xHiPP: eXtended Hierarchical Point Placement Strategy

Fábio Felix Dias, Rosane Minghim
Instituto de Ciências Matemáticas e de Computação
University of São Paulo
Av. Trabalhador São-Carlense, 400, São Carlos, SP, Brazil
Email: f_diasfabio@usp.br, rminghim@icmc.usp.br

***Abstract*—The complexity and size of data have created challenges to data analysis. Although point placement strategies have gained popularity in the last decade to yield a global view of multidimensional datasets, few attempts have been made to improve visual scalability and offer multilevel exploration in the context of multidimensional projections and point placement strategies. Such approaches can be helpful in improving the analysis capability both by organizing visual spaces and allowing meaningful partitions of larger datasets. In this paper, we extend the Hierarchy Point Placement (HiPP), a strategy for multi-level point placement, in order to enhance its analytical capabilities and flexibility to handle current trends in visual data science. We have provided several combinations of clustering methods and projection approaches to represent and visualize datasets; added a choice to invert the original processing order from cluster-projection to projection-cluster; proposed a better way to initialize the partitions, and added ways to summarize image, audio, text and general data groups. The tool's code is made available online. In this article, we present the new tool (named xHiPP) and provide validation through case studies with simpler and more complex datasets (text and audio) to illustrate that the capabilities afforded by the extensions have managed to provide analysts with the ability to quickly gain insight and adjust the processing pipeline to their needs.**

I. Introduction

Databases have steadily increased both in size and complexity, as a result of enhancement in data collecting and detection. There are several different sources of data in categories such as business, commerce, social network, biology, climate, image data and many others with valuable and strategic potential. These data have complex relations and structure as well as representation with varying numbers of attributes. Consequently, analyzing current datasets with the aim to extract meaningful and useful information has become a challenge.

Applying visualization techniques to inspect data is suitable to support data analyses and has been an integral part of data science activities. *Multidimensional Projections* are techniques frequently used in exploratory situations and they map data from an *n*-dimensional space into a *2*-dimensional or *3*-dimensional space. The resulting projections represent data instances as an individual graphical entity (such as small circles) and are thus sometimes called point placement strategies. Multidimensional Projections aim to reduce information loss by conserving, in projected space, properties or relationships such as proximity or neighborhood from the original (or attribute) space.

The application of specific projections depends on the type of data and tasks that analysts wish to perform, as well as the goal of the analysis [1]. There are several projections techniques, all of them with their advantages and disadvantages, as will be argued in the next section. However, while projections have greatly improved in processing time and in handling large attribute sets, one particularly difficult feature of projections is the management of visual spaces. Layouts become cluttered very fast, and groups of points that would in principle be distinguishable by the projection algorithm can become mixed. One example of point placement approach that handles visual scalability is the *Hierarchical Point Placement* (HiPP), which is capable of presenting several levels of detail of a dataset represented by its attribute set or by a similarity relationship. It eases exploration and visually organizes datasets of many different sizes under the same perspective [2]. It does so by integrating clustering and projection into the point placement algorithm. It is the only point placement strategy to date that handles these two features (data partition and relief of visual clutter). Its original formulation, however, is not very flexible in terms of adjusting to different clustering strategies and to different projection strategies. Additionally, the original tool is designed for handling interaction with text collections, while other multidimensional data can and should take advantage of its design.

This paper presents a new design for the HiPP method (called xHiPP) with the aim of improving its exploration capabilities and increasing its flexibility. The improvements allow better data representation and exploration, adapting to different analysis needs. As well as showing the new features and examples for data of varied nature, in this paper, a case study is presented, showing the application of xHiPP to acoustic data analysis, highlighting its capabilities to reflect global and local characteristics of the data under analysis.

This paper is organized as follows. Section II shows a representative sample of projection techniques and briefly reviews relevant concepts used in this work. Section III describes xHiPP and its methods. Section IV reports on experimental results obtained with the application of xHipp. Section V presents a discussion of the results. Section VI concludes the paper with some final remarks and directions for future work.

II. Background and Related Work

A. Multidimensional Projections

Most projection methods have been proposed to visualize high dimensional datasets maintaining the data neighborhood relationships. This goal is not always reached because of the great number of attributes on the original space [3] as well as the reduced space on a computer interface.

