Gradient visualization of grouped component planes on the SOM lattice

Georg Pölzlbauer[1], **Michael Dittenbach**[2], **Andreas Rauber**[1]
[1] Department of Software Technology
Vienna University of Technology
Favoritenstr. 11-13, Vienna, Austria
{poelzlbauer, rauber}@ifs.tuwien.ac.at
[2] eCommerce Competence Center – ec3
Donau-City-Str. 1, Vienna, Austria
michael.dittenbach@ec3.at

Abstract - *The Self-Organizing Map has been successfully applied in numerous industrial applications. An important task in data analysis is finding and visualizing multiple dependencies in data. In this paper, we propose a method for visualizing the Self-Organizing Map by decomposing the feature dimensions into groups with high correlation or selections by domain experts. Using Gradient Visualization we plot a vector field for each of these groups on top of the map lattice, with arrows pointing towards the nearest cluster center. We provide a real-world example from the domain of petroleum engineering and point out our technique's usefulness in understanding mutual dependencies hidden in the data.*

Key words - Visualization, Dual Gradient Fields, Groups of Component Planes

1 Introduction and related work

The Self-Organizing Map is a popular tool for exploratory data analysis and visualization. The key qualities of SOMs are vector quantization, which is related to clustering, and vector projection, which is done by mapping from a high-dimensional feature space to a low-dimensional output space. Self-Organizing Maps have been used in a wide variety of industrial and scientific applications [2].

The SOM is frequently used for clustering tasks, and thus it is important to know how the feature variables are actually responsible for forming cluster boundaries. In this paper, we investigate the influence of groups of variables on the cluster structure, selecting either correlated variables by clustering of the component planes, or choosing semantically similar ones based on domain knowledge. For visual inspection, we propose a method to plot the results on top of the map lattice so the observer can get a feeling on how and where each group contributes to the cluster structure. Since the SOM is so popular partly due to its powerful visualization methods, we aim to maximize the information communicated through a single visualization by combining different plots.

SOM visualizations [7] either show the map in relation to the data set, or they are derived from the model vectors. The most commonly used technique for showing the distribution of the

#	**Variable Name**	**Type**	**Description**
1	Proppant in Formation	param	Amount of propp. pressed into formation (lbm)
2	Average Pressure	geo	Pressure in Reservoir (psi)
3	Average Rate	param	Rate of pumping into formation (bbl/m)
4	Pad Fluid Vol. Pumped	param	Total volume required to pump proppant (bbl)
5	Total Volume Pumped	param	Total of fluids pumped (bbl)
6	Produced Gas	out	Quantity of gas obtained (MSCF)
7	NetPay	geo	Depth of gas reservoir (ft)
8	Formation	geo	One of two ground formation types
9	Proppant Type	param	One of two proppant types
10	Stimulation Costs	out	Total cost of operation ($)

Table 1: Description of variables for Fracture Optimization data set

data are hit histograms, where the number of data points that are mapped to each map unit is counted and plotted on top of the map grid. Recently, we have proposed an advanced graph-based method [3] that plots the connectivity of the data vectors after projection, creating strongly connected graphs on the map for densely populated areas in input space. The P-Matrix [5] shows the relative density of the map units based on the average distance between data vectors.

The second category of visualization methods, which are based solely on the SOM's model vectors, aim mainly at showing clusters and their boundaries. Most notably, distance matrices, like the U-Matrix, indicate how close neighboring map units are in input space. Another commonly used method which will be used extensively in this paper are component planes, which plot the value of an individual dimension of each map unit's model vector. Clustering of the model vectors themselves can be applied to visualize regions that are close in input space by any clustering method like k-means or hierarchical clustering. The U*-Matrix [6] is a hybrid of distance matrix and density visualization. Its focus lies on smoothing the cluster boundaries for large maps that are trained with more map units than training vectors.

Our work is also concerned with finding and visualizing correlation in data, thus it is related to canonical correlation analysis. Correlation is often measured by the Pearson Correlation Coefficient, a normalized value based on the covariance matrix. Other methods include the ranking correlation, which are more robust regarding outliers, most notably Kendall's and Pearson's Ranking Coefficients, and partial and multiple correlation, which measure complex interrelationships between more than two variables. For the visualization of groups of component planes presented in this paper we use a different technique, namely the gradient field method [4] which will be discussed in Section 3.

