Vision, Modeling, and Visualization (2014)
J. Bender, A. Kuijper, T. von Landesberger, H. Theisel and P. Urban (Eds.)

# Hierarchical Brushing of High-Dimensional Data Sets Using Quality Metrics

G. Albuquerque, M. Eisemann, T. Löwe and M. Magnor

Computer Graphics Lab, TU Braunschweig, Germany

**Abstract**

*In this paper, we present an interactive exploration framework that puts the human-in-the-loop with the application of quality metrics and brushing techniques for an efficient visual analysis of high-dimensional data sets. Our approach makes use of the human ability to distinguish interesting structures even within very cluttered projections of the data and uses quality metrics to guide the user towards such promising projections which would otherwise be difficult or time-consuming to find. Brushing the data creates new subsets that are ranked again using quality metrics and recursively analyzed by the user. This creates a human-in-the-loop approach that makes use of hierarchical brushing and quality metrics to support interactive exploratory analysis of high-dimensional data sets. We apply our approach to synthetic and real data sets, demonstrating its usefulness.*

## 1. Introduction

Visual exploration of high-dimensional data involves visualizing complex data sets in lower-dimensional projections, such that properties, relationships and functional dependencies in the data may be revealed. One of the most common representations for high-dimensional data sets is the *scatterplot matrix* (SPLOM), a matrix of pairwise projections for all dimensions of the data set [CLNL87, Cle93]. While SPLOMs are an intuitive approach to visualize data sets with more than three dimensions, higher-dimensional structures are lost this way. This issue has been partially alleviated by introducing *linking* and *brushing* techniques [McD82, BC87] which allow analysts to manually select samples in one chosen scatterplot and to inspect these samples highlighted in other projections. Brushing has since proven to be very useful for revealing multivariate interdependencies that would otherwise be practically invisible in matrices of bivariate projections. Unfortunately, the process of actually finding projections where brushing helps to gain knowledge about the data becomes more and more difficult as the number of dimensions in the data set increases.

*Quality metrics* (QMs) [BTK11, TAE*09] are automated approaches to support the exploratory visualization of high-dimensional data sets. They are sometimes also called indices [WAG05] or ranking functions [SS05], and can be used to appraise the value of a projection according to a specific criterion or user task. The idea of QMs is to rate each projection by its relevance and then reordering or highlighting projections and dimensions accordingly, limiting the number of projections that have to be investigated manually. A current drawback of most QMs is that they are computed on the projections of the whole data set. Projections considered to contain only noise may, however, contain important information if only subsets of the data are considered. Selecting these subsets can be arbitrarily difficult and often requires manual intervention.

In this paper, we introduce an interaction loop that iterates between applying quality metrics and brushing to extract meaningful subsets of the high-dimensional data. We present the prototype of an extensible framework that combines the interactivity of brushing with the established methods of quality metrics and regression analysis. We create a visual analytics loop that closely follows the widely accepted guide to visually explore high-dimensional data [KMSZ06, Shn96]: "analyze first, show the important, zoom and filter, analyze further, show details on demand". This is achieved by first applying quality metrics to give an overview of the information content of the 2D projections inside a SPLOM. The user may then zoom into relevant plots, select interesting structures by brushing, and repeat the process on the selected subset until all important structures are



DOI: 10.2312/vmv.20141284

found. Additionally, mathematical models may be derived from these selections using *Regression analysis* [FW06] and visualized together with its confidence levels. To evaluate our approach, we consider real and synthetically generated high-dimensional data sets to highlight the benefits and use of our approach.

## 2. Related Work

Exhaustive exploration of high-dimensional data sets [Asi85] quickly reaches its limits as the number of possible projections often grows exponentially with the number of dimensions. Quality metrics aid the user in finding interesting patterns and information in high-dimensional data by ranking the different projections according to different exploration tasks. From the early work of *Projection Pursuit* [FT74, Hub85] over the well known *Scagnostics indices* [TT85, WAG05] on to more recent metrics [BTK11, AEL*10, SNLH09, TAE*09, SSK06] QMs have become a well established field of research covering a large number of different visualization techniques and user tasks. All these measures can be used to support the visual analysis of high-dimensional data sets, e.g. by color-coding [SS05, AEL*09, LAE*12]. Commonly, they are applied in a preprocess, but this hinders their usage for interactive exploration beyond simple filtering and adaptive layouts [DW13, JJ09]. On-the-fly usage and re-evaluation of QMs in an interactive setting, as we propose in our framework has not been investigated yet, to the best of our knowledge.