It is possible to list projection techniques with somewhat distinct goals, such as *Multidimensional Scaling* (MDS) [4] that attempts to project data preserving distance relationships from the original space into the projected space. The *Force Scheme* [5] tries to lay out points based on geometric position and vector force displacement created between points. Distance reconstruction is also targeted here. The *t-Stochastic Neighbor Embedding* (t-SNE) [6] maps data similarities in original space to conditional probabilities and attempt to minimize some divergence function between probabilities. The *Least Square Projection* (LSP) [7] starts projecting control points (a sub-set of points) with *Force Scheme* technique and map the remaining data with a Laplace operator based on the control points first projected. The target for LSP is to reconstruct neighborhoods from the original space in the projected space. *Local Affine Multidimensional Projection* (LAMP) [8] places control points and builds a set of orthogonal affine mappings which follow the control points' layout. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) [9], [10], can also be used for projection purposes.

Projections that use multilevel approaches aid data exploration in several abstraction levels. Glimmer [11] is a global technique that projects a multidimensional data applying a multilevel approach, associated with GPU parallelism to improve the speed of computation. This technique organizes data into a hierarchical structure and recursively applies a strategy to combine and refine the levels. Authors reported that Glimmer had improvements in speed levels, numeric measurements, and visual quality when applied to synthetic and real datasets. The final visualization of Glimmer, though, is done in a single level.

B. Evaluation of Multidimensional Projections

In order to evaluate projections numerically, some measures can be applied. The *Stress function*, shown in Equation 1, computes the information loss of a projection process [12]. The range of values returned vary between 0 (lesser loss) and 1 (greater loss).

$$stress = \sqrt{\frac{\sum_{i<j}(\delta(\mathbf{y}_i, \mathbf{y}_j) - d(\mathbf{x}_i, \mathbf{x}_j))^2}{\sum_{i<j}\delta(\mathbf{y}_i, \mathbf{y}_j)^2}} \tag{1}$$

being:

- $\mathbf{x}_i$, the original space vector,
- $\mathbf{y}_i$, the projected space vector,
- d, the original space similarity function,
- δ, the projected space similarity function.

The *Silhouette Coefficient* [13], that measures cohesion and separation of clusters, is normally applied in the evaluation of cluster algorithm results. It can also be applied to assess projections of labeled data. This measure is calculated for each data point with Equation 5 and a value of the complete dataset is obtained with the average of all point silhouette values. Silhouette values measure data cohesion (Equation 2) and separation (Equations 3 and 4), and vary between -1 (fully overlapped groups) and 1 (fully separated groups).

$$a(\mathbf{x}_i) = \frac{1}{N_A - 1}\sum_{\substack{i \neq j \\ x_j \in A}} d(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

being:

- A is the $\mathbf{x}_i$ group,
- N_A is the quantity of items in A,
- d is a similarity function.

$$D(\mathbf{x}_i, C) = \frac{1}{N_C}\sum_{\substack{x_j \in C \\ C \neq A}} d(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

$$b(\mathbf{x}_i) = \min_{\forall C \neq A}\{D(\mathbf{x}_i, C)\} \tag{4}$$

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}} \tag{5}$$

Some other functions can be used to evaluate projections, such as *Neighborhood Hit* and *Neighborhood Preservation* that measure for each point, respectively, the proportion of point neighbors in projected space that belongs to the same class [7], and the proportion of neighbors in original space that remain neighbors in projected space [2]. Both of them vary between 0 and 1 (best projection result) and the final result for the projection is an average for all points.

C. Hierarchical Point Placement Strategy

Hierarchical Point Placement Strategy (HiPP) attempts to conserve data characteristics as clustering and segregation [2]. The structure generated by this technique affords exploration capability in several levels of detail. This ability enhances the analysis and allows the use of larger data collections than most projections. To layout items on a plane, HiPP follows takes three steps: *a)* building of a data hierarchy structure, *b)* projection of the data hierarchy and *c)* group spreading and removal of data overlap. HiPP authors highlighted that the computational complexity of the approach is $O(n\sqrt{n})$, being n the total number of data points.

Initially, a tree with the root node having all data instances as children is created. After this, the children level is split in k clusters with *Bisecting k-means* [14]. For each group generated, one node is added to the structure and all group items are associated with the respective group node. This split process is recursively applied to each new group node while the quantity of points in a group is greater than a threshold. This threshold is the square root of the total n. *Bisecting k-means* uses $k = \sqrt{m}$, with m being the number of items