The data set that we use in the rest of this paper is collected from the domain of petroleum engineering [9]. It has been collected to analyze the effects of pumping fluids and proppants into the gas field on the amount of gas produced. It consists of 199 samples, which correspond to gas wells, for which measurements of geological factors, parameters set by engineers, and performance indicators are taken in a total of 10 variable dimensions. We call it the "Fracture Optimization" data set. Table 1 provides a description of the variables, where the column "Type" indicates whether the variable describes a geological measurement, a parameter, or

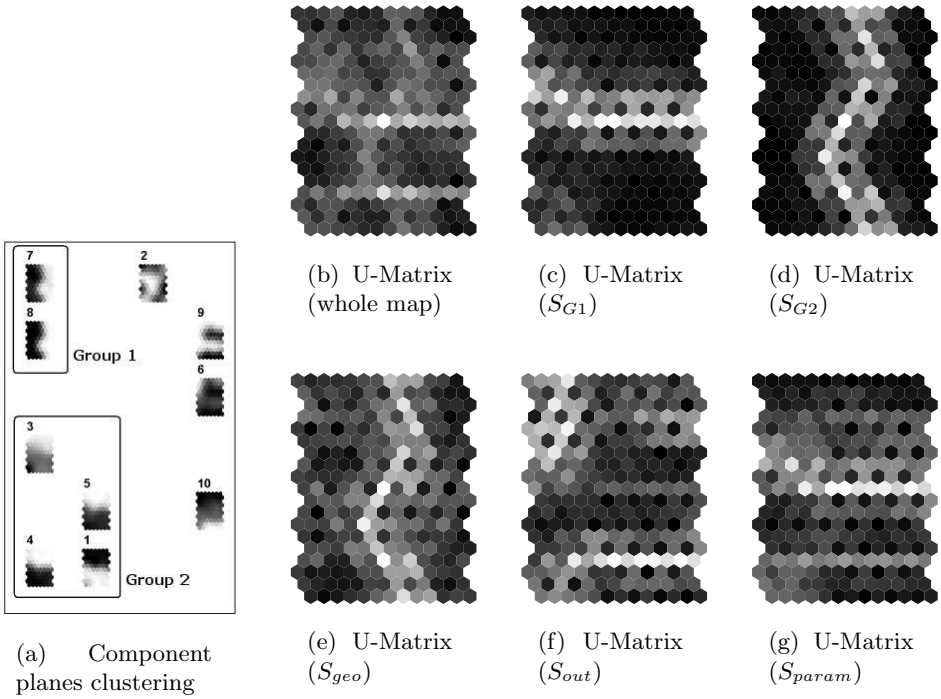
a result of the process. The SOM we use for experiments consists of 7×10 map nodes. The feature dimensions are weighted equally, we thus do not distinguish between geological, parameter and output variables, since the goal of this work is to find correlations rather than predict output values. Some of the variables are mutually dependent and correlated, for example "Total Volume Pumped" is the sum of "Pad Fluid Vol. Pumped" and other not explicitly measured fluids.
The rest of this paper is organized as follows: In Section 2, we discuss several options on how to find groups of similar feature variables. Section 3 introduces our method of visualizing the cluster structure of a SOM. In Section 4, we provide experimental results. Section 5 summarizes our work.

2 Groups of components

In this section, we will investigate how to obtain groups of variables and describe what we intend to do with them.
We begin with some formal definitions: The feature space $V = \Re^d$ contains the normalized data samples x and model vectors m; M denotes the set of model vectors (the codebook). The i-th component plane C^i of M is the projection of the codebook onto its i-th variable. Groups of component planes can be grouped and we denote S as the set that includes the indices of the selected components. We further need multiple groups of component planes, i.e. a partitioning of the available variable dimensions. Let g be the number of groups, and $S_{1...g}$ the corresponding sets of component plane indices, then $S_i \cap S_j = \emptyset, \quad i, j \leq g, \forall i \neq j$, i.e. a component plane must not be contained in more than one group, and $\bigcup_{i=1}^{g} S_i \subseteq S$ must hold. The union of the groups does not necessarily need to contain all the indices, thus not all of the component planes must be included. Further, we denote by $m^{(S_i)}$ the coordinates of model vector m of dimensions S_i. The codebook $M^{(S_i)}$ is then the set of all model vectors $m^{(S_i)}$. Note that $M^{(S_i)}$ is itself a Self-Organizing Map, and thus the U-Matrix, for example, can be computed and visualized. The projection is performed after the SOM has been trained, hence the same set of model vectors is decomposed into reduced sets of model vectors of the same map.
In the method proposed in Section 3, we investigate the correlation of groups of components, and aim at finding out how these groups contribute to the cluster structure. We further want to localize where the clusters differ, if only a subset of the components is considered. Thus, we need to find meaningful subsets of variable dimensions. There are mainly two ways component planes can be selected and pooled to be further investigated, either by picking variables that are highly correlated, or by choosing variables that are similar in terms of the domains where they have been measured. In the following subsections, we explain these choices in more detail.

