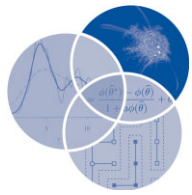
Advanced Review



Visual matrix explorer for collaborative seriation

Innar Liiv,[1]* Rain Opik,[1] Jaan Ubi[2] and John Stasko[3]

In this article, we present a web-based open source tool to support cross-disciplinary collaborative seriation with the following goals: to compare different matrix permutations, to discover patterns from the data, annotate it, and accumulate knowledge. Seriation is an unsupervised data mining technique that reorders objects into a sequence along a one-dimensional continuum to make sense of the whole series. Clustering assigns objects to groups, whereas seriation assigns objects to a position within a sequence. Seriation has been applied to a variety of disciplines including archaeology and anthropology; cartography, graphics, and information visualization; sociology and sociometry; psychology and psychometrics; ecology; biology and bioinformatics; cellular manufacturing; and operations research. Interestingly, across those different disciplines, there are several commonly emerging similar structural patterns. Visual Matrix Explorer allows users to explore and link those patterns, share an online workplace and instantly transmit changes in the system to other users. *WIREs Comp Stat* 2011 DOI: 10.1002/wics.193

Keywords: seriation; matrix reordering; collaboration; information visualization

INTRODUCTION

This article introduces a web-based tool for exploratory visual analytics—Visual Matrix Explorer (VME)—that enables dynamic evaluation and visualization of matrices, and the linking of different seriation results. The main motivation for this article was to change the tradition of literature review being a static textual result and to allow for an interaction between different theories and methods presented in overviews. Current article is complementing a recent literature review[1] on seriation and matrix reordering. We presently use the tool to evaluate, visualize, and link different seriation results from different disciplines, but the real value is much broader due to seriation being one of the fundamental learning components[2] to organize events, objects or other phenomena we are looking to understand.

VME is a tool for researchers investigating surveys and questionnaires, social networks, process execution logs or other data that can be expressed in a tabular form. It supports dynamic theory building and comparison, allowing the user to interactively explore and link any rankings of importance, interestingness, and focus (from any theory)—and finally settle for a suitable interpretation. The dynamic nature of the tool is additionally manifested in the capability of focusing on a subset of data and running operations therein.

This article is not just about comparing the results and theories of seriation and clustering, but—using the terminology of information visualization and interaction—is concerned with cross-disciplinary and cross-theory brushing.[3] VME gives the user a possibility to inspect, whether a collection of facts—a meaningful knowledge in one discipline—can yield relevant structures in other theories, adds new opportunities for exploratory data analysis and knowledge discovery in general.

The article is organized in the following way. In the next section, the difference and interplay between seriation and matrix reordering is described, followed by a brief discussion. In Section '*Comparing Theories*', the objective of finding related traits in different theories is elaborated. In Section '*The Functionality of VME*', the functionality of the tool is described. Section '*Illustrative Examples from Different Disciplines*' contains several accounts of exploration from different disciplines, each with unique investigative and analytical tasks, but a shared general goal.

*Correspondence to: innar.liiv@ttu.ee

[1]Department of Informatics, Tallinn University of Technology, Tallinn, Estonia

[2]Computational Systems Biology Laboratory, University of Georgia, Athens, GA, USA

[3]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

DOI: 10.1002/wics.193

SERIATION AND MATRIX REORDERING

As stated in '*Introduction*' section, seriation assigns objects to positions in a sequence, according to some predefined principle and objective. An intuitive example would be a shopping list that some people write down before a visit to a store. This list can be compiled in multiple ways. Depending on the organization, it can serve different goals—either to contain just the original information (what to buy?) or to include a categorization and ordering (seriation) in some way more reasonable for shopping (e.g., to prioritize or to minimize the walking distance in a shop). An interested reader is referred to Belknap's recent book about ordering lists.[4] Seriating a data table is a natural two-dimensional extension to organizing lists, where both rows and columns can be reordered.

While new methods of manipulating and storing data are being developed and brought into mainstream, the table (matrix) with its method—the spreadsheet—remains the predominant form of storing data,[5] already, since the ancient times.[6]

When finding a seriation for a matrix we can, depending on the optimization objective, perform it independently for the rows and the columns as suggested by Lenstra,[7] or do it in a dependent manner recommended by Niermann.[8] The combinatorial optimization problem is essentially one of finding a permutation, thus being in a factorial search space and using (often greedy) heuristics. As Figure 1 depicts the process of seriation by the classical example of Bertin's townships,[9] it can be observed that the result makes the relationships and patterns within the dataset more evident—most importantly without any dimensionality or other reduction in the data. Note, that after the initial data has been discretized, visualization transforms ones into black dots with zeroes forming the white background, using coding similar to a seriation package in R environment.[10]

The research of seriation and matrix reordering methods go back more than 100 years.[1] However, with a few exceptions, the research has only concentrated on choosing the best static representation, whereas according to the information visualization studies, the interaction with the representation should add a lot of extra value.[11,12] One example of an interesting, visually appealing and interactive interface of representing matrices is NodeTrix[13] which, concentrates on one-mode networks (graphs) and does not include support for most of the functionality described in the following sections. From the perspective of rigorous comparison of different vertex ordering and matrix permutation algorithms, the reader is referred to a recent study by Mueller et al.[14] VME, described in this article, can be thought of as a perfect environment for such comparisons, also enabling different kinds of interaction with such matrices, in order to highlight the differences and to support the final interpretation of the results.

COMPARING THEORIES

Is the psychoanalytic theory of Freud better than the self-concept theory of Rogers? While such questions, in any discipline, may be inadequate at general level, they are of primary concern for the explorer and experimenter. Even if the experimenter does not have any

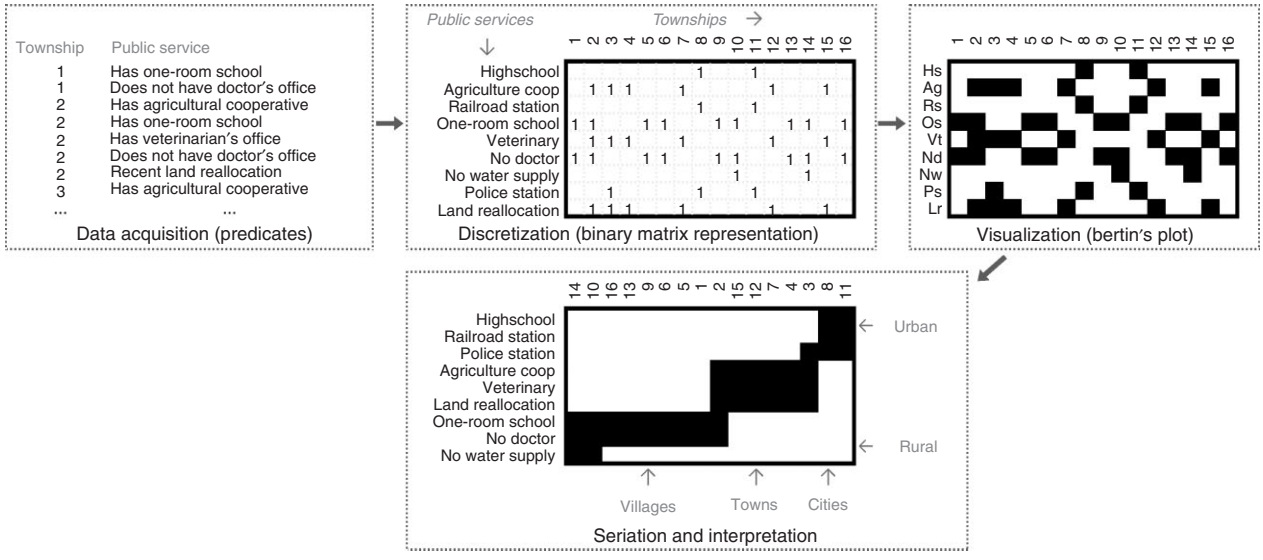


FIGURE 1 | Workflow from acquiring predicate data to visual knowledge mining.