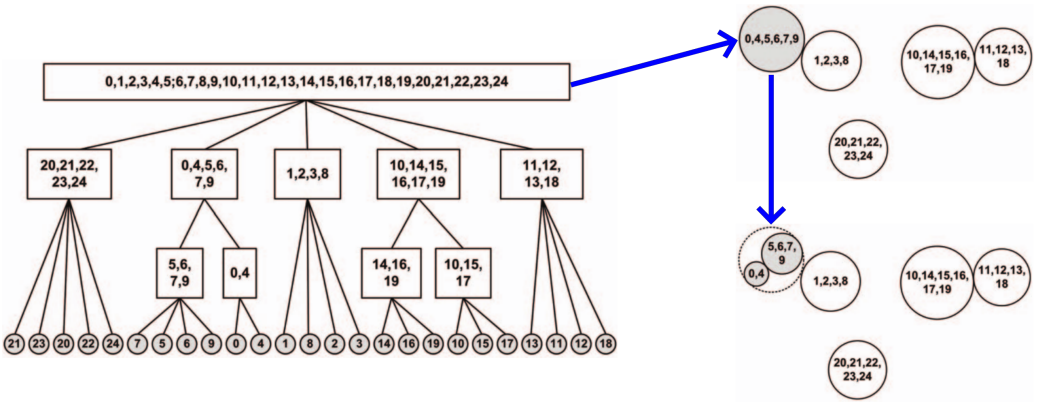


Fig. 1. HiPP partition, projection and interaction example. Adapted from [2].

within the node (group) that is being split. The left part of the Figure 1 illustrates the result of this step when applied to a small number of data instances.

In the second step, tree nodes are projected using the LSP projection. Initially, the first tree level is projected, placing circles, that represent group nodes, on a plane (Figure 1 right-top). In order to perform group projection, group centers are taken as a group attribute and the size of the circles is proportional to the number of items inside them.

The third step in HiPP applies a spreading algorithm to remove overlap in visualization space. A vector is created between every 2 projected nodes is used to spread them. These nodes are moved in the vector direction but in opposite directions. As in the projection process, the group nodes are represented for their centers. The right-top part of Figure 1 illustrates the spreading of group nodes (circles).

The right-bottom part of Figure 1 presents a result of user interaction with tree nodes. During the interaction, node levels are shown until tree leaves are reached.

In addition, for textual exploration, HiPP authors implemented text topic extraction based on term covariance. The topic terms are presented in each group during as labels during the interaction, and groups are colored based on topics.

HiPP also allows users to split and reassemble groups interactively. The strategy is effective when the user needs to tailor the clustering, but it has the limitation that it does not allow for flexibility of clustering and projection techniques or data types, or even evaluation based on clustering.

III. Materials and Methods - The xHiPP approach for multi-level projection-based visualization

This section presents the changes to HiPP that make up the features of the xHiPP point placement. We chose to show the features in that way to contrast with the basic approach and highlight the new strategy. Besides central aspects, such as flexibility of methods (both projection and clustering) and availability of summarization tasks for more dataset types, additional interaction and visualization functionalities were implemented. We also describe implementation details, datasets used for tests and tools employed for implementation.

A. eXtended HiPP

The first improvement of xHiPP relates to the implementation of the tree partition process. Instead of using just one clustering algorithm (*Bisecting k-means*), xHiPP implements three cluster techniques: *k-means*, *k-medoids* and a basic *Hierarchical Agglomerative Clustering*.

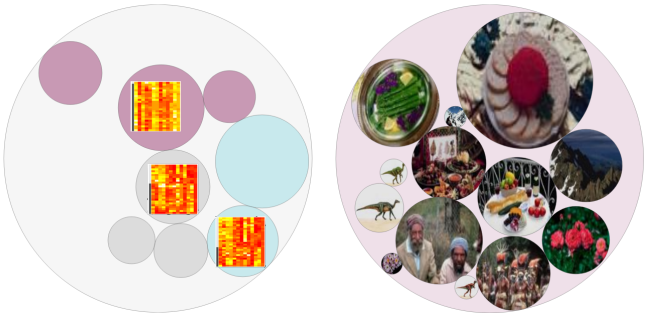
The next addition is linked with to the number of clusters. The original paper uses $\sqrt{n}$ to decide how many groups a node needs to be split in. Unfortunately, this value grows large when n is very large. For instance, with 1,000 instances it generates 31 groups. With 10,000 instances, it generates 100 groups. In our implementation the Sturges' Rule [15] was employed instead; it defines $k = 1 + 3.3 \times log(n)$. This measure results in a smaller number of clusters (e.g 10 to 1,000 instances and 14 to 10,000 instance). The user can modify this quantity to better adapt the data structure to the case at hand.

Another change was made in the projection process. Additionally to *LSP* to project external nodes, it is possible to apply *MDS*, *Force Scheme*, *t-SNE*, *PLMP*, *LAMP* or the *PCA*, all cited previously. To improve processing time, internal nodes are always projected with *Force Scheme*.