The selection of subsets of a multidimensional data set using *linking* and *brushing* techniques [McD82, BC87] is a common approach used in many visual exploration tools [STBC03, War94]. Inspired by brushing and linking, recent sketch-based approaches allow not only the exploration of multidimensional data sets, but also support the creation of synthetic data sets using different visualization methods to navigate in the multidimensional space [WRM13, ALM11]. Motivated by these techniques, we propose a visual exploration framework that allows for testing two-dimensional hypotheses using sketches. These sketches select interesting subsets of a data set and can then be used to find suitable models to describe the data in abstract form.

A common way of retrieving a mathematical description of given data is *regression analysis* [FW06]. In some cases the data is not well described by a single function, e.g. if the data contains intrinsic semantic clusters. To derive a suitable mathematical model of the dependencies a separation into semantic clusters is necessary beforehand. While spatial clustering is a well researched area, including e.g. k-means [Llo82], spectral clustering [NJW01], or nonlinear embedding techniques [CC10, vdMH08], semantic clustering within a single projection without prior knowledge still requires human interaction and understanding. This human interaction in the visual analysis process is an integral part of our proposed approach.

## 3. Overview

Exploratory visualization of high-dimensional data sets, where the user has no hypotheses about the information hidden in the data, is one of the main challenges of information visualization [KMSZ06]. This exploration process usually includes mapping the high-dimensional data in lower-dimensional embeddings to make the data amenable to the human visual system. Quality metrics are used to rank or reduce the combinatorically large amount of possible views to be analyzed. In this paper, we propose to close this visual analysis to a loop that allows to interactively analyze subspaces of high-dimensional data sets.

Our framework consists of three main parts, Figure 1. Firstly, the user may use QMs as ranking functions to quickly find dimensions or projections of the data with non-random structures (Section 4). In the second step the user chooses one or more of the projections in a SPLOM and sketches a prior model over the structures by selecting (brushing) data points (Section 4.1). In addition to the SPLOM visualization, a Multidimensional Scaling (MDS) [CC10] scatterplot of the data set is also available to help brushing and analyzing multidimensional structures. The selected points are marked in all projections, providing instant visual feedback. Different colors mark different groups of points. The resulting selection can then again be analyzed by QMs and further refined in an interactive exploration loop (Section 4.2). Finally, in the last step, an appropriate model may be fitted to sketched selections and visualized (Section 5), e.g. a two-dimensional Gaussian or polynomial functions. The visualization of this model includes additional information on its reliability which can be used in a next loop to refine the selection.

**Notation** Let $D \subseteq \mathbb{R}^n$ be a high-dimensional data space consisting of $N$ data points $p^i = (p^i_1, \ldots, p^i_n)$ with $i \in \{1, \ldots, N\}$. A view (projection/embedding) is a 2D orthogonal embedding $\pi_x \times \pi_y(D)$ of all $p^i$ to the $(x,y)$ coordinates with $\pi_x \times \pi_y(p^i) = (p^i_x, p^i_y)$, where $x, y \in \{1, \ldots, N\}$. For simplicity of discussion we omit the index parameter $i$ if not specifically required and use $x$ and $y$ as general indices of the projection axes when talking about a single 2D embedding. The displayed data in the 2D embeddings is normalized to make full use of the given image space.

## 4. The Quality Metrics Loop

Since the analyst might be interested in different aspects of the data set, choosing a user task and the appropriate quality metric is part of the exploration process. Different quality metrics can be added to our framework using a plug-in mechanism. We currently support the well-known Scagnostics indices [WAG05] together with the Class Density Measure (CDM) from [TAE*09].

Nine graph-theoretic Scagnostics measures for scatterplots were defined in [WAG05] to detect *outliers*, *convex,*

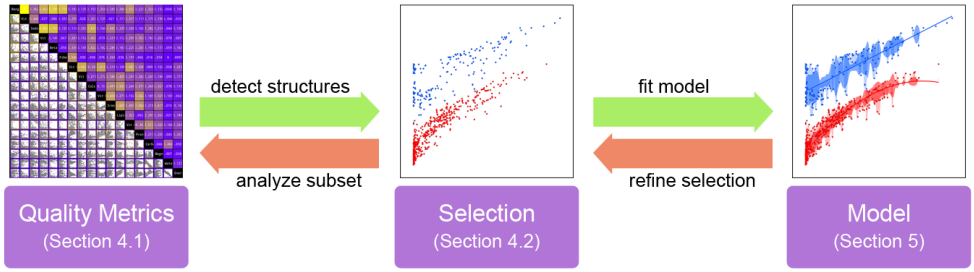



Figure 1: **Overview:** Our approach consists of two interaction loops intertwining automatic QMs with interactive brushing and automatic regression analysis.