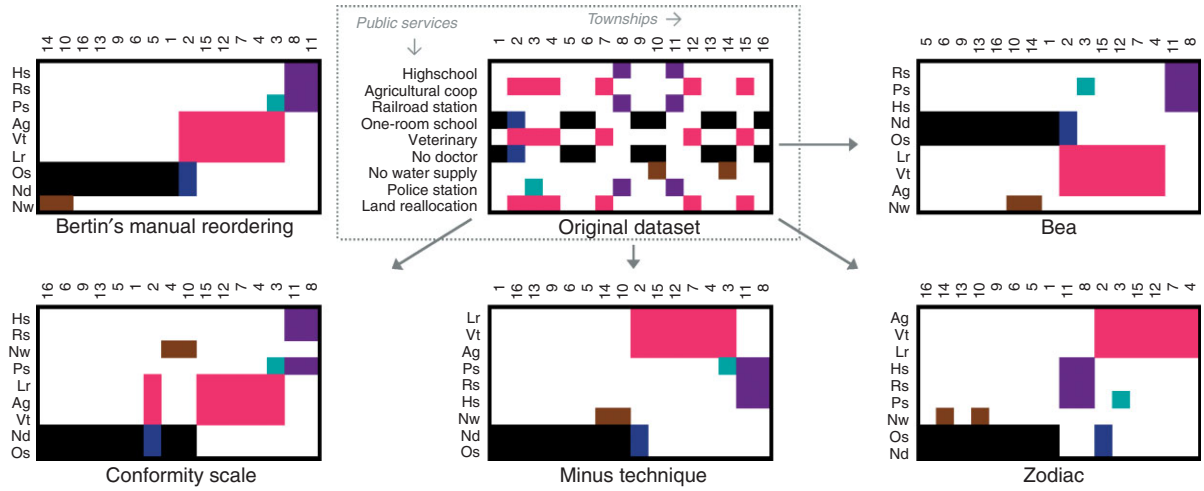


FIGURE 2 | A dataset rendered in different permutations.

preferences, there should be and efficient and objective way to analyze the differences. As stated by Kelly,[15] 'a theory may be considered as a way of binding together multitude of facts, so that one may comprehend them all at once.' Therefore, it is the facts, which allow explicit binding between different views and theories.

In the following example, the previous dataset is depicted in the top center position of Figure 2. This image also includes five other orderings from different algorithms originally meant to serve different disciplines. Technically, each of those can perform a combinatorial optimization using a different objective function, however fundamentally, all of those objective functions were tailored to emphasize as much similarities and visual clusters in the dataset as possible. These objective functions, therefore, were based on slightly different assumptions, biases and subjective beliefs.

While there may not be a single specific layout of the data that brings out the 'natural structure' of the dataset, we can identify different interesting aspects from each different view. An interested reader is referred to an extended historical review[16] in the search for the 'natural structure'. The tool presented in this article would allow comparing those different methods and approaches in order to find the natural system of things under observation.

In the view 'Bertin's manual reordering' (Figure 2) of data, VME is utilized for highlighting different subsets of data using different colors. Thereafter, the displacement of Bertin's blocks in other views of data becomes evident, allowing for a comparison of different results.

In Figure 2, magenta-colored region has the characteristics of a medium-sized rural settlement. Settlement 3 serves as a transition from a rural settlement to an urban township. The key attribute of this shift (police station) is separated by the BEA algorithm and colored in teal.

One dataset can have many different representations—in this case, permutations of rows and columns. In cellular manufacturing, 'similar' machines are placed adjacently, thereby improving the process flow,[17] by reducing worker round-trips in the factory, and minimizing the time required for transportation of materials between machines. In such a setting, rows of the matrix can represent machines while columns stand for processes, for which the machines are utilized. McCormick et al.[18] created a seriation method Bond-Energy Algorithm (BEA, Figure 2), which effectively maximizes the contiguous chunks, as clumps of data get placed next to each other.

Contiguous chunks can be interpreted as machines that should physically be placed together. If solitary bridges between chunks exist, they are, in the manufacturing workflow, usually interpreted as bottlenecks (in terms of time or material transportation).

When, from the field of social sciences, we consider social network analysis, an interesting parallel with the aforementioned manufacturing concepts can be drawn. The graph structure of a social network can be represented using an adjacency matrix, which in itself is easy to visualize. One difficulty in network analysis, though, is the calculation of a layout for a graph—as the nodes of the network do not contain any inherent ranking properties. The process of applying cellular manufacturing matrix seriation (searching for chunks) on the adjacency matrix yields a rank for each object. The chunks in the data and the bridges can, respectively, be considered as groups of friends and mediators in between, and the ranking information can be used while constructing the

layout. In network analysis, a clique is an often-sought structure, which, commonly defined as an inclusive group of friends, colleagues or individuals in general, helps to identify like-minded communities sharing the same knowledge, interests or ideologies. It should be noted, though, that often the graph-theoretic definition of a clique—a group where each member is connected with every other member—cannot be exploited, because real-world data rarely exhibits perfect structures. However, an exploration of reasons behind these irregularities could be of interest, for instance, if Alice socializes with Bob and Carol, what hinders Bob from befriending Carol? VME attempts to present a supplementary instrument for analyzing networks by utilizing aforementioned visual clustering in discovery of communities and anomalies therein.

THE FUNCTIONALITY OF VME

As stated in the brief discussion of seriation, research has traditionally concentrated on the best static representation of data, whereas VME enables interaction and, also, lets the user focus on a subset of interest. The functionality of the tool has been designed with the help and methodological guidance from the frameworks and taxonomies by Wehrend and Lewis,[19] Shneiderman,[20] Amar and Stasko,[21] and Amar et al.[22] that concentrate on the analytical tasks people undertake while investigating a dataset.

VME is a web-based tool that enables visualization and analysis of binary matrices. The input files are stored in a comma-separated values (CSV) format and may contain descriptive labels for rows and columns, as illustrated in the top-center plot of Figure 1. The system can be run in a web-browser supporting CSS, JavaScript, and Canvas elements.

Operation of VME is organized into workspaces. A workspace is a shared collaborative environment, which also acts as a tracker of actions taken by the user. The path of exploration and hypothesis discovery of an analyst is recorded in his 'trail of thought', which is shared across all browsers displaying the current workspace. There can be multiple workspaces of one file which is useful for following alternative trails of thought.