Perhaps the most impact in xHiPP is given by the next modification. The flow to generate the final projection in HiPP is exclusively clustering followed by projection. However, since most projections are very effective at approaching similar data points in projected space, they can be capable of partitioning the dataset with high precision. Besides first partitioning data to generate the hierarchical structure and then projecting this structure, in xHiPP it is also possible to first project the data and then create the hierarchy from a projected 2D space. For those datasets that are a good match to the processes of projection, we found that the precision is improved. More than that, it adds an extra level of flexibility for analysts to find the proper combination of methods fit their data.

Besides implementing functionalities for text exploration (such as the original topic modeling and the new word clouds), new functionalities were implemented to ease exploration for other data types, such as image, audio, and general datasets. The groups' medoids are used to represent image sets (see

Figure 2b). With audio files, if they have spectrogram image correlated with them, the process is the same as with images. Groups of ordinary data are summarized by heatmap images that map the description of data attributes distribution (Figure 2a).



(a) Heatmaps summarize some Wine dataset groups.

(b) Image medoids that represent Corel dataset groups.

Fig. 2. Group summarization examples.

For the exploration of text, word clouds were added to summarize the content of both groups and individual documents. It shows the one hundred most frequent words (Figure 4). Each group can also be examined by loading the appropriate data. Functionality for text, image, audio, and combination of audio and image were added to support examination of data points and groups. Lists of values are also presented for other general attribute sets. Colors represent node label (group nodes are assigned the color of the predominant label inside the group).

With these changes and additions, it is possible to use a combination of methods and identify which of them better represents the data collection under analysis.

Additionally to the visualization of the circles in a level of the structure, the tool also shows a Treemap of the hierarchy formed by the groups. That is very useful to evaluate the mixture of labels within groups.

Some original HiPP feature were not implemented. Text topic extraction was performed with *Latent Dirichlet Allocation* (LDA) [16] algorithm instead of evaluation of word co-occurrence. Joining and splitting clusters is not implemented since the goal of xHiPP is to find meaningful structure in the data instead of tailoring the structure.

Finally, as an implementation detail, the root node radius was estimated with the distance between data centroid and the farthest data item.

B. Datasets

In order to test xHiPP, the following datasets were employed. From *UCI Machine Learning Repository* [17]: *Iris* contains 150 items, 4 attributes and 3 classes that represent iris plant gender; *Wine*, which contains 178 items, 13 attributes and 3 labels that represent different Italian wines; and the *Banknote* dataset, which contains 1,372 items, 4 attributes, and 2 labels and represents images from genuine and forged dollar notes.

The Corel image dataset [18] contains 1,000 items, 150 attributes and 10 la labels such as bus, beach, flowers, etc.

A text dataset formed by RSS news from BBC, CNN, Reuters, and Associated Press was used [2] was also employed. This text collection has 2,684 items, 2,217 attributes and is not labeled.

Datasets for acoustic landscapes were employed as well. The complete dataset has 4,340 audio files represented by twenty-seven acoustic features. The data were labeled with the natural area where the audio was recorded; labels are *CostaRica1*, *CostaRica2*[1], *Ilheus*[2] and *Lajes*[3].

C. Material

The code for xHiPP was implemented anew. In this research the following tools were employed. To implement xHiPP the R[4] language was used with the projection package *mp*[5], the parallel programming package *doParallel*[6], and the *topicmodels* package[7], that pre-processes text and extracts their topics. Visualization results are displayed with Javascript library D3.js[8] associated with the *Shiny*[9] library from R that provides a web server. In summary, data partitioning and projection steps are run in R, and the spreading and rendering steps are executed with Javascript.

IV. Experimental results

This section reports the experimental results performed with the datasets cited in Section III-B. Summarizing the method for xHiPP evaluation, we generated results for each dataset. The first tests modified projection parameters (process order, cluster algorithm, and projection algorithm) and compared the visual and numerical results (*Stress*, *Neighborhood Hit and Preservation*, and *Silhouette Coefficient*) generated by them. Other experiments looked for ways in which xHiPP could help exploration tasks, showing results of analysis and some insights obtained. The last experiments report exploration results reported by xHiPP testers that applied our projection extensions to their own datasets.

The first experiment was performed with the **Iris** dataset. *K-means* and *Force Scheme* were used and the order process (clustering-projection, projection-clustering) was varied to evaluate its impact on results. The original data class segregation was maintained with the two process orders and the metrics values were better with the order projection first then clustering, as shown in Table I.