*skinny, stringy, straight, skewed, clumpy, striated* or *monotonically* changing structures. The Class Density Measure (CDM) [TAE*09] rates how well the different classes of classified data are separated in the different views. Class labels may be previously known, or in our case, user-selected classes can be interactively assigned to the samples of the data set using brushing.

### 4.1. Sketch-based Selection

When working with low-dimensional projections, higher-dimensional information is often lost. As a result, these projections are not always made up of exactly one statistical structure that can be properly represented using a single one-dimensional function. They may often contain several overlapping sub-populations that only become distinguishable in a higher-dimensional context. Brushing techniques are a popular approach for identifying and separating such structures. Figures 2a and 2c show an example of the importance of this user-guided structure separation.

Addressing this issue, our system allows organizing data samples into an arbitrary number of color-coded sets using brushing. Samples are simply colored directly in the SPLOM and/or MDS view and updated across all projections. The framework supports different brush sizes and interactive zooming in the SPLOM to reveal details or to facilitate brushing the data samples. The brushing is not restricted to a single projection but can be refined by brushing in all the available projections until the user is satisfied. Extracted subsets may then again be recursively partitioned into new sets using the same brushing technique. This allows separating a complex data set into its most basic sub-populations.

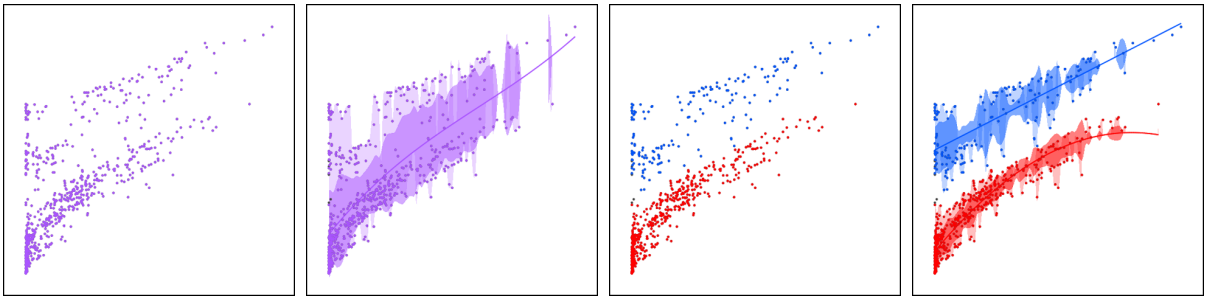
### 4.2. Interactive Quality-Metrics Loop

With increasing dimensionality of the data it becomes more difficult to find interesting projections within a SPLOM. Additionally, previously uninteresting projections may change to information-bearing projections when only a subset of the data is considered. To support the user in this high-dimensional data exploration process we incorporate QMs into an interactive feedback loop. The user may choose to analyze the currently active data subset using QMs at any time during the exploration process. The metric values are computed and displayed in upper right half of the SPLOM using a heat map. This map goes from dark purple for low values to yellow for high values of the QMs. Furthermore, to improve the visualization of the QMs ranking, we draw a small border around each of the scatterplots with the respective color, which eases orientation when zooming into the SPLOM, Figure 3. To further increase the readability of large SPLOMS we offer the possibility to automatically reorder the dimensions within the SPLOM to cluster highly ranked plots, similar to [WAG05, LAE*12]. Once the user selects a subset, QMs may be applied again to the newly defined subset, since previously unimportant projections now potentially contain structures of interest.

**Color Mapping** Additionally, our framework allows to change the mapping function of the QMs to the displayed color. This becomes important, e.g., when for a specific QM, few projections receive a high value while all the remaining receive comparably low values, Figure 3a. In our framework, the user may set an upper threshold that clamps the QMs values. The remaining values are linearly distributed among the color map, as e.g. in Figure 3b. This is especially important as the projection on which the brushing was applied often has the highest QM value in the next iteration of the loop suppressing the others. At any time the user may switch from this first visual analytics loop to the second to derive an abstract mathematical model from the selected data, Figure 1.