The workspace contains a side-by-side grid of graphical plots visualizing the dataset under different 'Permutations' (Figure 3), each of which is the result of applying one algorithm. The algorithms include:

- *orig*—stands for the original dataset;
- *countones*—our fast $O(n \log n)$ heuristic for larger matrices, based on sorting by the frequency of 'ones';
- *conf*—conformity scale; *minus*—minus technique, and *plus*—plus technique—algorithms from the Monotone Systems metaheuristic by Mullat[23] and Vyhandu.[24]
- *bea*—McCormick's BEA.[18]
- *roc2*—an enhanced rank order clustering by King et al.[25]
- *modroc*—an extension of the rank order clustering for group technology by Chandrasekharan and Rajagopalan.[26]
- *art*—a Carpenter–Grossberg neural network based clustering by Kaparthi–Suresh[27] and Kusiak–Chung.[28]

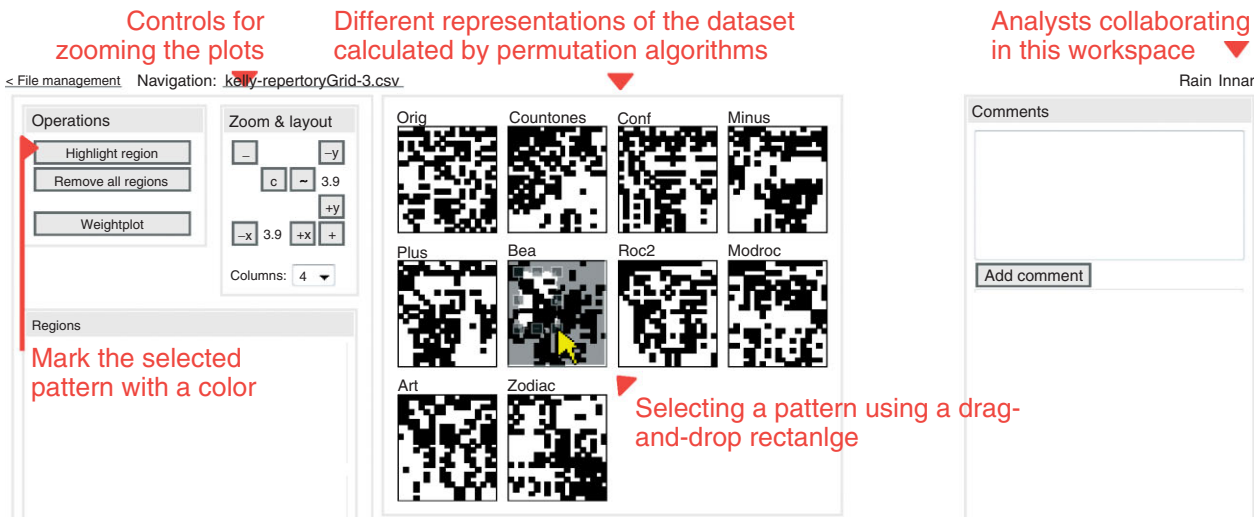


FIGURE 3 | An overview of the VME workspace.

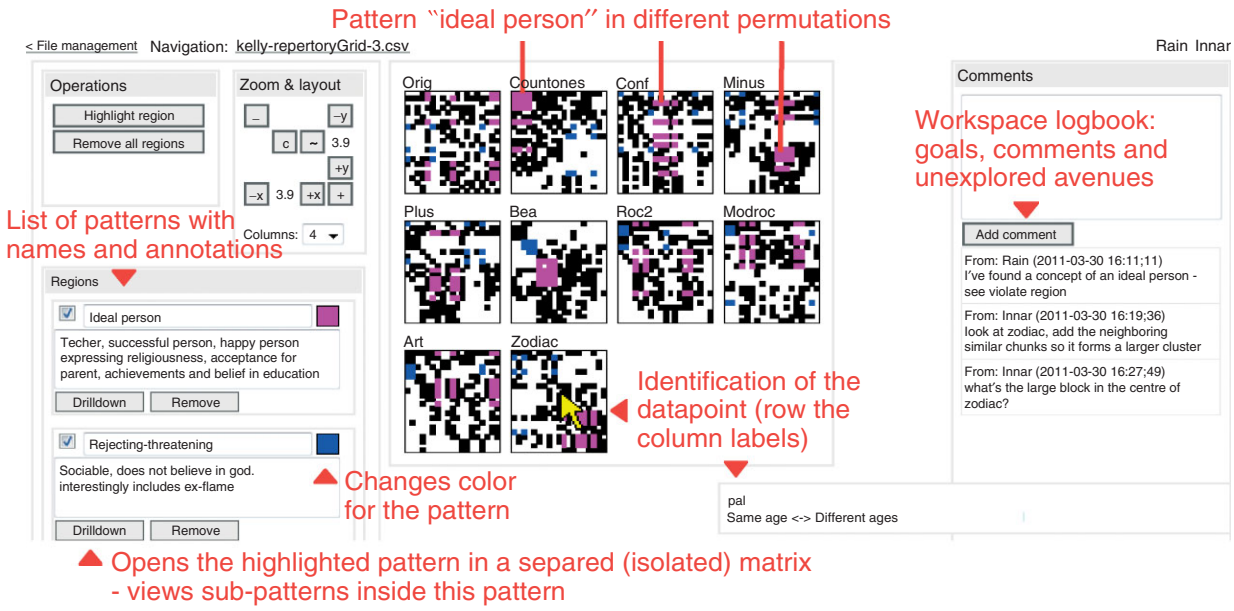


FIGURE 4 | Workspace with two highlighted patterns.

- *zodiac*—ideal seed method for part-family and machine-cell formation in group technology by Chandrasekharan–Rajagopalan.[29]

The system is open for extensions and new permutation algorithms can be included easily.

Depending on the size of the dataset and the size of the browser's viewport, a scaling factor may be necessary to fit images inside the window. In the case of large matrices, a single pixel on the screen may consist of multiple data points, similar to the information mural technique.[30]

A rectangular region of a permutation can be selected by clicking-and-dragging, as shown in the Figure 3. The selected region ('clump of points')—a subset of objects and attributes—is consequently brushed[3] with color on every plot, thereby enabling a comparison of results of different theories. In order to further identify and interpret the results, individual data points can be inspected with mouse which will result in the associated object and attribute name being displayed.

After consulting with data point names, a user can assign a descriptive name for the region. The system is designed to support easy capturing of hypotheses, for the purposes of exploration, interpretation, and reporting. Each pattern can be complemented with a detailed annotation which enables sharing the intent of the analyst instantly among fellow collaborating users. Newly identified patterns, their descriptions and explanations are delivered to all members of the workspace in a real-time fashion. A general commenting section provides a tool for driving the analysis process. It can be used as an accessible place for storing goals, unexplored alternative paths and general remarks. Figure 4 illustrates a workspace with two brushed regions along with annotations.

VME has a 'drilldown' feature that is in its nature similar to financial reporting and OLAP (see Ref 31 for an overview), in case of which transactions of a specified set of accounts are queried in detail. Drilldown can be used in order to reveal weaker patterns in the dataset (Figure 5, in pink). A weaker pattern, in our context, means a structural phenomenon of the dataset, which does not emerge clearly in the overall view; but if the structural constraints from other, more evident patterns are removed, will present interesting traits. In other words, we may say, that before drilling down we have a limited number of degrees of freedom for expressing patterns, as the rows or columns of the table cannot be split into two or more independent parts. Procedurally, the drilldown launches a new workspace that contains the results of all seriation algorithms applied only on the selected data. While considering Figure 5, we may also state, that if a set of variables and objects cause a certain pattern to emerge on many permutations, it would be of interest for the researcher to be focused on.

