.org/10.1109/tvcg.2007.70443

[19]: Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., & Carpendale, S. (2018). State of the art of sports data visualization. *Computer Graphics Forum*, *37*(3), 663–686. https://doi.org/10.1111/cgf.13447

[20]: Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

[21]: Siegel, S., & Castellan, N. J. (2003). *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill.

[22]: Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

[23] UCI Machine Learning Repository. (n.d.). Retrieved June 17, 2022, from https://archive.ics.uci.edu/ml/index.php

[24] Ventocilla, E., & Riveiro, M. (2020). A comparative user study of visualization techniques for cluster analysis of Multidimensional Data Sets. *Information Visualization*, *19*(4), 318–338. https://doi.org/10.1177/1473871620922166

[25] Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., & Liu, S. (2022). Revisiting dimensionality reduction techniques for Visual Cluster Analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 529–539. https://doi.org/10.1109/tvcg.2021.3114694

[26]: You, K. (2020, May 22). *Rdimtools: An R package for dimension reduction and intrinsic dimension estimation*. arXiv.org. Retrieved June 20, 2022, from https://arxiv.org/abs/2005.11107

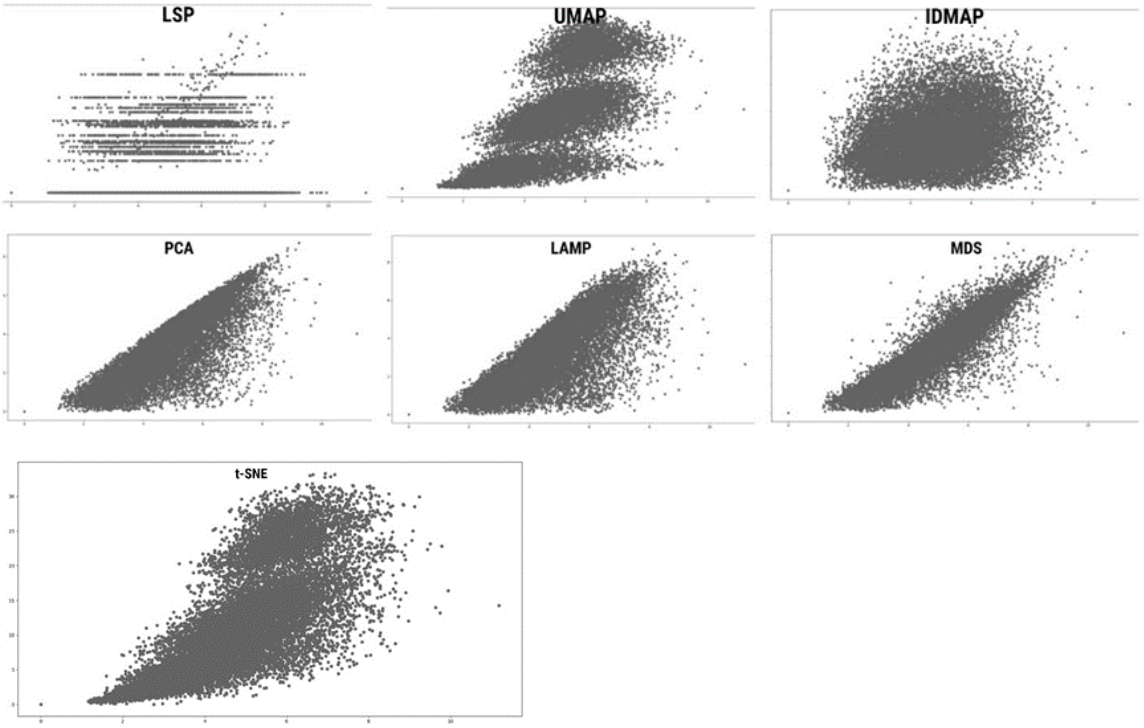
[27]: Zhan, M., Lu, G., Wen, G., Zhang, L., & Wu, L. (2019). A clustering algorithm via kernel function and locality preserving projections. *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. https://doi.org/10.1109/ssci44817.2019.9002683