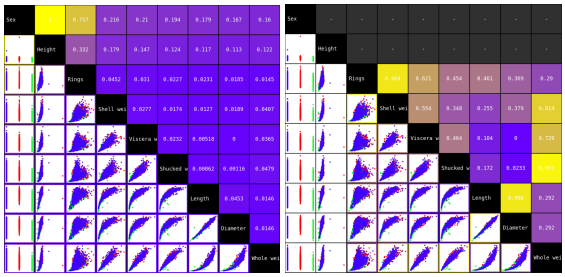
## 5. Structure Modelling

Once the user has found interesting structures using the brushing/QM loop, our framework offers further help to deduce a mathematical model from the selected data and to visualize it in a form that eases understanding and offers means to deduce the reliability of the model. Our framework currently supports the derivation of a model for two-dimensional curve-like structures and clusters. This model is defined by two main parts: A *structure model* that provides



(a) (b) (c) (d)

Figure 2: **Structure Modelling:** A demonstration of how brushing can be used to enhance fitting results and to help derive a suitable mathematical model. (a) shows an orthogonal data projection taken from the USDA Food Composition data set, with (b) being the corresponding result of a naive *polynomial* fitting. (c) shows the brushed data; the structures have been separated, so (d) exhibits two different regions and fitting curves, the blue using a *linear* and the red using a *logarithmic* prototype function. The darker areas within the structures mark the standard deviation from the model in a local window, whereas the lighter areas represent the maximum deviation.



(a) (b)

Figure 3: **QM Thresholding:** (a) Linear mapping of the QM to a color map. (b) Thresholding and scaling of the QM values before mapping them to a color helps to distribute them more efficiently among the color map. Values above the threshold are grayed out.

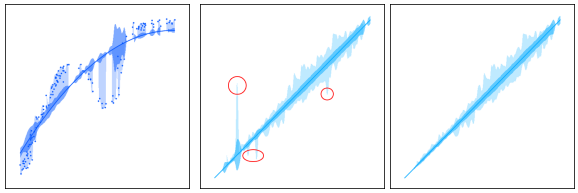
a general classification of the data distribution as e.g. logarithmic, linear or polynomial correlation between the dimensions, or the mean position of a cluster. We plot the model as a thin line/dot in the visualization; And second, a *deviation model* that is plotted as solid graphs to depict the local standard and maximum deviation of the data from the *structure model*. The extensibility of our framework makes the models exchangeable with other global or local regression techniques if required [Cle79]. Examples are given in in Figures 2b and 2d.

### 5.1. The Structure Model

The structure model describes the general trend of the data in the 2D $xy$-embedding by fitting a user chosen prototype function $f(p_x, a)$ to it using non-linear optimization [Mar63] where $a = (a_0, \ldots, a_n)$ is a set of model parameters of the prototype function. We currently support, but are not limited to, the following set of prototype functions representing logarithmic, polynomial, exponential and periodic behavior:

$$
\begin{aligned}
f_{\log}(x,a) &= a_0 + a_1 \log(a_2 + x) \\
f_{\text{poly}}(x,a) &= \sum_{i=0}^{n} a_i x^i, \quad n \in \{1,\ldots,5\} \\
f_{\exp}(x,a) &= a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x} \\
f_{\text{per}}(x,a) &= a_0 + a_1 \sin(a_2 + a_3 x) + a_4 \cos(a_5 + a_6 x)
\end{aligned}
$$

These have already been proven useful in the context of data retrieval [SBS11]. In addition, we added a 2D Gaussian function to describe clusters.



(a) (b) (c)

Figure 4: **Local Data Variation and Outlier Removal:** (a) The structure function matches the global data distribution well but lacks to represent the local variations. The deviation model makes the viewer aware of this fact by plotting the local standard (dark blue) and maximum (light blue) variance along the curve. (b) Outliers to a structure model appear as visible spikes in the visualization (marked with a red circle). (c) Result after removing the outliers in (b) before computing the structure model.