The time of calculation depends on the size of the dataset: for smaller datasets (less than 100 rows or columns) the results are displayed instantaneously. For moderate matrices (the number of rows and columns being between 100 and 1000), the permutations are displayed as soon as the corresponding calculation

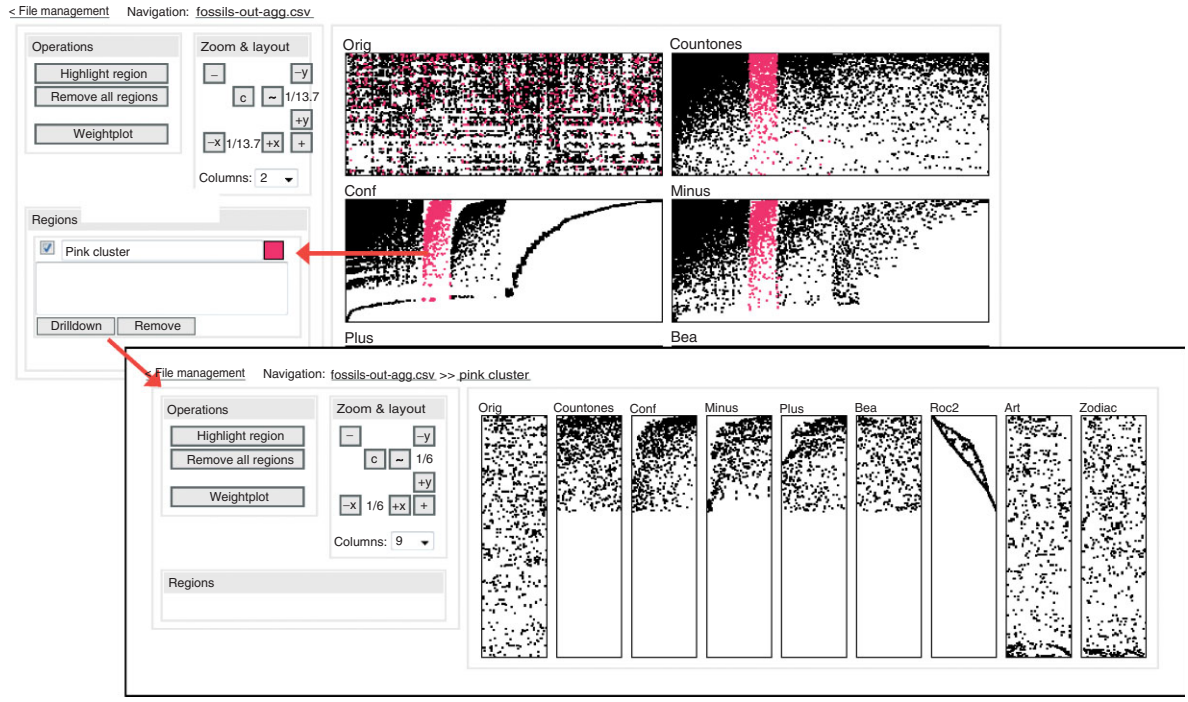


FIGURE 5 | Viewing the contents of the pink region—the results of a drilldown.

has completed. The inclusion of fast heuristic seriation methods provides user with a quick approximate overview, while more accurate albeit slower permutations are being processed. In the case of medium-sized datasets, ranging up to 5000 rows and columns, the computational complexity of seriation algorithms impacts the responsiveness of the drilldown operation, for most algorithms have time-complexity of $O(n^3)$ or more.

We have successfully used count of ones and conformity scale methods to seriate large binary matrices of approximately 1M rows by 1k columns, although experiments show that algorithms with time-complexity beyond $O(n^2)$ do not exhibit practically applicable execution times. As an example, the computation times for the Mammals dataset[32] described below (1597 rows and 3873 columns), vary from several seconds for simpler algorithms (*countones*, *roc*, *art*) to 20–25 min for cubic-time permutations (*plus, minus, bea*) on a virtual Intel® Pentium® G6950 2 GHz processor with 3.5 GB of memory.

The computation of permutations is handled server-side and the system is designed to launch several algorithms in parallel. The processing components can be distributed across several computers, which can include cloud-based computation services such as Amazon Elastic Compute Cloud. This architecture has relatively modest requirements for the client computer displaying the web-based user interface. However, the proposed tool is currently not suitable for visualizing large datasets because of prolonged data transfer times and limited client-side in-browser processing capabilities, being hindered mainly by browser's JavaScript engine speed.

All plots can be transformed to a spreadsheet format for detailed offline analysis. The exported files are also in the CSV format.

ILLUSTRATIVE EXAMPLES FROM DIFFERENT DISCIPLINES

In order to put forth two exemplary applications of working with VME, we are firstly going to turn to the field of psychology. The dataset depicted in Figure 6 is a classical example from Kelly's theory of Personal Constructs.[15] We are considering an interviewing technique called repertory grid, which allows a psychotherapist to identify semantic constructs of an interviewee by coming to a consensus on a common set of concepts. In order to construct the matrix, the psychotherapist enumerates a set of individuals which will range from those with concrete roles, like a mother or a friend, to abstract individuals, possessing certain values, like, for example, an ethical person. The interviewee is asked to formulate his personal constructs that are to be assigned to the figures. A construct is a contrasting or sometimes discordant pair of terms (as perceived by the

Figures																			Constructs	
Threatening person	Pitied person	Self	Mother	Happy person	Rejecting person	Ex-pal	Pal	Sister	Rejected teacher	Ethical person	Spouse	Brother	Attractive person	Boss	Ex-flame	Father	Successful person	Accepted teacher	Emergent pole	Implicit pole
			√	√						√		√		√			√	√	Achieved a lot	Hasn't achieved a lot
√	√	√	√	√	√		√		√	√		√	√	√	√	√	√	√	Believe in higher education	Not believing in too much education
					√			√							√				Don't like other people	Like other people
√		√		√	√		√		√	√	√	√	√					√	Think alike	Think differently
√		√			√	√		√			√	√	√			√	√	√	Not athletic	Athletic
			√								√		√		√				Both girls	A boy
		√	√	√			√	√	√	√		√	√			√	√	√	Both have high morals	Low morals
√	√				√	√	√	√							√				Both friends	Not friends
				√			√		√			√						√	Understand me better	Don't understand at all
				√				√	√			√		√			√		Same age	Different ages
				√			√		√	√				√			√	√	More understanding	Less understanding
											√		√		√				Both girls	Not girls
√				√	√										√		√		Don't believe in god	Very religious
							√	√	√			√		√			√	√	Higher education	No education
√	√		√	√	√							√		√				√	More sociable	Not sociable
√		√	√	√			√	√	√	√	√					√	√	√	More religious	Not religious
							√	√	√			√					√	√	Same sort of education	Completely different education
		√	√	√			√	√	√	√	√			√		√	√	√	Parents	Ideas different
	√		√	√				√	√	√				√		√	√	√	Believe the same about me	Believe differently about me
		√					√		√	√		√	√		√		√		Both appreciate music	Don't understand music
			√								√		√		√				Both girls	Not girls
			√						√	√		√		√		√		√	Teach the right thing	Teach the wrong thing

FIGURE 6 | An example repertory grid.