2.1 Correlated variables

One way of obtaining interesting combinations of variables is selecting them based on high pairwise correlation. This can be performed either manually or automatically. Once a SOM has been trained, a convenient way to find groups is to manually select similar variables by visual inspection of component planes. The problem here is that the variables are usually not ordered and comparison becomes increasingly difficult with higher number of variables.



(a) Component planes clustering

(b) U-Matrix (whole map)

(c) U-Matrix (S_{G1})

(d) U-Matrix (S_{G2})

(e) U-Matrix (S_{geo})

(f) U-Matrix (S_{out})

(g) U-Matrix (S_{param})

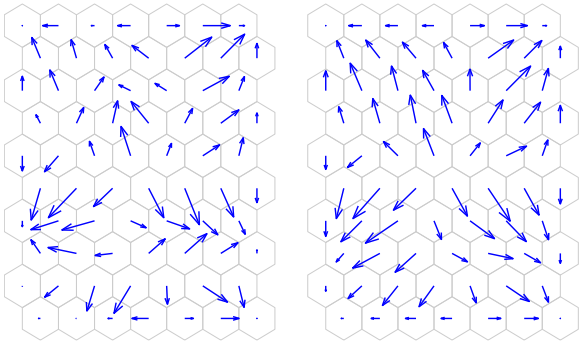
Figure 1: 7×10 SOM trained on Fracture Optimization data

An interesting approach [8] applies reorganization of the component planes such that similar component planes are located close to each other. This is performed by yet another SOM, where the input samples are component planes from the original SOM. The measure of distance between component planes i and j can be defined, for example, as the absolute value of the correlation of each map position, formally

$$d_{comp}(i,j) = -\|\gamma(C^i, C^j)\| \tag{1}$$

where γ is a suitable measure of correlation, usually the Pearson correlation coefficient. Because positive and negative correlation are not distinguished, the absolute value is taken. An advantage of this method is that an ordered presentation of similar components is automatically generated and provided to the observer. A disadvantage, however, is that the actual choice of grouping variables is left to the user.

A more thorough approach [1] includes comparing component planes and scatter plots. Another method would be to apply hierarchical clustering with the above defined distance measure d_{comp}. In this paper, however, we have chosen the interactive approach by reorganizing component planes. An example of this is depicted in Figure 1, where a SOM trained on the Fracture Optimization data set is shown. Figure 1(a) shows the results of reorganizing the component planes according to mutual similarity. We identified two clusters of component planes, the first one consisting of 4 variables in the lower left corner of the reorganized plane, the second one with 2 components in the upper left part. We are not interested in the remaining component planes due to their high dissimilarity from each other and the two clusters. The index sets are $S_{G1} = \{1, 3, 4, 5\}$ and $S_{G2} = \{7, 8\}$. The U-Matrix can be applied



(a) (b)

Figure 2: Gradient visualization with parameters (a) $\sigma = 1.5$, (b) $\sigma = 3$

to show the decomposition of local dissimilarities of map units on any of the reduced SOMs, as depicted in Figures 1(c,d). The U-Matrix of the whole map can be seen in Figure 1(b). Please note that it is not the simply the sum of the 2 matrices, since the number of variables differs in each group, and some variables are not included in the groups. However, a good impression is given on which variables are responsible for each of the cluster boundaries.

2.2 Semantically similar variables

Another option for choosing groups of variables is selecting variables that belong semantically together. Since the data is collected from real-world domains, groups of measurements from related sources may be of interest while these are not necessarily correlated.
For example, consider the three categories from the Fracture Optimization data set, where variables either describe geological factors, parameters applied during the engineering process, and output variables. The corresponding index sets are $S_{geo} = \{2, 7, 8\}$, $S_{param} = \{1, 3, 4, 5, 9\}$, and $S_{out} = \{6, 10\}$. U-Matrices are visualized in Figure 1(e)–(g). When compared to the groups found by the clustering selection of variables, the boundaries of the semantic groups seem less clearly drawn. In the former case, choosing correlated variables implicitly stresses the cluster boundaries. The semantics-based grouping, however, is more interesting in context of analyzing industrial data. For example, analysts are interested in comparing effects of geologic and engineering parameters independent from the output, to find out whether it is necessary to further optimize the engineering process given certain geological factors, or whether it is more important to carefully select the position of the well in the first place. Naturally, the clusters found in the previous subsection have a certain degree of similarity and overlap with the semantic classes.