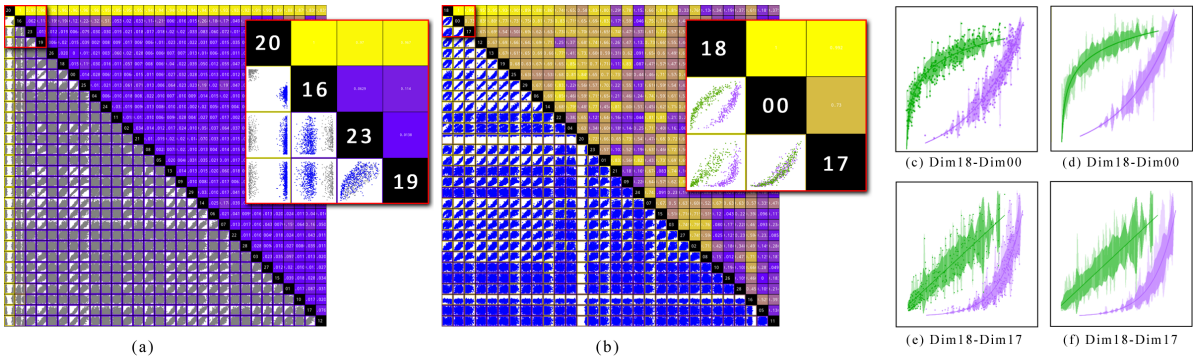



Figure 5: **Synthetic data set:** (a) Overview of the data set, ranked by the *Clumpy* metric. Inset: Selecting a subset of the data set (highlighted in blue). (b) Overview of the isolated subset, ranked by the *Skinny* metric and reordered to cluster best results. Inset: Dim18-Dim0 reveals previously hidden clusters that can be separated using brushing. We fit different model functions to the selected structures. (c) Between Dim18-Dim0, a logarithmic function is fitted for the green selection and an exponential function to the purple selection. (d) Resulting model for Dim18-Dim0 without the sample points. (e) Between Dim18-Dim17, a line is fitted to the green selection and a exponential function to the purple selection. (f) Resulting model for Dim18-Dim17 without sample points.

### 5.2. The Deviation Model

In addtion to the structure model (Section 5.1) we make use of classic deviation models [Cle79] giving the analyst a visual feedback on the local variance of the data and how well it fits the prototype function. Figure 4 shows an example where the overall fitting quality is good but varies drastically on a local level.

We superimpose the curve of the structure model with solid shapes representing the local *standard deviation* $\sigma$ and *maximum deviation* $\sigma_{max+}$ and $\sigma_{max-}$ from the structure model in positive and negative $y$-direction, respectively. The values can be computed using standard local regression techniques [Cle79]. For the special case of fitting a 2D Gaussian, the shapes represent the standard deviation and two times the standard deviation. The final visualization gives a clearer visual feedback on the local reliability of the chosen structure model and removes the visual clutter of the original scatterplot.

### 5.3. Outlier Removal

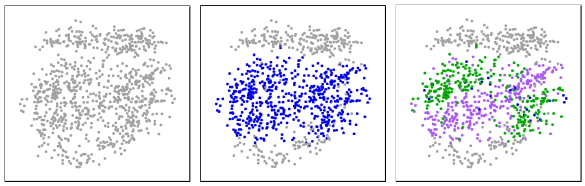
An interesting feature of this visualization is that outliers to the structure model appear as clearly visible spikes in the local standard and maximum deviation, Figure 4(b). While outliers are often indications of errors in data acquisition and processing, they may also represent real instances of extreme divergence. In either case, noise and outliers can have a severe negative impact on automated regression models. We allow the user manually deselect outliers to adjust the model according to his/her needs, Figure 4(c). Moreover, users are not limited to creating a single fitted model per structure, but may instead create an arbitrary number of models for a number of variations of the same structure.

Brushing the data can have a strong effect on the underlying model of the data. Therefore, the analyst has to be aware and very careful during selection and must always have his/her goal in mind. E.g. removing outliers is helpful if one wants to know the behavior of the data in general. One example could be the removal of defective parts in an assembly line [SSK06] to get a model of the correctly produced parts.

## 6. Use Cases

In this section we evaluate the analytical capabilities of our sketch-based framework. First, we use a synthetic data set to exemplify the suggested workflow. We then show a real-world application.