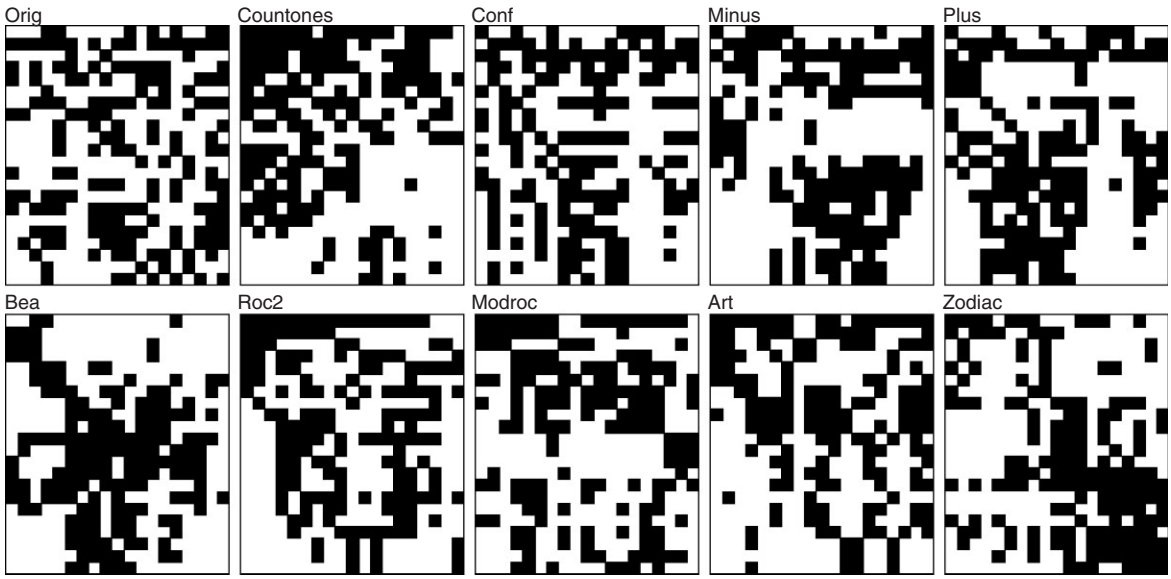


FIGURE 7 | Repertory grid visualized in multiple permutations.

interviewee), such as 'understanding person—ignorant person'. However, as the interviewee may have several meanings for the term 'understanding', it is up to the psychotherapist to identify the sets of opposite pairs—for example, to distinguish between 'understanding—ignorance' and 'understanding—disagreement'. The constructs will thereafter be applied to individuals, starting from concrete persons—family members and loved ones—and moving on to more abstract images of a threatening, attractive, or happy person. For a data analyst, Kelly's repertory grid technique can be considered as an interesting attribute discovery process for objects under investigation.

An example of the resultant table is depicted in Figure 6 which is fed into VME for seriation and visualization. In Figure 7, each tick of the grid is rendered with a black dot.

As we did in the case of Bertin's townships (Figure 2), we are now going to work with multiple

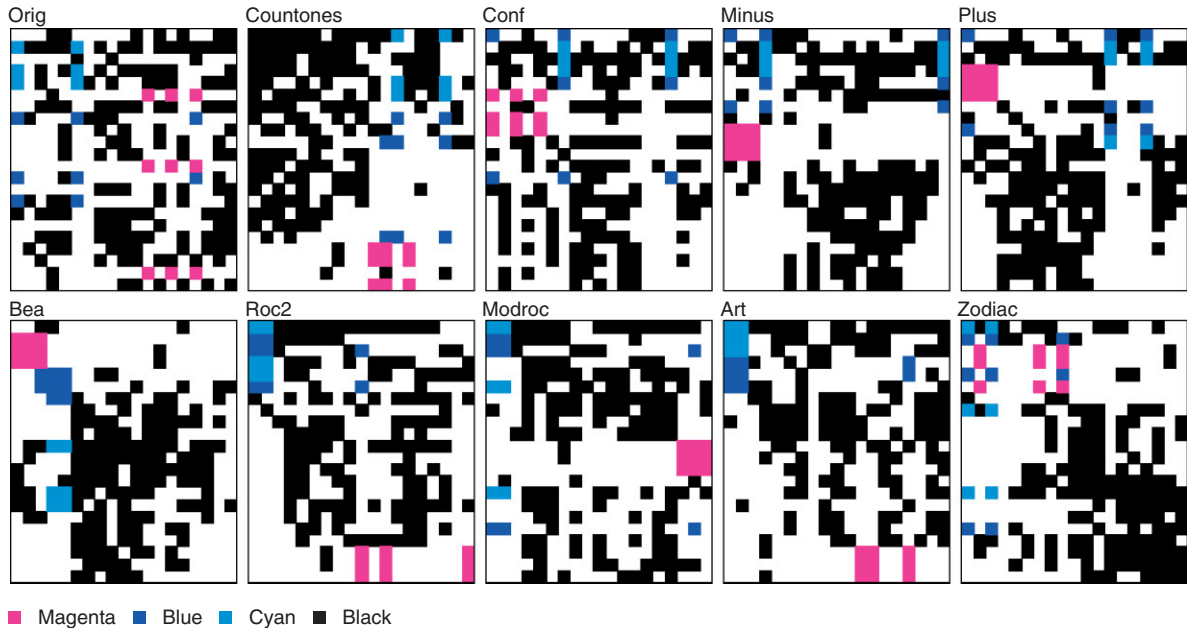


FIGURE 8 | Two highlighted patterns: magenta for a concept of loveable woman and blue for a rejecting-threatening person.

permutations. We will try to highlight relevant information therein, as relatedness is manifested in the adjacent placements by seriation algorithms.

In the *bea* permutation, the top-left corner contains two connected clumps (colored in magenta and blue in Figure 8), which in terms of cellular manufacturing can be described as sequential work cells. Consulting the attributes of the topmost cluster (in pink), it may be hypothesized to represent persons of opposite sex, who are of interest—a spouse, ex-flame, or simply an attractive person. All persons considered are female and will thereafter be marked 'loveable woman'.

Moving down from the pink cluster, the next clump is interesting. Ex-flame is placed side-by-size with an image of rejecting and threatening person. Constructs that are manifested by these roles are sociability, lack of belief in God and friendship with the interviewee. This cluster is marked with blue color in Figure 8. However, other permutations suggest including additional constructs for better comprehension of the concept 'rejecting person'.

In the *roc2* permutation, the aforementioned properties lie in a contiguous clump in the top-left corner. *Roc2* aims to solve the same cellular manufacturing problem as *bea*, however, the algorithm tries to organize the matrix in a block-diagonal form.[25] By expanding the blue-colored region, three additional descriptions of a rejecting person can be found: belief in a higher education, thinking like the interviewee and not being athletic. This operation of expanding a cluster in a parallel permutation, however well justified by the optimization algorithm's similarity measure, seems hard to vindicate in the psychological context. However, given the emotional and irrational nature of personal constructs, these controversial clusters provide a topic for a further interview.

On the *bea* permutation (Figure 9) a third, fairly solid block—that is located in the middle of the matrix—has been colored in violet. The individuals that take part in this pattern are successful person, happy person, and teacher. Looking at the represented constructs, achievements, belief in higher education, high morality, religiousness, and acceptance as a parenting role, are involved. Thus this person could be conceptualized as an 'ideal person'.

We have so far considered two processes—one being clarifying (by building roles and constructs) and the other being interpreting (e.g., by considering the permutation matrices). These can, however, be reciprocal. As the solid block mentioned above contains one inconsistency, the psychotherapist can now further investigate, whether great achievements are a prerequisite for someone to be considered a rejected teacher—as a correction would suggest.

By further trying to expand the current contiguous regions, we turn to the *zodiac* plot. The nearest similar blocks (in plum) can be attached to the violet region. This addition expands the interviewee's concept of an ideal person with two roles: mother and an ethical person. Note that this large cluster is unobtainable with a single rendering of the matrix, but in turn is seemingly justifiable in the context of our interpretation.