Appendix A

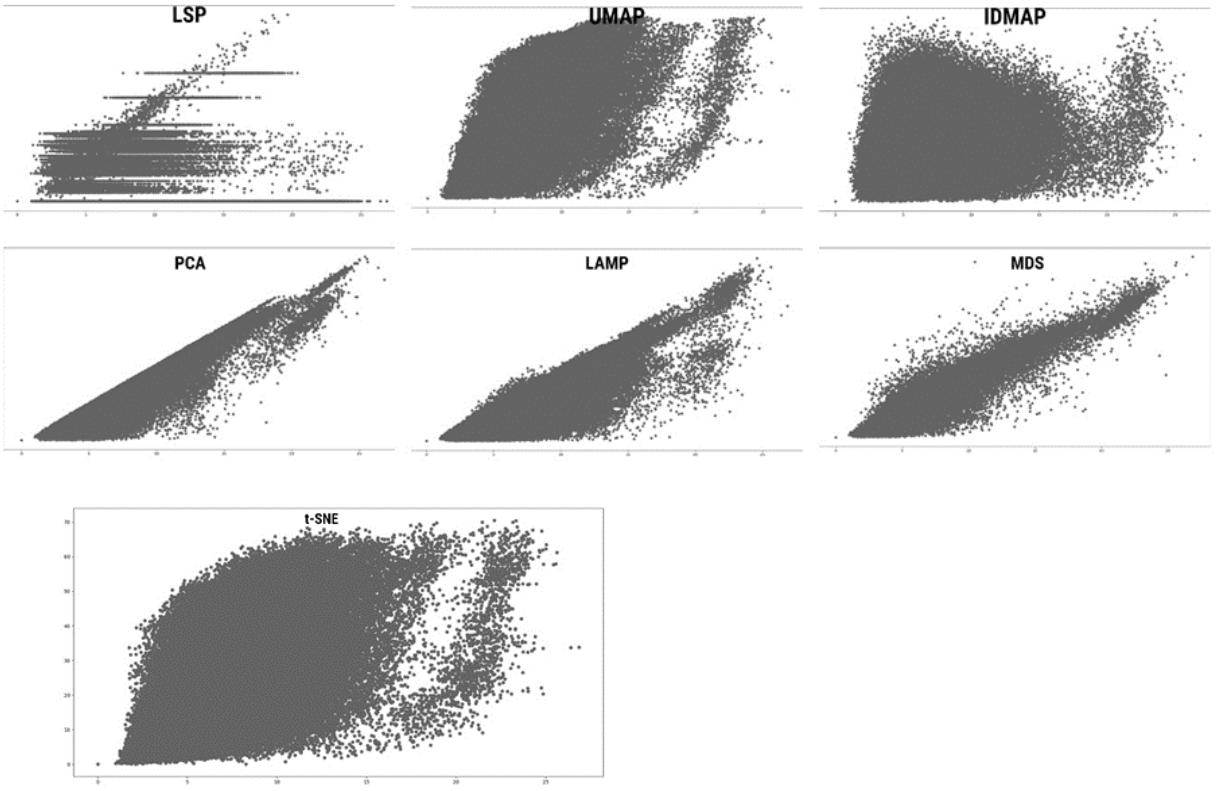
Table 1. Features description of both data sets us

id	Ride Features	Range/Values
1	Gender	Male, Female
2	Weight (kg)	[45,136]
3	Height (cm)	[152,208]
4	Age	[16,77]
5	Total Distance (km)	[5,601]
6	Total time move (secs) (TTM)	[0,123000]
7	Total time stop (secs)	[0, <TTM]
8	Average Speed (km/h)	[10,51]
9	Uphill (meters)	[0,9000]
10	Downhill (meters)	[0,9000]
11	Average Temperature (C)	[-10,58]
12	number of slopes	[0,500000]
13	Average Heart Rate Zone	[>1, 9]
14	avgHeartRateZonePerUser	[1.88, 9]
15	speedHeartRateRatio	[1.7, 14]
16	elevationPerDistanceRatio	[1.39, 50.42]
17	timeRatio	[1.11,14121]