**Synthetic Data Set** We use a synthetic data set containing 30 dimensions and 900 samples generated with the tool presented in [ALM11]. Visual analysis starts by first displaying the entire data set using a SPLOM. We then apply the scagnostics quality metrics [WAG05] to all scatterplots. Figure 5a shows an overview of the data set, ranked by the *Clumpy* metric. We can observe well-defined clusters in the scatterplot (Dim20-Dim16) with the highest value for the *Clumpy* metric that is marked in yellow. We brush the right cluster (blue) and apply the *Skinny* metric to the selected data subset. The best-ranked scatterplot for the *Skinny* metric is the scatterplot between Dim18-Dim0. In this scatterplot we can visualize two well defined clusters that are brushed with purple and green, Figure 5b inset. Figure 6 shows the MDS visualization for each of the selections steps. In a last iteration we use these user selections as classes for the CDM metric [TAE*09] revealing views that best separate these selections and guiding us to other structures across the SPLOM,



(a) (b) (c)

Figure 6: **Synthetic data set:** (a) MDS view of the the entire data set. (b) MDS view of the synthetic data set showing the first selection (blue). (c) MDS view of the synthetic data set showing the last selections (green and purple).

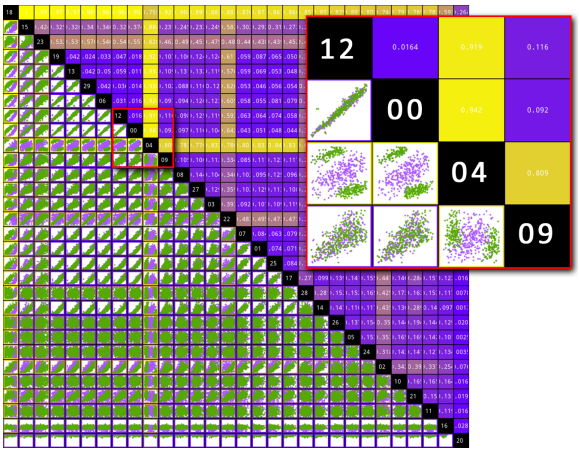



Figure 7: **Synthetic data set:** Overview of our synthetic data set evaluated by the *CDM* metric using the three selections as classes (green and purple). The best ranked scatterplots show the projections that better separate the classes.

as can be seen in Figure 7 inset. The selections can then be used to fit a desired model function. Figures 5c-f show models that were fitted to the selections in scatterplots Dim18-Dim0 and Dim18-Dim17.

**USDA Food Composition Data Set** In our second test case we use the USDA data set from [TMF*12], which contains different nutrients from a collection of food samples. It is composed of 722 records(foods) with 18 attributes, including diverse nutrients of samples such as Vitamin D, Vitamin B12, Iron, Carbohydrate, Lipid, Protein, Water, etc. We begin the analysis of the data by first computing the scagnostics metrics over all two-dimensional projections and visualizing the results. Figure 8a shows the result for the *Clumpy* metric, where we observe a few scatterplots with higher quality values than others. The plot displaying *Manganese* and *Vit C* has the highest value for this metric. It the shows two rough clusters, one in horizontal and one in vertical direction, from which we select the bottom one by brushing (green), Figure 8a inset.

Applying the scagnostics metrics again only on the selected subset and visualizing the results for the *Monotonic* metric, we see high-ranked plots that show correlations between the Water, Energy, Carbohydrate and Lipid dimensions, Figure 8b. We brush a second cluster (red) in the plot displaying *Water* and *Energy* that has the highest value for the *Monotonic* metric, Figure 8b inset. Figure 9a shows the MDS projection of the data set at this step. We then select the remaining points (blue, Figure 9b) and apply the CDM metric to the selected data, treating each of the three selections as a separate class, Figure 10.

Using both SPLOM and MDS views in parallel, we have a better control about selections, and interesting structures that may be hidden in orthogonal views can be found as well. Figure 8c-f shows two of the models that we fitted to two of the best ranked scatterplots. Between the dimensions Energy-Water projection two polynomial functions of second order are fitted to the red and green selections, respectively, and for the Energy-Lipid projection a line is fitted to the red selection and a polynomial function of second order is fitted to the green selection. Looking at the data marked in red in the inset of Figure 8b and 10, it shows typical food behavior. The more water it contains the less energy, the more lipid the more energy, and the more lipid the less water. These relationships become clearer in the selected data subset than in the complete data set, as it is not entirely true for samples with a low amount of water, see the green selection in Figure 10.

## 7. Conclusion and Discussion

In this work we have presented a quality metrics- and regression analysis-based human-in-the-loop visual analytics approach to find relevant information content in high-dimensional data sets. Color coding based on QMs guides the analyst to interesting projections within a SPLOM. Our interactive brushing, structure modelling, and visualization approach helps in selecting interesting substructures and to derive a suitable model description of the underlying data. The visualized model can then be used to detect and remove outliers or to create a feedback loop with the QMs to guide the further extraction and data analysis.