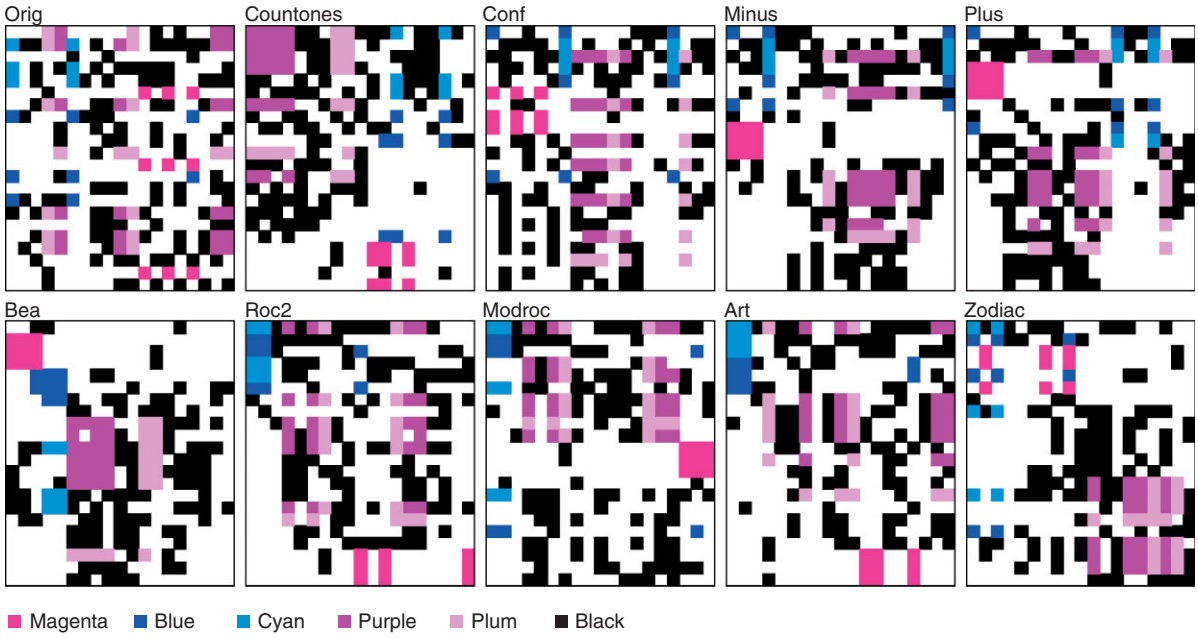


FIGURE 9 | Violet and plum regions representing a concept of an ideal person.

Secondly, we are going to consider an example from paleontology. The Neogene of the Old World[32] database contains taxa of land mammals in various localities across Europe. The original dataset includes a number of attributes: the taxon of the finding down to the species level, the geographical coordinates of the locality, an estimate on the size and the body mass of an animal, an assessment of the dietary habits of an animal, etc. (Figure 10).

The aggregated fossil matrix is much larger than the previous examples we have considered: roughly 3800 species mark the columns, while 1500 sites are on the rows of the matrix. Judging by the dimensions of the table, an investigative strategy of studying the findings of a certain species or a comparison of taxa between two sites would be too time-consuming. However, as visual clusters tend to be formed as a result of seriation, the aforementioned brush and drill-down techniques can aid in dividing the data table into smaller, clear-cut parts. A high variety of seriation algorithms implemented grants additional freedom: albeit certainly not being rigorously provable, the experience shows that at least one of the permutations usually yields some discriminating clusters.

In Figure 11, the most clear-cut seriation results are visible on *conf* and *plus* permutations (less tangibly on *minus*). The most typical findings are located in the corner of the table (in black), dissimilar species, in terms of distribution amongst localities, are colored.

Sites ↓ Species →	Artiodactyla->Anthracotheriidae-	Artiodactyla->Bovidae->Caprotragoides->stehlini	Artiodactyla->Bovidae->Eotragus->artenensis	Artiodactyla->Bovidae->Eotragus->clavatus	Artiodactyla->Bovidae->Hypsodontus->pronaticornis	Artiodactyla->Bovidae->Protragocerus->chantrei	Artiodactyla->Bovidae->Tethytragus->cf. koehleri	Artiodactyla->Bovidae->Tethytragus->koehlerae	Artiodactyla->Bovoidae->Amphimoschus->artenensis	Artiodactyla->Bovoidae->Amphimoschus->cf. Hispanomeryx->indet.	Artiodactyla->Bovoidae->Cainotheriidae->Cainotherium->miocaenicum	Artiodactyla->Cervidae->Euprox->furcatus	Artiodactyla->Cervidae->Euprox->minimus	Artiodactyla->Cervidae->Heteroprox->larteti	...
Lid:20001: Steinheim, Germay (48 42 0 N, 10 3 0 E)												1		1	
Lid:20002: Sansan, France (43 54 0 N, 0 30 0 W)				1										1	
Lid:20003: Pasalar, Turkey (40 0 0 N, 28 30 0 E)		1			1			1		1					
Lid:20004: La Grive St. Alban, France (44 0 0 N, 0 0 0 E)						1	1								
Lid:20006: Artenay, France (48 6 0 N, 1 54 0 E)	1		1						1						
Lid:20007: Bézian, France (44 12 0 N, 0 54 0 E)											1				
...															

FIGURE 10 | Findings of mammal species by location, a subset.

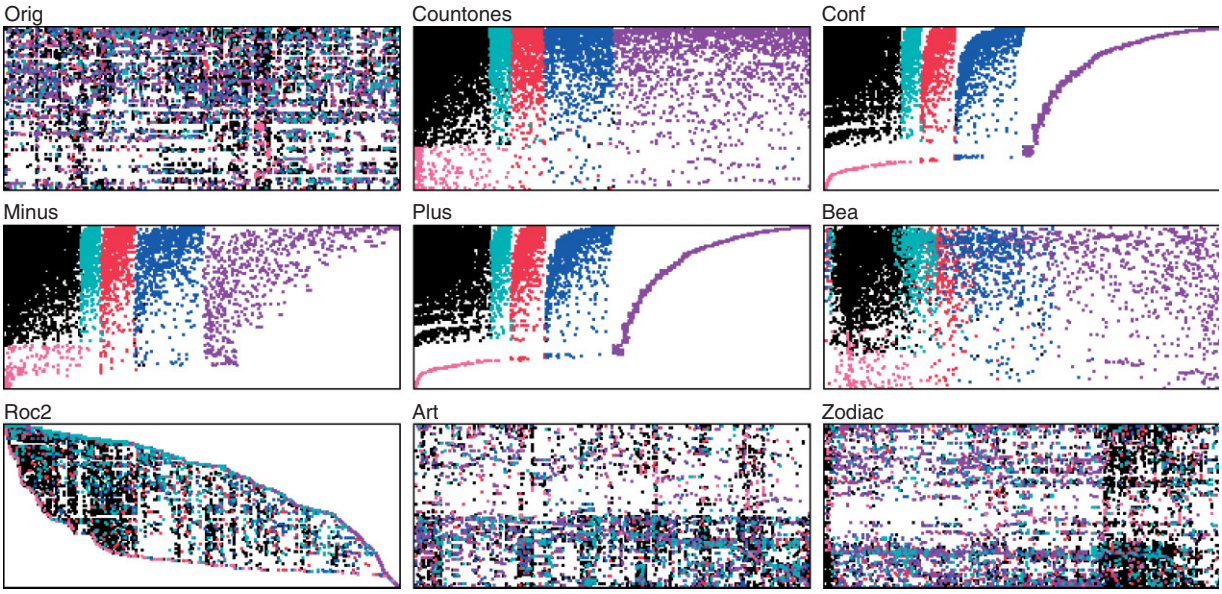


FIGURE 11 | A visualization of the fossils dataset, rows represent sites, columns mark species.