Another experiment was run with the **Wine** dataset. In this test, clustering-projection order and *Force Scheme* projection were applied and the cluster algorithms were varied to evaluate their results. Table II shows that the *k-means*

[1]Terrestrial audios collected in two areas of *La Selva* Biological Station, Costa Rica

[2]Underwater audios collected in south coast of Bahia State, Brazil

[3]Underwater audios collected in *Laje de Santos* Marine State Park, on the coast of São Paulo State, Brazil

[4]https://www.r-project.org/

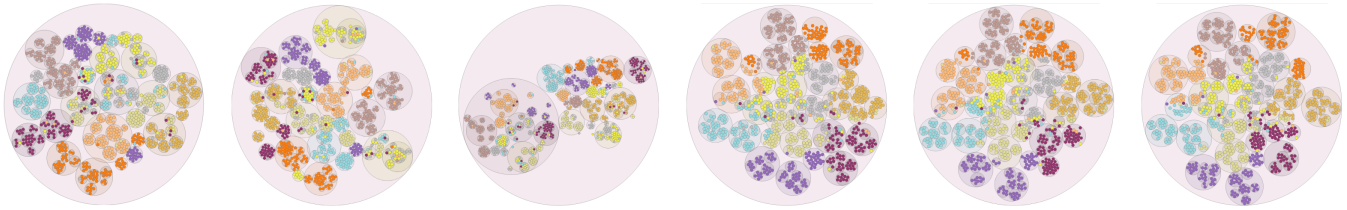
[5]https://cran.r-project.org/web/packages/mp/index.html

[6]https://cran.r-project.org/web/packages/doParallel/index.html

[7]https://cran.r-project.org/web/packages/topicmodels/index.html

[8]https://d3js.org/

[9]https://shiny.rstudio.com/



(a) *k-means* + PCA. (b) *k-medoid* + MDS. (c) Hierarchical + *Force*. (d) t-SNE + *k-means*. (e) t-SNE + *k-medoid*. (f) t-SNE + Hierarchical.

Fig. 3. The best results for Corel dataset, varying clustering and projection algorithms, as well as process order. Colors represent each data label.

TABLE I
NUMERIC VALUES GENERATED BY TESTS WITH **IRIS** DATASET, VARYING PROCESS ORDER.

	Stress	**N. Preserv.**	**N. Hit**	**Silhouette**
Clustering-projection	0.7061	0.5483	0.8644	0.4707
Projection-clustering	0.4113	0.7494	0.9483	0.5562

algorithm presented the best silhouette result, the *Hierarchical* algorithm reached the best *Neighborhood Preservation* value, and the *k-medoid* algorithm presented the best *Stress* and *Neighborhood Hit* values. Even with these variations, the projections positions, and the classes segregation were not visually perceptible.

TABLE II
NUMERIC VALUES GENERATED TO TESTS WITH **WINE** DATASET, VARYING CLUSTER ALGORITHMS.

	Stress	**N. Preserv.**	**N. Hit**	**Silhouette**
k-means	1.3453	0.5808	0.6534	0.1907
k-medoid	1.0894	0.5959	0.6634	0.0923
Hierarchical	1.3134	0.6154	0.6124	0.1775

The next experiment was conducted with the **Banknote** dataset. In this test, clustering-projection order and *k-means* were used, and the projection approaches were varied. Table III shows that *Force Scheme* obtained the best *Silhouette* values, MDS reached the best *Neighborhood Preservation* value, PCA obtained the best *Stress* value and t-SNE presented the best *Neighborhood Hit* value. Even with these variations, the projections positions, and the classes segregation were not clearly visible.

TABLE III
NUMERIC VALUES GENERATED TO TESTS WITH **BANKNOTE** DATASET, VARYING PROJECTION ALGORITHMS.

	Stress	**N. Preserv.**	**N. Hit**	**Silhouette**
Force	0.9442	0.6261	0.9168	0.1914
MDS	0.9371	0.6346	0.909	0.1922
PCA	0.9297	0.6182	0.9003	0.1939
t-SNE	0.9391	0.6265	0.9195	0.1945

Another experiment was performed with the **Corel** dataset. In this test, clustering, projection and process order were varied. All variations generated 42 combinations. The best visual results are presented in Figure 3, and their numerical results are shown in Table IV. In tests that used clustering-projection (Figures from 3a to 3c), the combination of *k-means* and PCA reached the best values of *Stress* and *Neighborhood Preservation*. On the other hand, *k-medoid* and MDS presented the best *Silhouette* and *Neighborhood Hit* values. Tests that applied projection-clustering (Figures from 3d to 3f) reached the best tests results, always with t-SNE projection, and their measurement values are all equivalent.

TABLE IV
THE BEST NUMERIC VALUES GENERATED TO TESTS WITH **COREL** DATASET, VARYING CLUSTER, PROJECTION AND PROCESS ORDER.