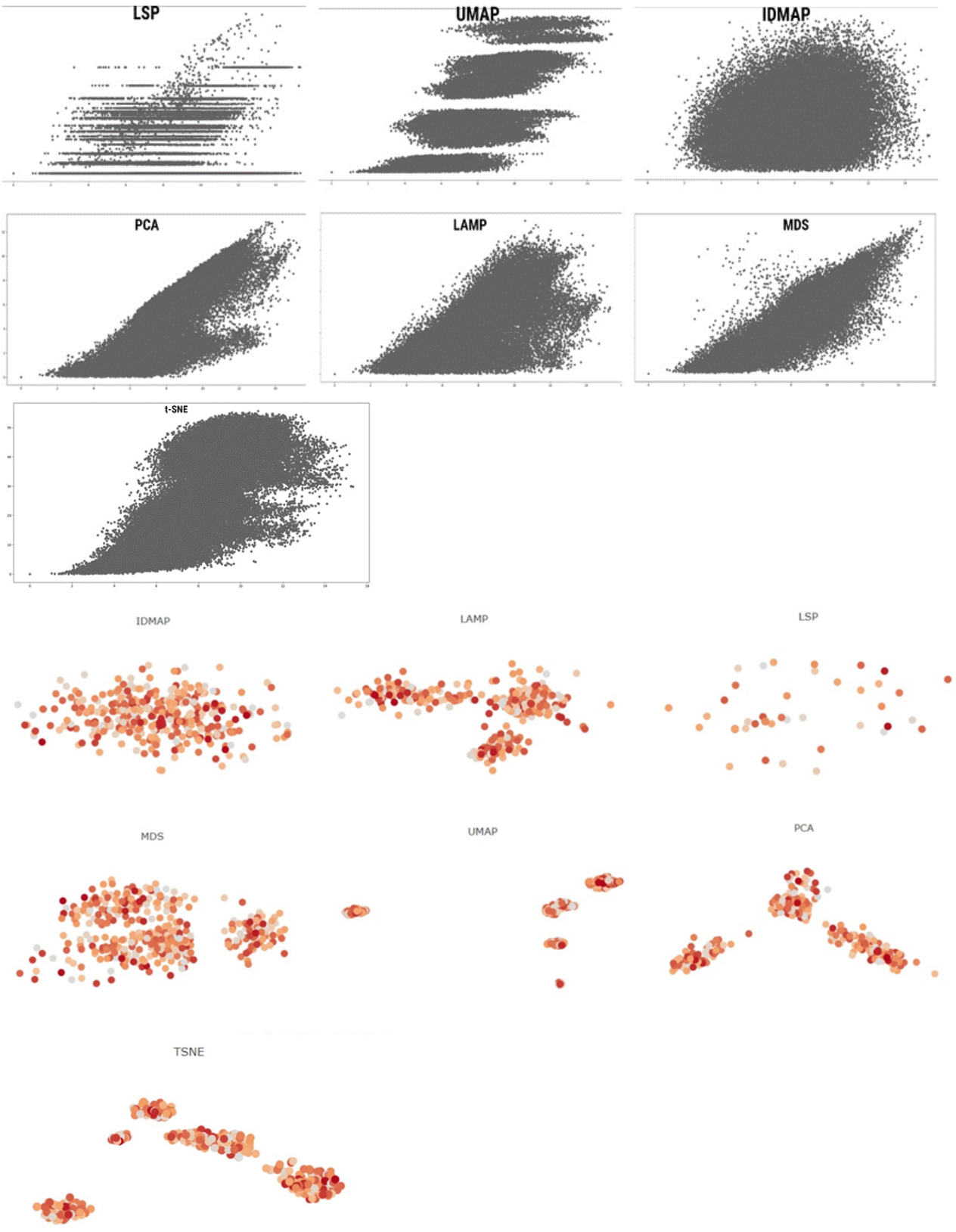
2. Shepard's diagrams and projections for Wine dataset.