**Limitations and Extensions** Our proposed framework is easily extensible, e.g. adding more prototype functions or multidimensional regression analysis is straightforward in our current framework and can be easily added as additional prototype functions. The same is true for Gaussian processes or non-parametric models.

The QMs do not enforce a certain behaviour, they simply give hints on which projections appear interesting. The usefulness of QMs in general has been studied by many researchers and we rely on their expertise. The same holds for brushing and subset selections. Additionally, brushing and

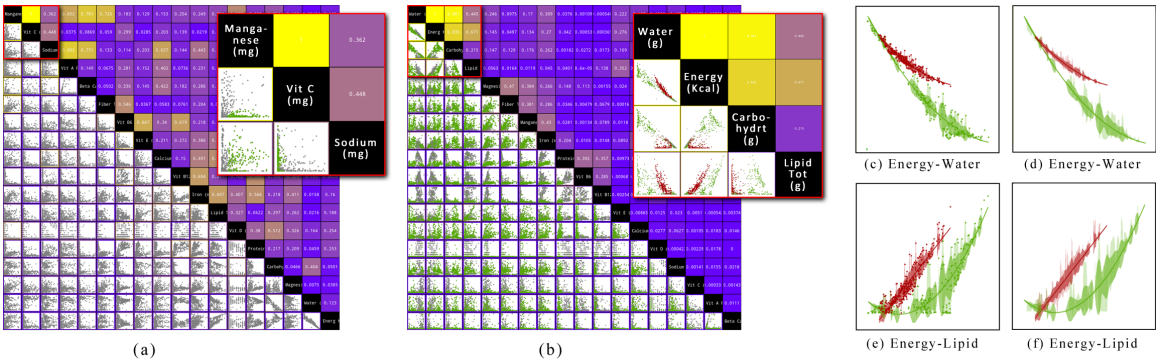



Figure 8: **USDA Food Composition Data Set:** (a) Overview of the food data set using the *Clumpy* metric. A subset of the data is selected for the scatterplot displaying the Manganese-Vit C dimensions, shown in the inset in (a). (b) Highly ranked plots (marked yellow) using the *Monotonic* metric on the selection show correlations between the Water, Energy, Carbohydrate and Lipid dimensions. (c) Two polynomial functions of second order are fitted to the red and green selections in the Energy-Water plot. (d) Resulting model for Energy-Water plot without the sample points. (e) A line is fitted to the red selection and a polynomial function of second order is fitted to the green selection in the Energy-Lipid plot. (f) Resulting model for Energy-Lipid without sample points.

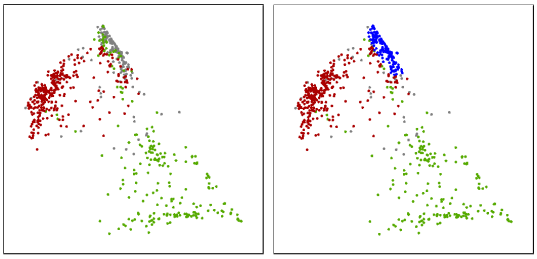



Figure 9: **USDA Food Composition Data Set:** (a) MDS view of the food data set showing the first selections (green and red). (b) The remaining subset of the data is selected in the MDS view (blue).


Figure 10: **USDA Food Composition Data Set:** Overview of the food data set evaluated by the *CDM* metric using the three selections as classes (green, red and blue). The best ranked scatterplots show the projections that better separate the classes.

subset selection requires visual analytics experts as otherwise arbitrary relations could be brushed into and extracted from the data. Our framework is, however, the first to combine brushing and QMs in a generic, extensible, and interactive framework.

Lastly, there are always technical improvements to be made. Methods to automate the selection of an appropriate prototype function may be further investigated as this will become more important as their number grows. There may also be ways to automate the process of outlier elimination—as discussed in Section 5.3—to a reasonable extent. Additionally, an automatic selection of the best suited QMs may be implemented to guide the user to the most evident trends in the data. Based on the insights gained from this work we would like to extend our framework to incorporate additional QMs and visualizations. For this purpose we plan to make our framework available in the public domain.

## 8. Acknowledgments

The authors gratefully acknowledge funding by the German Science Foundation from project DFG MA2555/6-2 within the strategic research initiative on Scalable Visual Analytics. We also want to thank the anonymous reviewers for their many valuable suggestions.