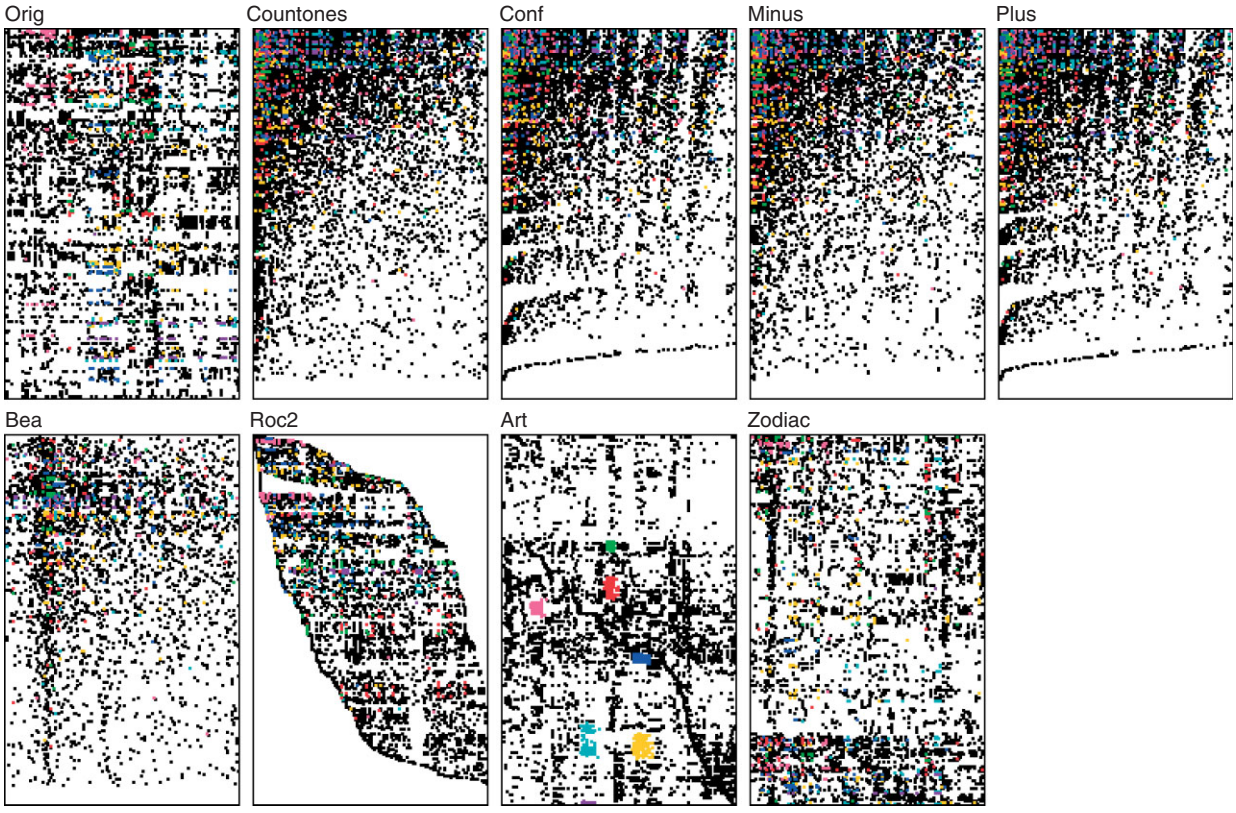


FIGURE 12 | Drilldown of the black dense region. Art and zodiac permutations introduce clumps that were not detectable on the original matrix.

According to the Monotone Systems meta-heuristic algorithms (*conf*, *plus*, *minus*), the dense black part contains the strongest influencers. Drill-down operation on that part reveals a fern-like clustering on the submatrix as well (Figure 12, plots *conf* and *plus*), thus constituting a fractal-like structure. As the submatrix has more degrees of freedom, cellular manufacturing algorithms art and zodiac are able to produce fairly solid rectangular clumps.

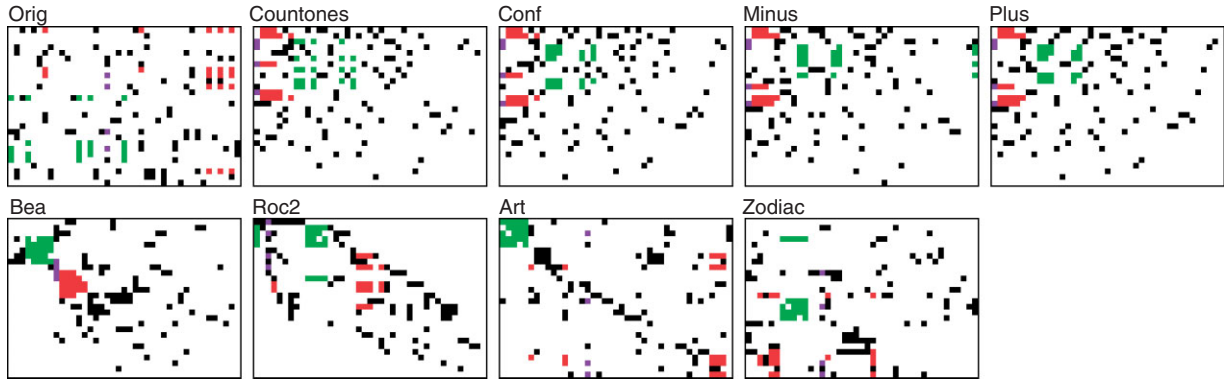


FIGURE 13 | Two sites of insectivores and rodents in Germany and France.

Figure 13 represents the blue-colored region of Figure 11. On the *bea* permutation we see clusterings with familiar shape (here colored in green and red), which are findings of several rodents and insectivores in France and Germany, the similarity of which should further be investigated. The purple line connecting the two clumps (in cellular manufacturing terms, a bottleneck), is an emphasized species (family *Soricidae*, with a common name shrew), being present at both sites.

Returning to the initial dataset in Figure 11, the loosely organized subset—in violet—can be considered as follows. The diagonal in the *plus* permutation features small connected blocks instead of a smooth curve—a hint that the datapoints can be divided into disjoint clusters of species across all sites. Drill-down of the violet region—in Figure 14—confirms the suggestion as several permutations present 5–6 disjoint clusters—a subject of discussion with fellow analysts.

CONCLUSION

This article presented a collaborative exploratory data analysis tool, VME, to analyze and compare different views of the same data. VME can also be considered as a tool to interact with seriation literature reviews and comparisons. There are a lot of tools for software visualization and information visualization, but, so far, no tools for comparing different scientific results, theories, assumptions, and biases. As an example, a bottleneck in cellular manufacturing seriation result may stand for a mediator in social network analysis.

VME also records the analytical activities and allows for web-based collaboration, as multiple users can operate on the same workspace, share comments and details of the exploration process.