	Stress	**N. Preserv.**	**N. Hit**	**Silhouette**
k-means + PCA	0.9169	0.4408	0.6792	0.1052
k-medoid + MDS	0.9288	0.4265	0.7001	0.1329
Hierarchical + Force	0.9334	0.329	0.493	-0.0407
t-SNE + k-means	0.9083	0.5541	0.7762	0.3177
t-SNE + k-medoid	0.9083	0.5541	0.7762	0.3177
t-SNE + Hierarchical	0.9083	0.5541	0.7762	0.3177

The next experiment was executed with the **News** dataset, with the goal of evaluation xHiPP in exploring text data. Figure 4 illustrates the projection results with *k-means* and *t-SNE* as clustering and projection techniques, and projection-clustering order. Tests with clustering-projection order generated an unbalanced structure with more than 20 levels. The figure shows examples of terms associated with group topics, as well as word clouds generated to the complete dataset and to a specific group.

The group highlighted in Figure 4 contains items from different topics. The word cloud supports the perception that text with distinct topics maintains similar word groups.

An experiment was conducted with the **Ecological acoustic** dataset, with the goal of evaluating xHiPP capability in exploring this type of data. Figure 5 presents projection result with *k-means* and *t-SNE* as clustering and projection approaches, and projection-clustering order. The projection illustrates the separation of labels present in the data, mostly between terrestrial and underwater sounds. The data corresponding to the terrestrial area is also separated because of characteristics of collecting areas. The picture also shows representative spectrograms of some groups projected. At the bottom and bottom-right, the displayed spectrograms indicate a group that contains audios with vessel sounds patterns. On the right, the dark gray level in spectrograms shows rain and an insect sound pattern. On the bottom-left side, the spectrogram depicts a sound of fish and humpback whales patterns. On the left,

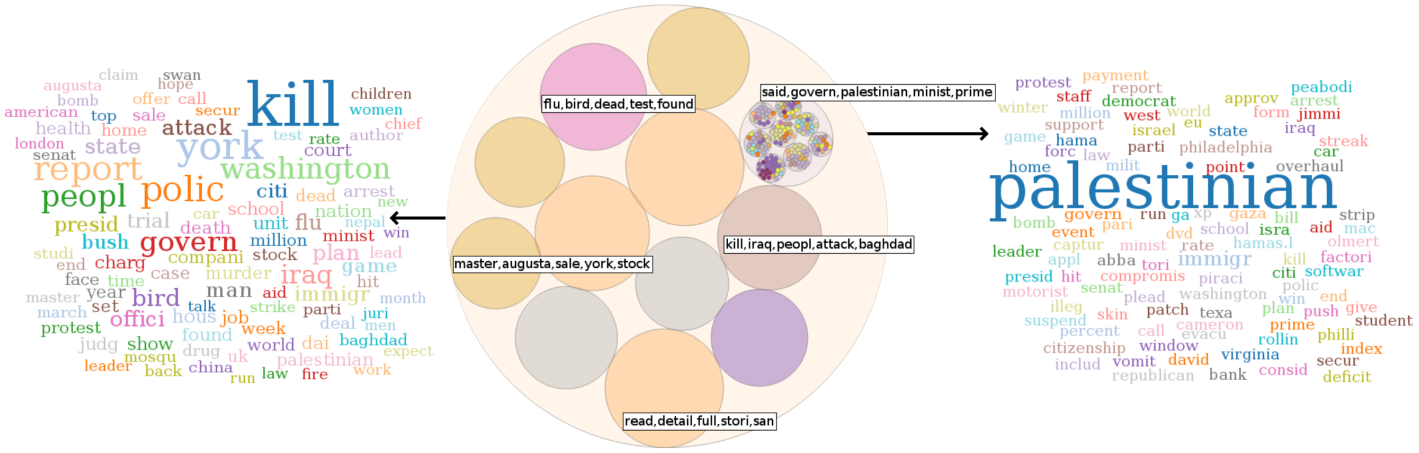


Fig. 4. xHiPP projection of news dataset. Item and groups color represent text news topics. Examples of terms associated with group topic are presented. Word cloud on the left refers to the complete dataset and the one on the right represents the group expanded. Word colors have no semantic meaning.

the spectrogram shows nocturnal sounds such as those of insects, amphibians an others, and the top spectrogram mainly highlights bird sounds.

A. Results reported by testers

These users tests aimed to verify the performance of xHiPP additions in real user analysis, and the importance of both projection layout and projection metric evaluation. Users were presented to a simple xHiPP implementation, where they could change parameters (process order, clustering and projection algorithms, and other original xHiPP attributes), visualize metric values and freely interact with data projected.