## References


[AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M.: Quality-based visualization matrices. In *Proc. Vision, Modeling and Visualization (VMV) 2009* (2009), pp. 341–349. 2

[AEL*10] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M.: Improving the visual analysis of high-dimensional datasets using quality measures. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)* (2010). 2

[ALM11] ALBUQUERQUE G., LÖWE T., MAGNOR M.: Synthetic generation of high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics (TVCG, Proc. Visualization / InfoVis) 17*, 12 (2011), 2317–2324. 2, 5

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143. 2

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing scatterplots. *Technometrics 29*, 2 (1987), 127–142. 1, 2

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2203–2212. 1, 2

[CC10] COX T., COX A.: *Multidimensional Scaling, Second Edition*. Taylor & Francis, 2010. 2

[Cle79] CLEVELAND W. S.: Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association 74* (1979), 829–836. 4, 5

[Cle93] CLEVELAND W.: *Visualizing data*. AT&T Bell Laboratories, 1993. 1

[CLNL87] CARR D. B., LITTLEFIELD R. J., NICHOLSON W. L., LITTLEFIELD J. S.: Scatterplot matrix techniques for large n. *Journal of the American Statistical Association 82* (1987), 424–436. 1

[DW13] DANG T. N., WILKINSON L.: Scagexplorer: Exploring scatterplots by their features. In *IEEE PacificVis* (2013), pp. 1–10. 2

[FT74] FRIEDMAN J., TUKEY J.: A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on C-23*, 9 (1974), 881 – 890. 2

[FW06] FREUND R. J., WILSON W.: *Regression Analysis*. Academic Press, 2006. 2

[Hub85] HUBER P. J.: Projection pursuit. *The Annals of Statistics 13*, 2 (1985), 435–475. 2

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 993–1000. 2

[KMSZ06] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in visual data analysis. In *In Proceedings of the Tenth International Conference on Information Visualization* (2006), pp. 9–16. 1, 2

[LAE*12] LEHMANN D. J., ALBUQUERQUE G., EISEMANN M., MAGNOR M., THEISEL H.: Selecting coherent and relevant plots in large scatterplot matrices. *Computer Graphics Forum 31*, 6 (2012), 1895–1908. 2, 3

[Llo82] LLOYD S.: Least Squares Quantization in PCM. *IEEE Transactions on Information Theory 28* (1982), 129–137. 2

[Mar63] MARQUARDT D.: An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics 11*, 2 (1963), 431–441. 4

[McD82] MCDONALD J. A.: *Interactive graphics for data analysis. oai:cds.cern.ch:146591*. PhD thesis, Calif. Univ. Stanford, Stanford, CA, 1982. Presented on Aug 1982. 1, 2

[NJW01] NG A. Y., JORDAN M. I., WEISS Y.: On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14* (2001), MIT Press, pp. 849–856. 2

[SBS11] SCHERER M., BERNARD J., SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (2011), ACM, pp. 363–372. 4

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (1996), VL '96, pp. 336–343. 1

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis 2009) 28*, 3 (2009), 831–838. 2

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidinemsional data. *Information visualization 4*, 2 (2005), 96–113. 1, 2

[SSK06] SCHNEIDEWIND J., SIPS M., KEIM D.: Pixnostics: Towards measuring the value of visualization. *Symposium On Visual Analytics Science And Technology 0* (2006), 199–206. 2, 5

[STBC03] SWAYNE D. F., TEMPLE LANG D., BUJA A., COOK D.: GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis 43* (2003), 423–444. 2

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNORK M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2009), pp. 59 –66. 1, 2, 3, 5

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D. A.: Subspace Search and Visualization to Make Sense of Alternative Clusterings in High-Dimensional Data. In *Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2012), pp. 63–72. 6

[TT85] TUKEY J., TUKEY P.: Computing graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics 85. Nat. Computer Graphics Assoc.* (1985). 2

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *Information Visualization (InfoVis) 2005. IEEE Symposium on* (2005), pp. 157–164. 1, 2, 3, 5

[War94] WARD M. O.: Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings of the IEEE Symposium on Information Visualization* (1994), pp. 326–333. 2

[WRM13] WANG B., RUCHIKACHORN P., MUELLER K.: SketchpadN-D: WYDIWYG sculpting and editing in high-dimensional space. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2060–2069. 2