3 Linking gradient field visualizations

In this section, we describe the gradient field method [4] for visualizing cluster structures. The key components are vectors computed and plotted on top of the SOM lattice. Each map

node i is assigned an arrow a_i that points to the nearest cluster center. The length of a_i is the ratio between the dissimilarity of the model vector m_i to the area it points to and the dissimilarity to the area it points away from. Thus, long arrows indicate that the map node is close to a cluster boundary, while short arrows are found either in cluster centers or in rare cases directly between two clusters with approximately equal distances to both clusters. The algorithm takes into account the neighborhood kernel and the width parameter σ, that specify the influence of nearby and distant nodes on the map. Higher levels of σ emphasize the global cluster structure by weighting a relatively large region surrounding the map units, while lower values stress the fine details of the clusters. The resulting visualization gives an overview on the cluster structure, where centers and interpolating units are located, and the degree of similarity of nodes to its surrounding area.
An example of the gradient field method is shown in Figure 2 for $\sigma = 1.5$ and $\sigma = 3$. The lower value of σ corresponds closely to the U-Matrix visualization, which depicts distances between neighboring map units. From Figure 2(b) it can be seen that there is a major boundary separating the map horizontally, and several minor clusters with centers most probably located in the corners. The differences to the U-Matrix are that a wider region is taken into account for computing distances in feature space, and that the vectors of the gradient field point to nearby cluster centers.
In the previous section, we have seen U-Matrices of groups of variables. We wish to extend and combine the gradient field method by linking two such visualizations that show the clustering structure to maximize the amount of information communicated through a single plot. By depicting two vector fields simultaneously, we intend to determine in which regions of the SOM discrepancies between the vectors occur to aid data exploration. In [4], for each model vector m_i a vector a_i is computed. The computation depends solely on the model vectors, thus we extend this definition to subsets S_1, S_2 of the variables. The gradient field algorithm is performed on maps with codebooks $M^{(S_1)}$ and $M^{(S_2)}$. We denote $a_i^{(S_1)}$ as the vector derived for model vector $m_i^{(S_1)}$, and analogously for S_2. Then, the result can be visualized by plotting two gradient fields with different colors. A normalization has to be performed to adjust the length of the arrows according to the number of underlying variables to weight larger sets heavier that smaller ones:

$$\hat{a}_i^{(S_1)} = a_i^{(S_1)} \frac{|S_1|}{|S_1| + |S_2|} \tag{2}$$

and analogously for S_2. $|S|$ denotes the cardinality of set S. Experimental results and a thorough discussion of the properties are provided in the next section.

4 Experiments

In this section, we discuss our combined gradient field method with both types of partitions of the input variables as shown in Section 2. Figure 3 shows the results for a gradient field visualization with kernel width $\sigma = 2$. In Figure 3(b), Groups 1 and 2 are shown. It can be seen that the variables S_{G1} are responsible for the horizontal division of the map, while S_{G2} split the map vertically. This was also visible in the U-Matrices in Figures 1(c,d), but here it is combined in a single plot. Orthogonal angles between arrows from different groups, as in Figure 3(b), indicate that the groups are almost independent.