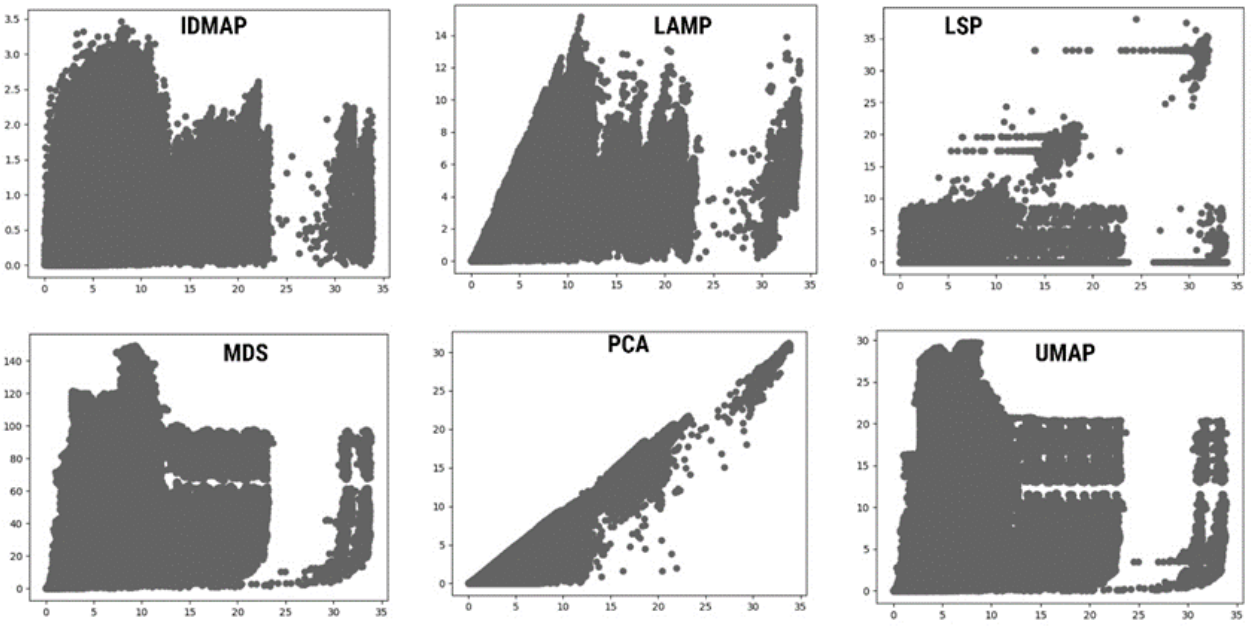
3. Shepard's diagrams and projections for the BCW dataset.