We also emphasize collaboration on two different levels. First, that concerning the technical aspect of several people working simultaneously in a web-based environment. Second, allowing people

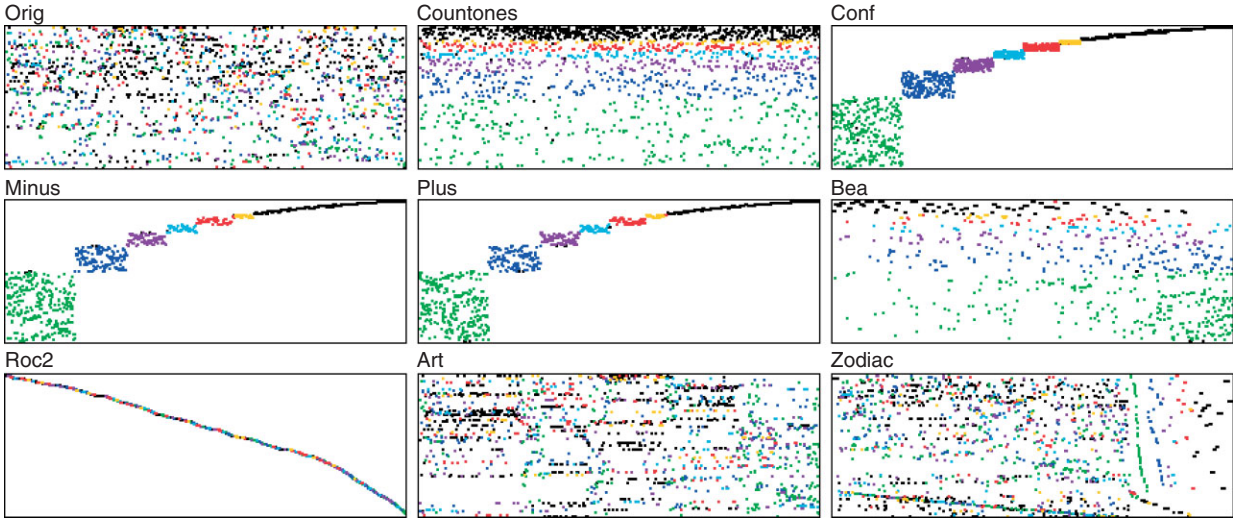


FIGURE 14 | Details of the comet-tail violet pattern in the initial matrix.

from different disciplines to work with the same dataset and to share a set of patterns, whereas maintaining their disciplines' traditional view.

In this article, VME has been applied to three datasets—those of Bertin's classical example of townships,[9] the theory of personal constructs in psychology[15] and European land mammals in paleontology.[32] Several interesting avenues of future research can be identified. Regardless of the analytics topic and task, one might be interested in logging the process of investigative data analysis. From such logs, when reasonably annotated and structured, it might be possible to extract a metaprocess for data exploration and collaborative data exploration. Certainly, a creative investigator can not be modeled, but most of the methodological repetitive steps a regular data investigator takes, could help to build a data mining, information visualization and interaction methodology using a bottom-up strategy.

VME is available as an open source project on SourceForge[33] for research purposes.

REFERENCES

1. Liiv I. Seriation and matrix reordering methods: an historical overview. *Stat Anal Data Mining* 2010, 3:70–91.
2. Inhelder B, Piaget J. *The Early Growth of Logic in the Child*. London: Routledge & Kegan Paul; 1964.
3. Becker RA, Cleveland WS. Brushing scatterplots. *Technometrics* 1987, 29: 127–142.
4. Belknap RE. *The List: The Uses and Pleasures of Cataloguing*. New Haven: Yale University Press; 2004.
5. Knight D. A Real alternative to spreadsheets. *Proc. of 2-nd Int. Symposium on Spreadsheet Risks, EUSPRIG2001*. Amsterdam, Holland, 2001.
6. Abraham R. End-User software engineering in the spreadsheet paradigm, Doctoral Thesis, Oregon State University, 2007.
7. Lenstra JK. Clustering a data array and the traveling salesman problem. *Oper Res* 1974, 22:413–414.
8. Niermann S. Optimizing the ordering of tables with evolutionary computation. *Am Stat* 2005, 59:41–46.
9. Bertin J. *Graphics and Graphic Information Processing* (Translated by Berg WJ and Scott P). Berlin: Walter de Gruyter; 1981.
10. Buchta C, Hornik K, Hashler M. Getting things in order: an introduction to the R package seriation. *J Stat Softw* 2008, 25:1–34.
11. Yi JS, Kang Y, Stasko JT, Jacko JA. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans Visual Comp Graph* 2007, 13:1224–1231 (Paper presented at InfoVis'07).
12. Pike WA, Stasko JT, Chang R, O'Connell TA. The science of interaction. *Inform Visual* 2009, 8:263–274.
13. Henry N, Fekete J, McGuffin MJ. NodeTrix: a hybrid visualization of social networks. *IEEE Trans Visual Comp Graph* 2007, 13:1302–1309.
14. Mueller C, Martin B, Lumsdaine A. A comparison of vertex ordering algorithms for large graph visualization. *Proceedings of Asia-Pacific Symposium on Visualization*. Sydney, Australia, 2007, 141–148.
15. Kelly GA. *The Psychology of Personal Constructs*. New York: Norton; 1955.
16. Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman; 1963, 11.
17. Burbidge JL. Production flow analysis. *Prod Eng* 1963, 42: 742–752.
18. McCormick WT, Schweitzer PJ, White TW. Problem decomposition and data reorganization by a clustering technique. *Oper Res* 1972, 20:993–1009.
19. Wehrend S, Lewis C. A problem-oriented classification of visualization techniques. *Proceedings of VIS'90* 1990, 139–143.
20. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings of 1996 IEEE Conference on Visual Languages*, 1996, 336–343.
21. Amar R, Stasko J. A knowledge-task based framework for design and evaluation of information visualizations. *Proceedings of InfoVis 2004*, 2004, 143–149.
22. Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization. *Proceedings of IEEE InfoVis'05*, Minneapolis, 2005, 111–117.
23. Mullat JE. Extremal subsystems of monotonic systems I. *Autom Remote Cont* 1976, 37:758–766.
24. Vyhandu L. Some methods to order objects and variables in data systems. *Trans of Tallinn University of Technology* 1980, 482:43–50.
25. King JR, Nakornchai V. Machine-component group formation in group technology: review and extension. *Int J Prod Res* 1982, 20:117–133.
26. Chandrasekharan MP, Rajagopalan R. MODROC: an extension of rank order clustering for group technology. *Int J Prod Res* 1986, 24:1221–1233.
27. Kaparthi S, Suresh NC. Machine-component cell formation in group technology: a neural network approach. *Int J Prod Res* 1992, 30:1353–1367.
28. Kusiak A, Chung YK. GT/ART: using neural network to form machine cells. *Manufact Rev* 1991, 4:293–301.
29. Chandrasekharan MP, Rajagopalan R. ZODIAC—an algorithm for concurrent formation of part-families and machine-cells. *Int J Prod Res* 1987, 25:835–850.

30. Jerding DF, Stasko JT. The information mural: a technique for displaying and navigating large information spaces. *IEEE Trans Visual Comp Graph* 1998, 4:257–271

31. Codd EF, Codd SB, Salley CT. *Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate*. San Jose: Codd & Date; 1993.

32. Fortelius M. Neogene of the Old World Database of Fossil Mammals (NOW public release 030717). University of Helsinki, 2003. Available at: http://www.helsinki.fi/science/now. (Accessed July 2, 2011).

33. Visual Matrix Explorer source code. Available at: http://sourceforge.net/projects/vismatexplorer. (Accessed July 2, 2011).