Gradient visualization of grouped component planes on the SOM lattice

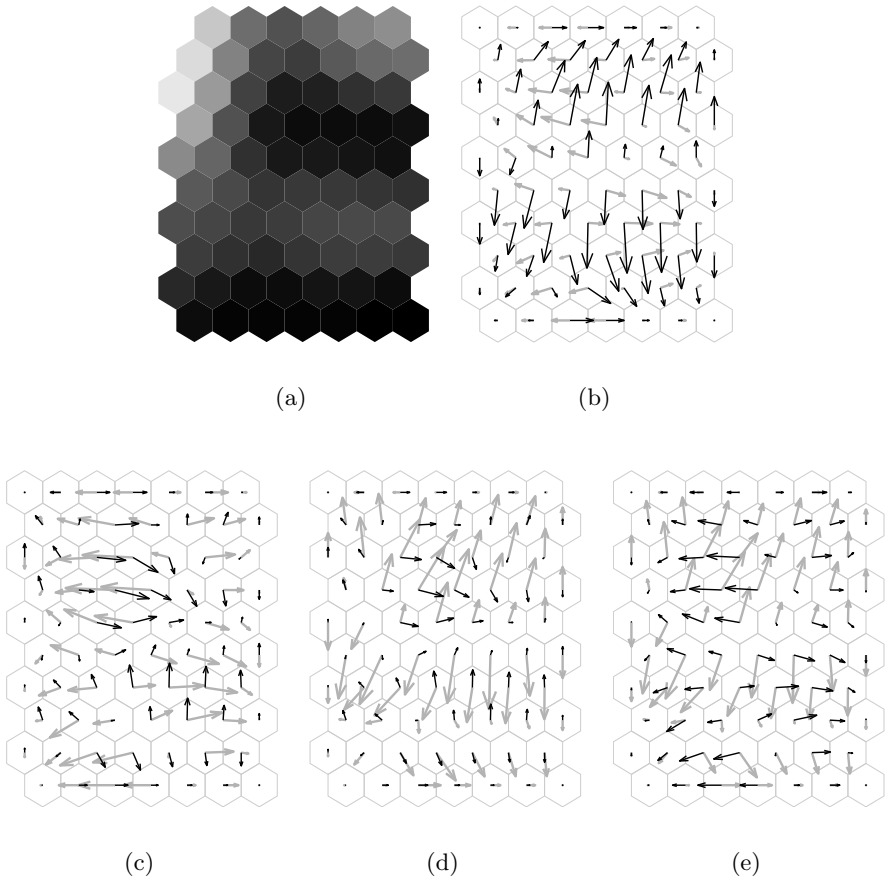


Figure 3: (a) Component Plane: "Produced Gas"; (b)–(e) Gradient visualization with parameters $\sigma = 2$ on pairs of variable groups; in the following enumeration, the first group is indicated by black, and the latter by gray vectors: (b) S_{G1} vs S_{G2}, (c) S_{out} vs S_{geo}, (d) S_{out} vs S_{param}, (e) S_{geo} vs S_{param}

The visualizations for the semantic groups show less defined cluster boundaries, but the results are more interesting for the fracture optimization process. By looking at Figure 3(e), it can be seen that the geological factors are divided vertically, thus the changes occur in the horizontal direction. For the engineering parameters, the changes occur mostly in the vertical direction. The important component plane "Produced gas", depicted in Figure 3(a) shows that the highest gas output can be expected from wells which map to the upper left corner of the SOM, which is thus the most desirable position. It can be derived that the engineering parameters can influence the position of a well in vertical directions, so the knowledge gathered from this visualization is that a well with certain geologic conditions can be moved "up" with right choices of parameters, but not to the left or right.

One property of our combined gradient field method that due to clarity reasons cannot be shown in black-white print, the combination with yet another visualization. A grayscale background visualization can be used and red and blue arrows for groups of variables. The background can be any of the typical SOM plots, like U-Matrix or hit histogram, or an

important component plane, like "produced gas" as in the example discussed above thus combining several different views of the SOM in a single plot.

5 Conclusion

In this paper, we have introduced a method for visualizing pairs of group of variables with a gradient field visualization. The goal of this technique is to get insight into the contribution of these groups to the cluster structure. We have seen an application on data collected from the petroleum industry, where variables have been grouped either according to correlation or by similar context. The results have shown the decomposition such that the engineering tasks can be adjusted to maximize the gas output.

References

[1] J. Himberg. Enhancing som-based data visualization by linking different data projections. In *International Symposium on Intelligent Data Engineering and Learning (IDEAL'98)*, 1998.

[2] M. Oja, S. Kaski, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural Computing Surveys*, 3:1–156, 2001.

[3] G. Pölzlbauer, A. Rauber, and M. Dittenbach. Advanced visualization techniques for self-organizing maps with graph-based methods. In *Intl. Symp. on Neural Networks (ISSN'05)*, 2005.

[4] G. Pölzlbauer, A. Rauber, and M. Dittenbach. A visualization technique for self-organizing maps with vector fields to obtain the cluster structure at desired levels of detail. In *International Joint Conference on Neural Networks (IJCNN'05)*, 2005.

[5] A. Ultsch. Maps for the visualization of high-dimensional data spaces. In *Workshop on Self Organizing Maps (WSOM'03)*, 2003.

[6] A. Ultsch. U*-matrix: a tool to visualize clusters in high dimensional data. Technical report, Dept. of Mathematics and Computer Science, Philipps-University Marburg, 2003.

[7] J. Vesanto. Som-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126, 1999.

[8] J. Vesanto and J. Ahola. Hunting for correlations in data using the self-organizing map. In *International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99)*, 1999.

[9] G. Zangl and J. Hannerer. *Data Mining: Applications in the Petroleum Industry.* Round Oak Publishing, 2003.