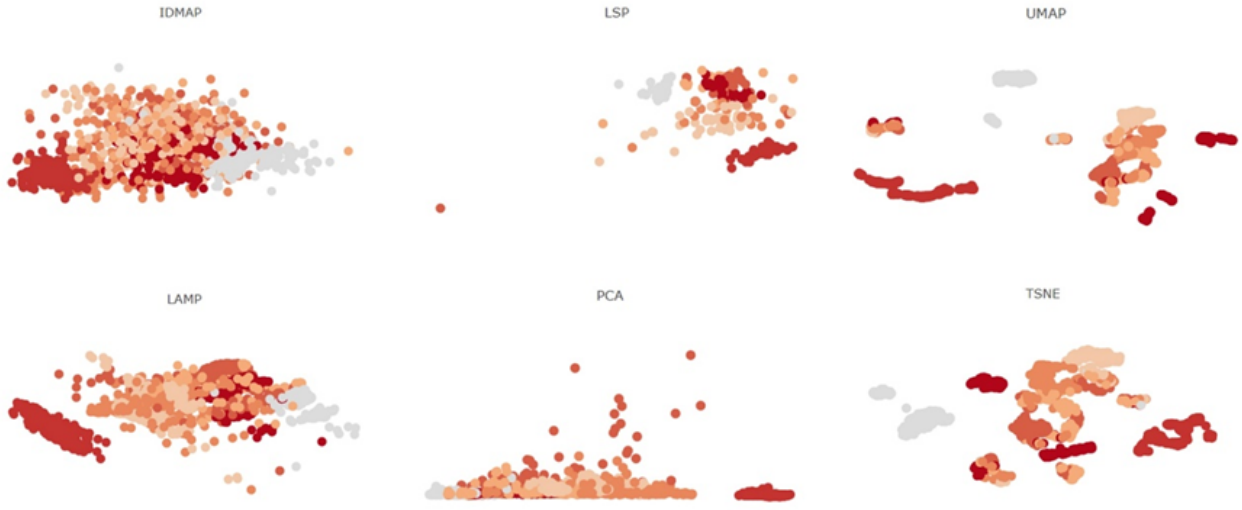


4. Shepard's diagrams and projections for the Dermatology dataset.



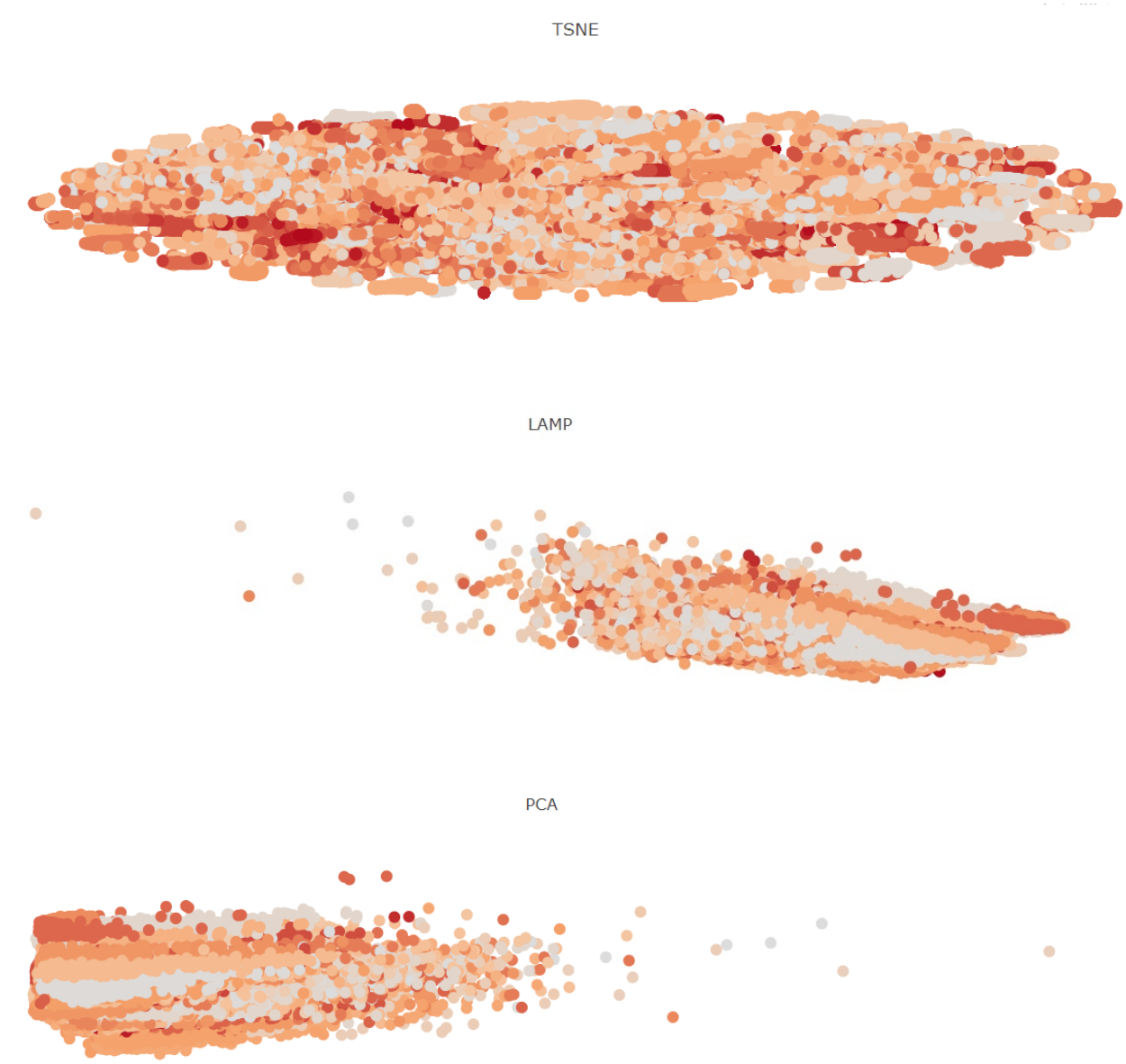
5. Shepard's diagrams and projections for the Segmentation dataset.