The first tester used xHiPP to explore **Medical Trauma** data with a sample of 1,000 patients described by 13 orthopedic trauma indexes and clinical attributes, classified according to 8 survival/non-survival risk status. Tester reported that xHiPP interface is simple since brief instructions were necessary to interact with it. In his tests, the cluster-projection order (*k-means* with *Force Scheme*) generated consistent clustering with a good visual layout as reflected by data labels. Projection-cluster order, with any algorithms combination, presented visual clutter; however, the cluster configuration can represent abstraction levels that is interesting to the data probed.

Some observations about data could be drawn. For example, a patient labeled with *good recovery* was placed in a group that contains mostly patients with the outcome of *death*. Analyzing data attributes, this situation occurred because the patient was admitted to the hospital in a critical situation, but a medical team could turn patient condition from critical to *good recovery*. Conversely, some clusters with predominance of *good recovery* and *limited recovery* labels have some *death* outcomes. This happened due to patients being in good conditions when admitted and later developing complications evolving to a critical stage.

The second tester employed xHiPP to explore **Time series** data with 309 observations collected daily by a High-Volume Air Sampler, between April 2014 and April 2015. Each observation is described by 20 attributes, such as weather conditions, the concentration of atmospheric Particulate Matter (PM), the month of collection (12 columns generated by One Hot Encoding[10] technique applied to the month) and the year of collection (2 columns generated by One Hot Encoding technique applied to year). Observations were labeled with year seasons. With brief explanations, the user was able to interact with xHiPP. In his tests, both cluster-projection and projection-cluster order generated consistent data layout and the combinations with *k-means* and *t-SNE* presented slightly better results than other combinations.

Groups generated have samples from different months or seasons, but with similar PM level range. This occurs owing to the local climate dynamics that leverage PM levels (when it is too dry and cold, pollution particles tend to stay more concentrated). Even in the same month or season, some days appear with a combination of temperature and humidity levels that may fit into the men of another season.

With this type of exploration, the user could start the analysis to understand months and seasons with similar pollution patterns.

Both users found the approach very useful during their initial data analysis and for finding and explaining unusual patterns. Their knowledge about their data was ratified and some new questions, that could be investigated, arose. During their analyses, the visual data distribution and aggregation were more important than the available metric values (Section II-B).

V. Discussion

The option to vary the process order (clustering-projection, projection-clustering) in xHiPP allows for enhanced data representation, as it is shown in Table I and with user tests. The possibility of choosing clustering techniques and associate them with several projection techniques could enhance projection efficacy. For instance, PCA has a visual inconsistency

[10]Approach used in data analysis to turn categorical values into binary representation.

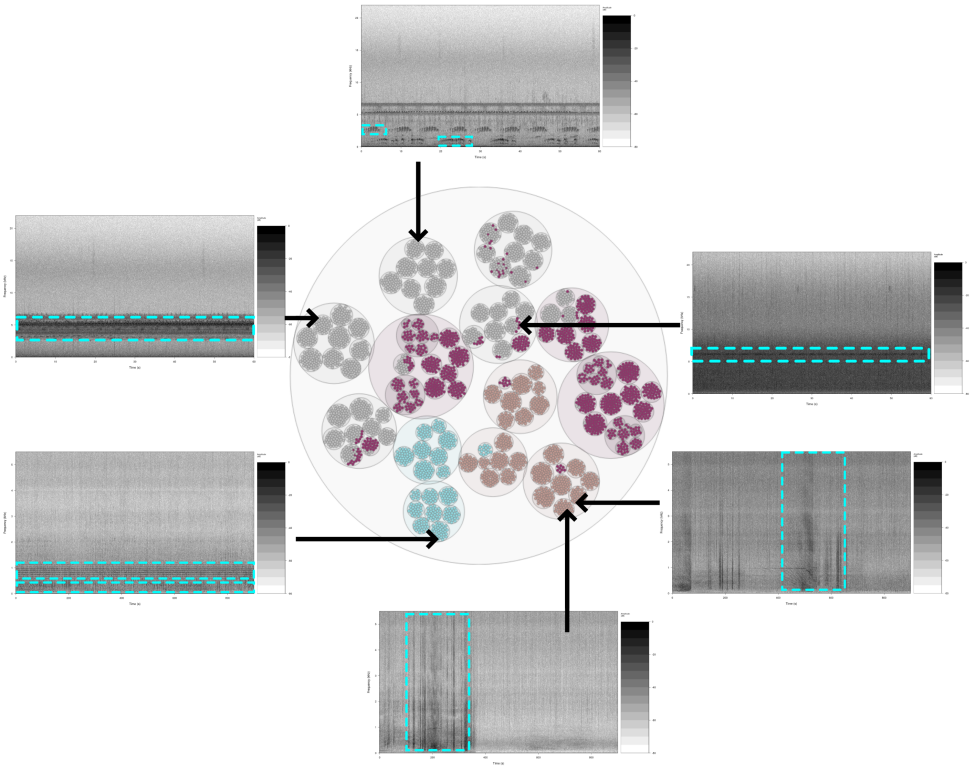


Fig. 5. xHiPP projection of acoustic dataset. The colors represent the labels used: purple (*CostaRica1*), gray (*CostaRica2*), light blue (*Ilheus*) and light brown (*Laje*). The spectrograms show the main content of some groups presented. Highlighted spectrograms areas get attention to distinct sound patterns.

when non-linear data attributes are used, notwithstanding the use of appropriate cluster algorithm can aid it to reach better results, as it is shown in Figure 3. Additionally, taking advantage of t-SNE segregation can improve the results of simple clustering algorithms such as Hierarchical, and the capacity for discrimination during data exploration.

The hierarchical structure can contain several levels in a scenario that cluster algorithms generate an unbalanced structure as was the case with our text dataset. This situation drives users to expand more levels to explore and assess datasets. Inversion of process order was capable of eliminating part of the problem, generating a balanced structure with fewer levels and more adequate exploration. Additionally, the word cloud extended the description made by group topics, enhancing text summarization and helping analyses of text content.

The employment of group medoids to summarize cluster data was capable to guide users to specific patterns into an audio dataset. Users can focus on patterns of interest around issues related to control of species diversity, monitoring of preservation areas and their changes, and so on. The use of the xHiPP in this scenario could present both differences among regions and singularities of groups presented.

Heatmap summaries helped testers to understand how attributes were distributed inside groups. With this information, testers were capable to perceive why items from different classes were put together and why groups are closer or farther from each other.

Unfortunately, when the instances amount (more than 1,000) and the attributes quantity (more than tens) are large, the process efficiency decreases. This happens due to the computational complexity of the projection $O(n\sqrt{n})$, mentioned in Section II-C, the complexity of clustering algorithms (*k-medoid*), and the R language characteristics, even with parallel programming techniques employed. Another bottleneck is the slowness of D3.js library to render thousands of data (more than 5,000). Regardless, the framework can be used for larger datasets in regards to visual space organization.

The tool code is available for users at https://github.com/fabiofelix/xHiPP.

VI. Conclusion

This paper proposed xHiPP, a multi-level strategy for multidimensional data representation, point placement, and exploration. Although based on the HiPP approach, a series of extensions with the aim of improving its data representation and exploration assets led to a new approach for testing combinations adequate to data sets and tasks of different nature and achieving good data representation for user-centered pattern analysis. The choice of cluster and projection algorithms, as well as the process order, can lay out different datasets in a coherent reproducible way. The presentation of internal group patterns also well-conducted data examinations. As limitations, processing time and visual data manipulations need to be improved in order to deal with larger datasets than the ones

used here. Future work will deal with these drawbacks as well as with the choice of a platform with better memory management.

ACKNOWLEDGMENT

This research was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq). The authors wish to thank: CAPES; professors Linilson R. Padovese from Polytechnic School of the University of São Paulo, Brazil, and Bryan C. Pijanowski from Purdue University, Indiana, USA, for their ecological data; and testers Erasmo Artur da Silva and Evandro Ortigossa that collaborate with useful feedback and ideas to improve ordinary data summarization.

REFERENCES

[1] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen, "Perception-based evaluation of projection methods for multidimensional data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 81–94, Jan 2015.

[2] F. V. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1229–1236, 2008.

[3] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualizaton: foundations, techniques, and applications*, 2nd ed. CRC Press, 2015.

[4] J. B. Kruskal and M. Wish, *Multidimensional scaling*. Sage, 1978, vol. 11.

[5] E. Tejada, R. Minghim, and L. G. Nonato, "On improved projection techniques to support visual exploration of multi-dimensional data sets," *Information Visualization*, vol. 2, no. 4, pp. 218–231, 2003.

[6] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[7] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 3, pp. 564–575, 2008.

[8] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato, "Local affine multidimensional projection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2563–2571, 2011.

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[10] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[11] S. Ingram, T. Munzner, and M. Olano, "Glimmer: Multilevel mds on the gpu," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 2, pp. 249–261, 2009.

[12] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[13] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining. 1st," 2005.

[14] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[15] H. A. Sturges, "The choice of a class interval," *Journal of the american statistical association*, vol. 21, no. 153, pp. 65–66, 1926.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[17] M. Lichman, "UCI Machine Learning Repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[18] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.