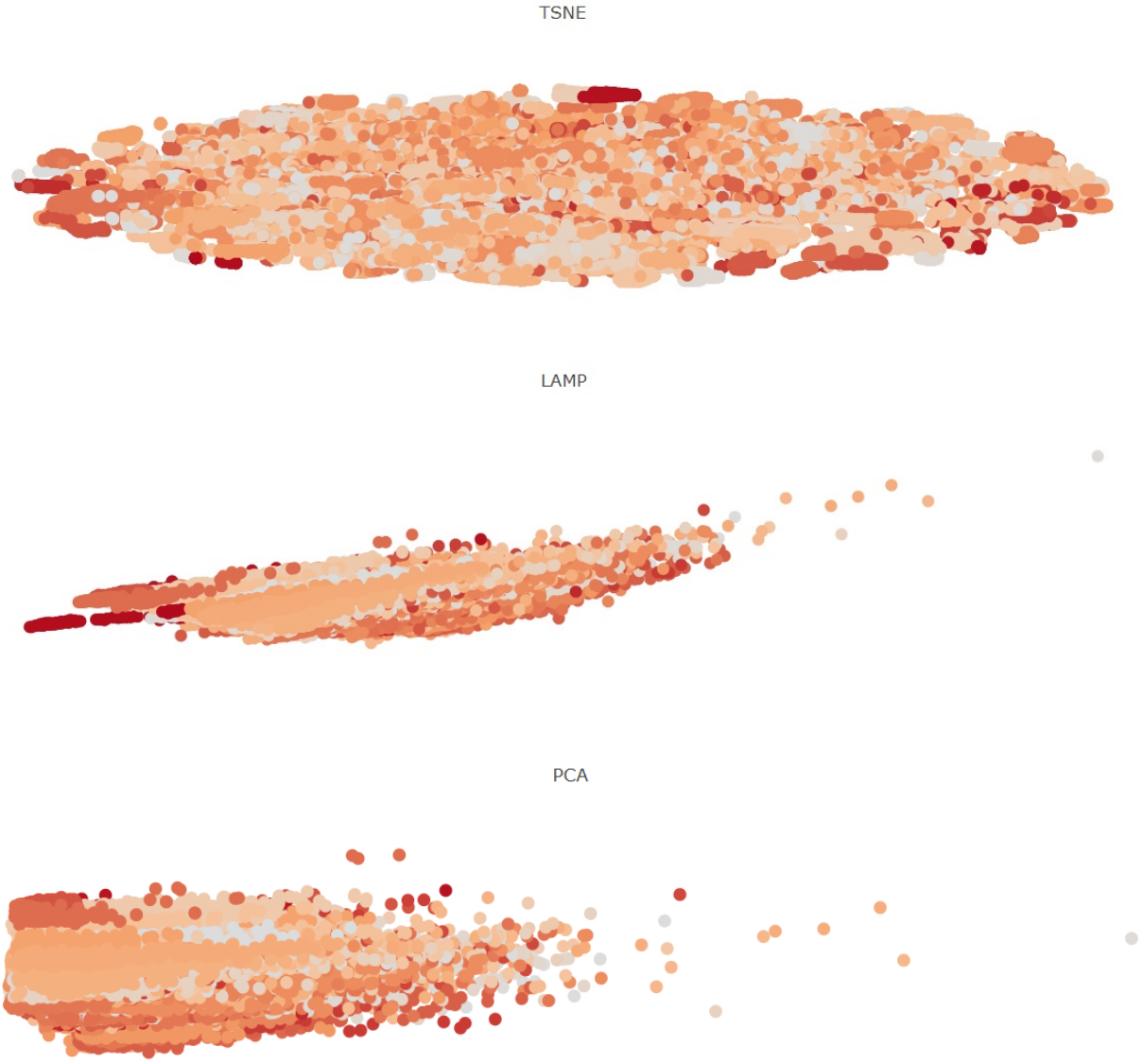


6. CASE STUDY RESULTS: Cyclists' dataset

AGE:



WEIGHT:



LENGTH: