# Enhancing the Visualization Process with Principal Component Analysis to Support the Exploration of Trends

**Wolfgang Müller[1], Thomas Nocke[2], and Heidrun Schumann[2]**

1) Media Education and Visualization Group
University of Education Weingarten
D-88250 Weingarten, Germany
muellerw@ph-weingarten.de

2) Department of Computer Science
University of Rostock
D-18051 Rostock, Germany
(nocke/schumann)@informatik.uni-rostock.de

## Abstract

This paper describes the integration of the Principal Component Analysis into the Visualization Process. Although, the combination of Principal Component Analysis (PCA) and visual methods is a common approach to the analysis of high-dimensional datasets, it is mostly limited to a pure preprocessing step for dimension reduction. In this paper we will discuss, how PCA results can be used to control all steps of the visualization pipeline to generate more effective visual representations, and thus, a higher degree of understanding of the PCA values as well as of original data.

*Keywords*: Information Visualization, Visualization process, Principal Component Analysis.

## 1 Introduction

The amount of data in public and corporate databases is growing rapidly day by day, and databases with gigabytes or terabytes of data are no longer unusual. For long the visual representation of such data was proven to be of high value in exploring this data, detecting correlations as well as trends and outliers. For doing so, the high dimensional data must be somehow converted to low dimensional geometry for display. One well-established method in this context is the Principal Component Analysis (PCA).

PCA is a transformation technique that aims to identify main factors accounting for variances in the data. Identifying these factors leads to a more compressed description of correlations in the data and, thus, for a better understanding of the underlying features. Therefore, it is a powerful approach to extract general trends in a data set. Moreover, since PCA provides principal components ordered by their significance, it also offers an excellent basis for data dimension reduction in case of multidimensional data. This can be achieved by omitting less relevant trends in the data set and concentrating on main principal components.

The combination of PCA and visual methods is a common approach to the analysis of high-dimensional datasets in many application domains (see e.g. Yang 2003, Komura 2004, Landgrebe 2002, Santos and Brodlie 2004). However, nearly all of these approaches use the PCA as a kind of preprocessing computing significant trends in the data and visualize them. This may particularly cause 2 problems:

1. The PCA may generate principal components which are difficult to interpret and which sometimes even appear orthogonal to what intuitively seems to be the dominant trend in the data. A subsequent rotation of principal components is therefore sometimes approriate (see Joliffe 1986 for more details). Also, the necessary assumption of global linearity of the data oftentimes proves inadequate. (Müller and Alexa, 2004) therefore suggest an interactive, visually-guided improvement of the results of a PCA.
2. The coordinate transformation into principal component space involved with the PCA makes it difficult to relate identified trends to original variables. In general, it is often not easy to explain the variation of data with respect to a principal component.

Our aim is to reduce these problems by a stronger integration of the PCA into the visualization process. That means, we want to discuss, how the PCA can be used to support the different steps of the visualization pipeline to generate more effective visual representations, and thus, a higher degree of understanding of the PCA values as well as of original data.

The remainder of this paper is structured as follows. Section 2 describes the background including the fundamentals of the PCA as well as of the visualization process. In section 3 we look at each step of the visualization process, and discuss the main aspects of combining these steps with PCA. Moreover, we demonstrate the usefulness of our approach by different examples. We conclude with a summary and an outlook on further work in section 4.

## 2 Background

### 2.1 Fundamentals of PCA

The Principal Component Analysis is is a technique commonly applied to *reduce* the number of variables and to *detect structure* in the relationships between variables in multi-dimensional data sets. In the PCA the extraction of principal components amounts to a *variance maximizing rotation* of the original variable space. That is, the PCA provides a transformation of the original data space in such a way that the first coordinate of the resulting principle component space would resemble most of the variance in the data set, the second variable the most of the remaining variance, and so on.

More formally, we assume that our original data set is given in matrix notation

$$Y = \begin{pmatrix} y_{11} & y_{21} & y_{d1} \\ y_{12} & y_{22} & y_{d2} \\ & & \\ y_{1n} & y_{2n} & y_{dn} \end{pmatrix} \quad (1)$$

where each column is associated with a single variable, thus representing an individual dimension of the data set. Consequently, $d$ describes the number of dimensions. Each row $y_i$ , $i = 1 \ldots n$ represents a different case in the data set. For instance, each row may stand for a different time step. In the following, we assume a normalized and mean-centered data matrix $Y$. There are several ways to determine the principal components for $Y$. An often applied approach is to apply a *Singular Value Decomposition* (SVD). The SVD factors $Y$ directly into

$$Y = C\Lambda^2 W \quad (2)$$

where $W$ represents the loadings that is the basis vectors $w_i$ of the transformed vector space. $\Lambda^2$ represents a diagonal matrix with the significance of each principal component corresponding to the amount of variance described. Finally, $C$ represents the scores $c_i$, that is the coordinates of the data elements in the transformed vector space.

Once this information is obtained it can be exploited in several ways. First of all, the principal directions in W may be interpreted as prominent trends in the data. Also, the loadings in $W$ provide information on the correlation of different variables with these trends. For data reduction, on the other hand, we can simply neglect less significant axes in the principal component space and reduce the analysis to the remaining dimensions.

### 2.2 The Visualization Process

Santos and Brodlie (2004) introduce an extended data flow model to accommodate the visualization process (see figure 1). In contrast to the common visualization pipeline they replace the preprocessing step by two separate components distinguishing between data analysis and filtering. The data analysis component is computer-centered and provides some computational methods to analyze or enrich the data. Typical functions for this purpose are interpolation functions, PCA or Multidimensional Scaling (Santos and Brodlie, 2004). The filtering-component is user-centered, and provides different functionality for selecting data of interest. Thus, the portion of data to be visualized can be extracted. This is done in two levels. First, an n-dimensional window with upper and lower bounds is specified within the data domain. Moreover, a focus point within these bounds is set. In doing so, the representation within the window can be controlled, e.g. by accenting data of interest. Although (Santos and Brodlie, 2004) assume a fixed order of these two components (data analysis first, then interactive data filtering, whereby the data selection process can be repeated more than once), it can be expedient to swap these steps. For example, it can be useful to apply a PCA after a data selection step.

If a PCA is applied to all variables of a data set, the interpretation of the visualized PCA-components could be difficult. In this case, natural references such as time in the case of temporal data, or spatial dimensions in the case of spatial data, are also a target of the coordinate transformation into principal component space. Consequently, they are no longer available as a point of reference to relate to. Therefore, it can make sense, to select variables for PCA-processing, while leaving others as frame of reference.

The next step in the visualization pipeline is the mapping of selected and computed data to be visualized onto a geometrical representation considering the given focus point. This is the most crucial step during the visualization process deciding about the expressiveness and effectiveness of a visual representation. Different approaches were developed to support the mapping step, and to generate an appropriate visualization design (see e.g. Mackinlay, 1986; Senay, and Ignatius, 1994; Roth et al 96; Fujishiro et al, 2000; Zhou+ 2002; Almar and Stasko, 2004; Tang et al., 2004). However, none of these approaches apply PCA results to control the mapping. We will discuss in section 3.3 how PCA results (loadings from the matrix $W$ and eigenvalues from the matrix $\Lambda^2$) can be used to define and parameterize the mapping, and in doing so, creating intuitive visual representations for a better insight of the data.

The last step of the visualization process is the rendering step. Here, image data are created for display on a monitor. Usually, this step is not in focus of recent research. However, it is worth to investigate, how a combined handling of data and PCA-values can improve the display. For example, showing trends in the data by rendering principal components requires a mental mapping of axes of the principal component space onto data variables from the user. Therefore, an important step in the application of the PCA is the adequate labeling of the resulting axes. This includes the determination of adequate axes labels describing the background of a principal component, and providing suitable axis value marks to provide an adequate relationship to the original data space. To achieve this, it is necessary to identify in how far certain variables correspond to principal components. Nearly all PCA visualizations forbear from

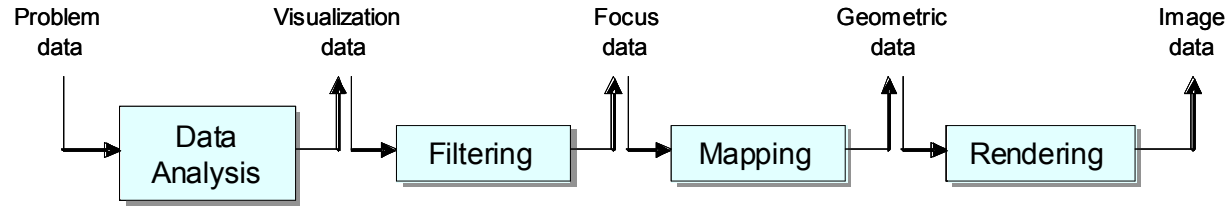


**Figure 1: The visualization process, adapted from (Santos and Brodlie, 2004)**

doing that. We will get back to this problem in section 3.4.

After discussing the separate steps of the visualization pipeline, finally we want to consider the process as a whole. The visual exploration of data sets has a high degree of interactivity. Thus, the visualization process can not be understood as one pipelining queue. Rather, different steps are repeated more than once to create different visual representations supporting different exploration tasks. For this, (Shneiderman, 1996) introduced the so-called *Visualization Mantra* "Overview first, Zoom and Filter, then Detail on Demand". Hereby, the fundamental procedure of visually exploring data is defined.

Starting with a general overview image representing general trends of a data set, zooming is applied to focus on regions of interest. Moreover, information hiding is provided to filter out non-relevant data. Finally, details on demand have to be added by request of a user.

PCA can provide useful information to support the *Visualization Mantra*. Since PCA provides principal components ordered by their significance, it offers an excellent basis for data dimension reduction, and thus for creating overview images. Moreover, Zooming can be controlled by PCA values, e.g. to set the focus on interesting trends or even on outliers. Similarly, certain trends or outliers can be filtered out depending on the exploration task. Finally it must be possible to add details, e.g. adequate annotations.

In sum, we can notice that a stronger interrelation between PCA and the visualization process can be very useful, and therefore it would be worth for further investigations. We will discuss this topic in more detail in the next section.

## 3 Integration of PCA into the Visualization Process

### 3.1 Data analysis and PCA

The first step of the visualization pipeline is the data analysis step. As already mentioned, PCA is a common procedure in this context, especially to achieve dimension reduction. The PCA provides principal components ordered by their significance, and thus less relevant trends in the data set can be omitted concentrating on main principal components only. Moreover, we want to provide a better insight of the PCA-results, and in doing so, a better insight of the features of the original data, by a stronger coupling of PCA and visual methods. Figure 2 shows our visual representation of the loadings of the matrix $W$ (see formula 2) for a demographical data set. High loadings in a column act as an indicator for a high relevance of the associated axis in the PCA-space representing, and thus, as a relevant trend of the given data set. The values in a row show the influence of the different variables on the trends represented by the PCA-axes.

More precisely, in Figure 2 we see a major trend represented by the principal component 1 (PC1). All the positive loadings (in blue color) in PC1 indicate a direct proportional relationship for the variables *literacy*, *infant mortality* (*Babymort*),…, and *life expectation*. *Population* and *density* do not influence this trend. The second trend (PC2) is constituted by *gross domestic product* (*GDP)* and *life expectation* and is indirect proportional to *infant mortality*, *death rate* and *birthrate*. This example illustrates how the user gets fast insight into the main trends in a high dimensional data set. However, he gains as well information about more hidden trends. Instance for such an "outlier trend" are the oppositional loadings of Life expectation of men and women in the principal component PC9.

A better understanding of the insights of the PCA can be used to control the following steps of the visualization pipeline (data selection, mapping and rendering) to generate more expressive visual representations. Furthermore, this provides a powerful basis to create overview images as well as detailed displays. To explore relevant trends for example, the user can switch to further views e.g. line charts or scatter plots (see section 3.3).

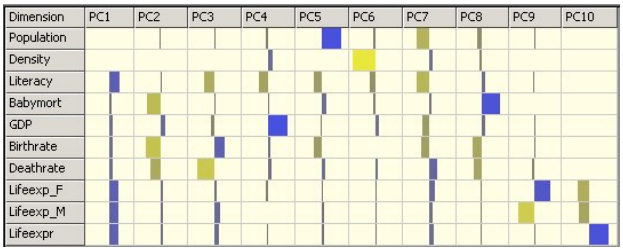


**Figure 2: Visualization of the Loadings matrix W for a demographical data set**

### 3.2 Filtering and PCA

The second step of the visualization pipeline is the filtering process. Here, variables of interest are selected for further procedure by the user. Thus, this process is highly interactive and user-centered. With regard to the PCA we distinguish two strategies:

1. Filtering before executing PCA,
2. Filtering after executing PCA.

Filtering before executing the PCA includes the selection of variables as well as of data objects (defined by columns resp. rows in the input matrix *Y*) for PCA processing. Usually, natural references such as time in the case of temporal data, or spatial dimensions in the case of spatial data, are excluded from PCA to leave them as a frame of reference. Moreover, special variables or data objects of interest can be selected for PCA-processing. Contrarily, high correlated variables or outliers (variables as well as data objects) can be excluded from PCA. For supporting the user to include or exclude variables and data objects for PCA, a visual representation of the original data can be very helpful.

Figure 3 illustrates how the user gets hints to filter out certain data objects (table rows) or variables (table columns) by example of a data table visualization of the demographic data set from figure 2. For instance, the user can examine high outlier values of *population* and *density* (the first two columns)

To exclude these outliers from PCA, the referring data objects respectively the referring variables have to be deselected. Furthermore, this data set includes three *life expression* variables (women, men, all, the last three columns), which are expectedly highly dependent. This fact will strengthen certain trends calculated by the PCA (may be more then the original data express). It is up to the user to exclude such dependent variables from the PCA.

Filtering after executing the PCA has high potency to guide users through a variety of trends in an overview and detail manner. Using a tabular loadings visualization (see figure 2) the user gains an overview about the most important trends and the involved variables. Based on such a loadings table he can select principal components, original variables or combinations of both. This is a flexible approach to analyze the trends and the contributions of certain variables to these trends by appropriate visual representations in more detail (see section 3.3).

Furthermore, PCA does not only calculate the trends, but even orders them by their significance (PC1 for highest to PCn for lowest significance). Additionally, the normalized PCA loadings (by the eigenvalues from the matrix $\Lambda^2$) can be used as a quantitative measure to express the significance information. This information is very helpful to guide users through the "trend space". That means, the calculated measures can be used to visually direct the user's attention to the most relevant trends.

Figure 4 shows a *Table Visualization* with normalized loadings. The first trend (PC1) - primarily established by *literacy* and *life expression* - has a high significance for almost all variables. Normalized loading values for the next relevant PCs are strongly decreasing. Thus, the user can select low dimensional subspace of the original space for visualization purposes concentrating on the exploration of main trends. In this case, for instance, a simple line chart can represent the first principal component.

### 3.3 Mapping and PCA

Mapping is the most crucial step in the visually guided process of analyzing data. It involves both the selection of an appropriate visualization technique and the mapping of data values onto expressive and effective visual attributes, such as position, marker sizes, and marker color.

When applying a PCA we enrich our original data by a principal component representation making additional aspects of the data explicit and, thus, available in the mapping process. Consequently, the gained information may influence our decisions on what aspects of the data to map onto visual attributes (i.e. PC values vs. original data).

Visualizing the data from principal component space instead of mapping the original data follows the main objective of the PCA to reduce the number of dimensions of the original data space and to present prominent trends in the data only. We will demonstrate this strategy by an example. Figure 5 displays a reduced version of the countries data set in the original data space in terms of a Scatterplot Matrix.


**Figure 3: Data table visualization of the demographic data set (with table lens, see Kreuseler (2002))**

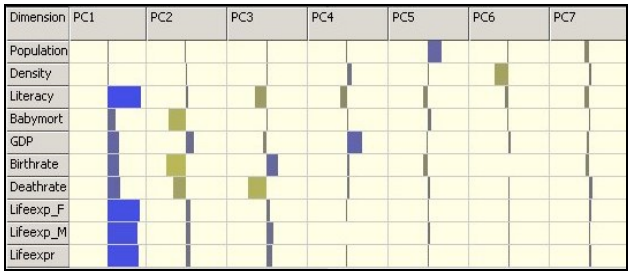


**Figure 4: Normalized loadings visualization of the demographic data set**

We had to reduce the original data set to generate this intuitive representation. The visualization of the whole data set (in this case 10 variables) or of even higher dimensional data sets will generate rather complex visual representations with many recurring dependencies and thus, quickly exhaust users. A big drawback of manually filtering variables is the loss of possibly relevant trends in the data. The question is how to decide which original variables to show and which not, or in the case of the demographic data set: why do we display the variable *literacy* and not the variable *population*. To answer this question, PCA can be a good guide to select (depending) variables keeping the remaining trends in mind (see section 3.2). For instance, in figure 5, the selection of variables is based on the main PCs (see figure 4).

In contrast to the PCA controlled visualization of variables in the original space, we may decide to utilize the data in principal component space for a direct visualization. That is we utilize the PCA score matrix data and map those transformed coordinates directly to visual attributes.

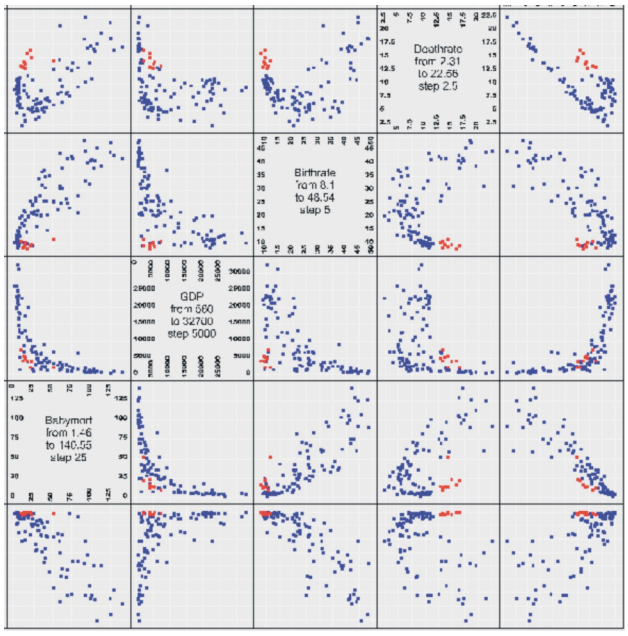


**Figure 5: Scatterplot matrix of a subspace of the demographic data set; countries with high death rate, low birth rate, high literacy and high life expectation highlighted (all are former east European countries)**

In the case of the dataset discussed before, for instance, a scatter plot matrix of the first three principal components presents the most noticeable trends in the data directly (see figure 6). Interestingly, in our example there is a strong correlation between the trends represented by the first two principal components. This is usually a sign for a more complex correlation between the corresponding variables that cannot be described easily in terms of a variance in a linear direction.

As an alternative to these straightforward approaches we suggest to apply a table visualization of the scores ordered by the values of the first PC (see figure 7). The advantage of this kind of visualization is that users can directly identify data objects with certain PC values, and thus, get deeper inside about the meaning and significance of the trends. For example, we can see that the last three rows (representing the countries *Japan, Norway* and *Luxembourg*) are prominent representatives of the first two trends.

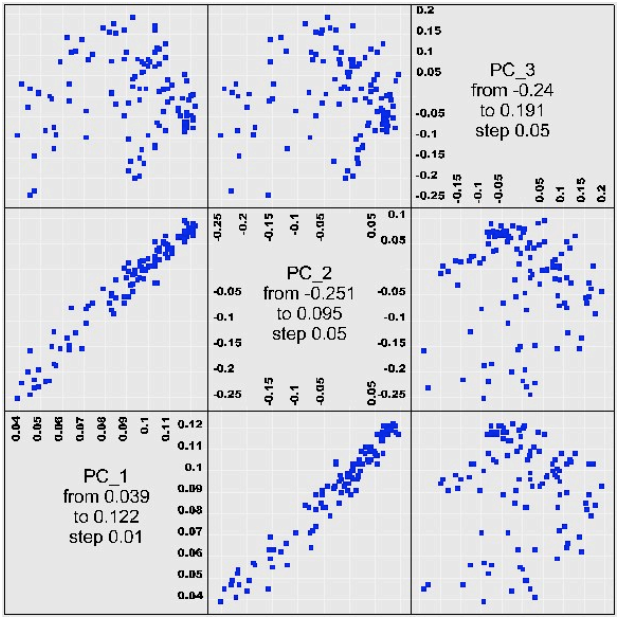


**Figure 6: Scatter plot matrix with the first three principal components**

Besides presenting the most significant principal components it may also be appropriate to visualize the least significant ones, for instance to identify outliers or erroneous data.

In general, the presentation of principal component scores results in an abstraction of the original variables and in abstract axes. Both make it difficult to relate the visible patterns to trends with respect to the original variables.

This represents a specific form of a correspondence problem. Consequently, graphing data in principal component space requests for an adequate labeling of the resulting PC axes. We will come back to this issue in chapter 3.4.

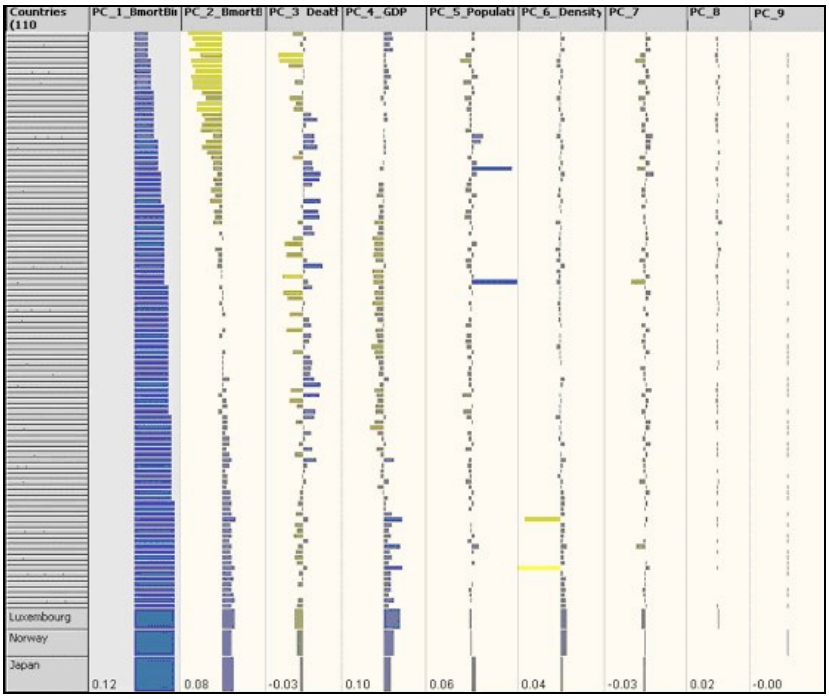


**Figure 7: Table visualization of the normalized PC scores of the demographic data set ordered by the first PC**

A useful aid for the interpretation of original and PCA data provides the possibility to combine both in a single plot. For instance, in the example of the demographic data set a similar scatter plot relating *literacy* and *life expectation* to the first principal component can clarify the correspondence of the first identified trend to these variables. Figure 8 shows a strong correlation between these three variables, whereby the correlation between *literacy* and PC1 is more scattered than the correlation between PC1 and *life expectation*.

For a combined representation of PCA values and original data we have to decide which variables to present directly. A common approach in this context is to distinguish between independent and dependent variables. Usually, dependent variables are processed by PCA, whereas independent are displayed in the original data space. In our case it can make sense to show PCA values of the demographic data set in their spatial frame of reference (see figure 9).

Another approach is to combine PCA loadings instead of the scores to the original data. Different from the score data the loadings represent the prominent trends as identified by the PCA directly. We may utilize this information to depict the identified trends in the context of the original data (Müller and Alexa 2004). Figure 10 depicts an example for such an approach.

### 3.4 Rendering and PCA

So far we discussed approaches based on a direct visualization of original data and/or principal components and their factors. We mainly concentrated on combining original and PCA data in visualizations and on performing a direct mapping of PCA data to visual attributes. Another promising approach is to utilize the trend information made explicit by the PCA for tuning the rendering process. We denote this *implicit visualization* of PCA values.

An implicit visualization may especially aid in the filtering and the mapping stage. For instance, variables can be mapped automatically onto visual attributes based on the correspondence of the variable to a major trend as identified by the PCA and the effectiveness of certain visual variables (e.g. position vs. color, see Mackinlay 1986). This opens up a new field of automatically controlling the mapping of data variables based on detected trends in the data.

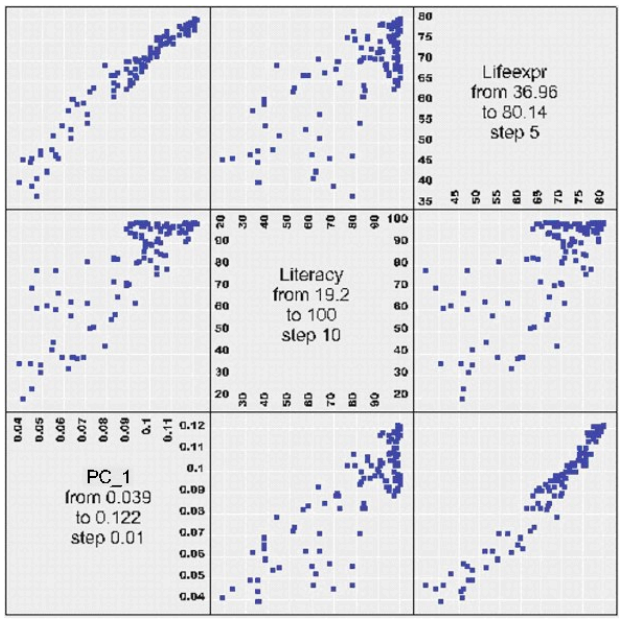


**Figure 8: Scatter plot relating the first PC to the variables literacy and life expectation**

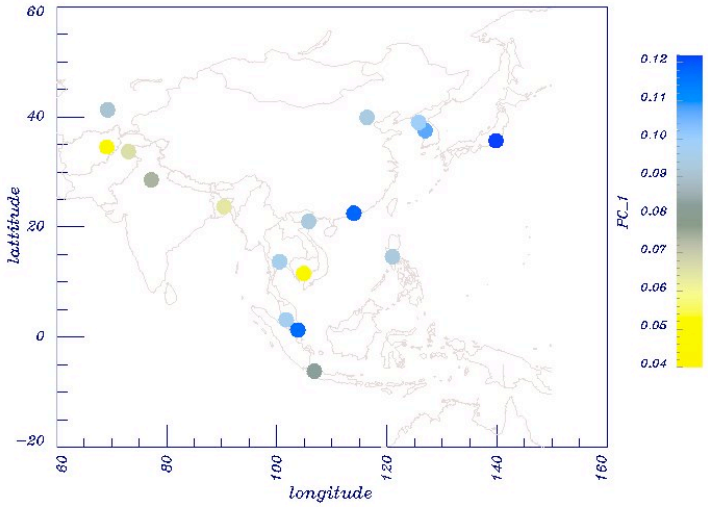


**Figure 9: Spatial visualization of the first PC of the demographic data set (cutout of Asian countries; PC1 is represented by a colored circle at the capital location**

### 3.4.1 Focusing and arrangement

Another application of principal component scores is their utilization for an indirect visualization, for instance for brushing & linking. In figure 11 high values with respect to the most prominent data trend are selected (left), and highlighted in a scatter plot visualization (right) representing three original data variables (two axes and color). Besides a simple highlighting, further details can be presented for the selected data elements, e.g. annotations or semantic zooming. Brushing and linking of PC score values is not limited to standard graphs but can also be utilized in the context of other information visualization techniques for multi-dimensional data. Figure 12 depicts an example for focusing on information in a table view representation by applying a table lens. Data objects corresponding to high trend values are rendered with more detail (by larger row space).

The additional information available in terms of the PCA scores can be used to steer the rendering process either manually or automatically by setting for instance the focus on most prominent trends or outliers automatically in an initial view (see e.g. figures 11 and 12).

Besides changing the focus and highlighting certain objects, PCA can be applied to automatically arrange original variables or data objects as well. Figure 13 depicts the table visualization from figure 3 reordered by the first PC.

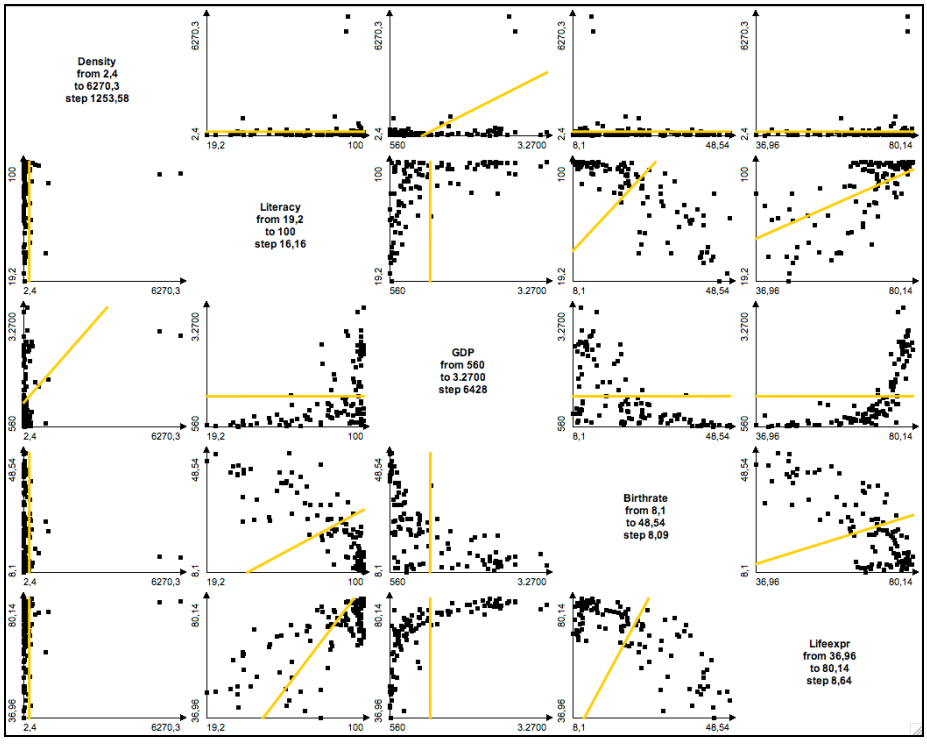


**Figure 10: Scatter plot matrix of the original demographic data set with PC1 directions**

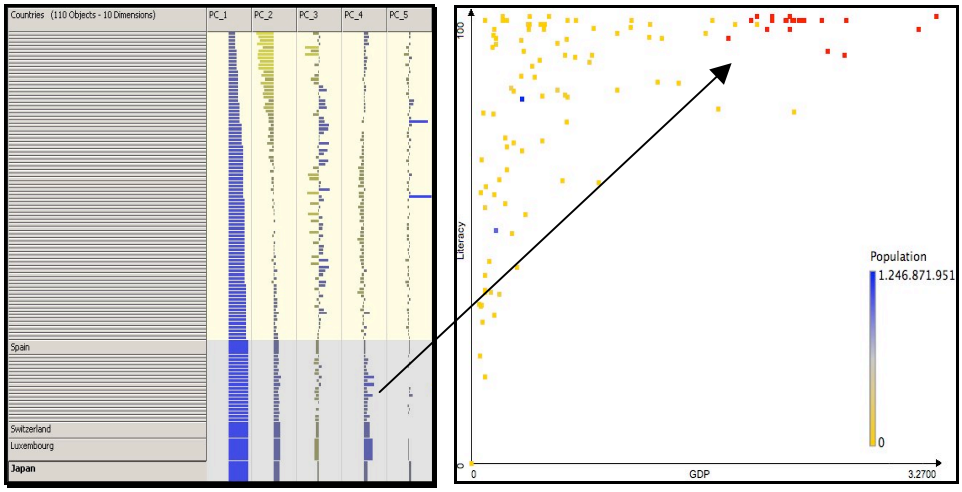


**Figure 11: Example of Brushing& Linking utilizing PCA score data. Selecting high PC score values in the table view allows to link extreme values with respect to a prominent trend in the data with original data (scatter plot visualization of the variables literacy (y-axis), GDP (x-axis) and population (color))**

### 3.4.2 Annotation

As already discussed, a general problem in the visualization of principal components is to relate the identified trends to the original variables (see section 3.3). The principal component space represents an abstraction of the original data space and its basis vectors contain usually aspects of several data variables.

As one solution to this problem, we propose combined visualizations of original variables and PC (figure 8) and implicit visualization of PCs in the original space (figure 11 and 13) to enable users to interactively approach to the meaning of the trends calculated by the PCA, and thus, improve the understanding of the rather abstract PCs.

A second solution to the mentioned problem is to use the information given by the significances (eigenvalues) and the loadings to encode the meaning of the PC more explicitly in the visualization. Challenge in this context is to enrich the rather abstract visualizations in the PC space, especially by generating meaningful axes labels.

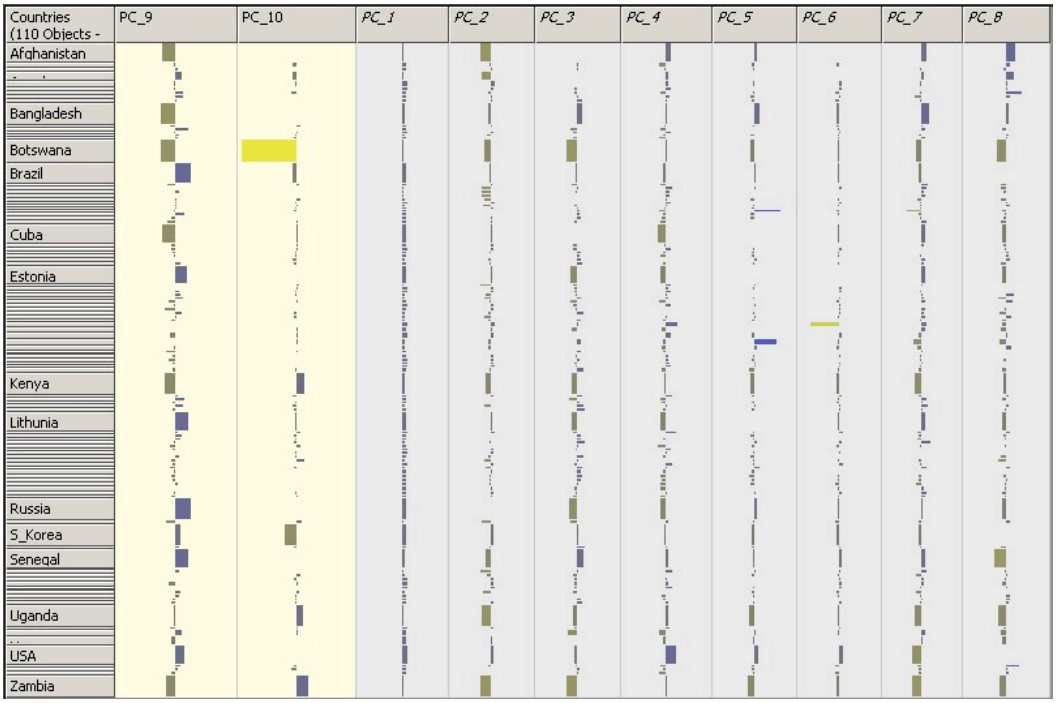


**Figure 12: Table Visualization with adapted table lens function to focus on countries with extreme score values in the trends PC_9 and PC_10**


**Figure 13: Data table visualization of the demographic data set, reordered by first PC (with table lens, see Kreuseler, 2002)**

Figure 14 displays a first solution to this problem, enriching PC axes labels by the variables they mainly represent (applying a threshold on normalized loading values), ordered by their significance. However, determining when a trend based on a principal component corresponds to a variable or not is not always easy. (Joliffe 1986) recommends to utilize the PCA loadings and to accept a correspondence in case of 70% accordance of original variable and principal component. Unfortunately, we often may have the situation that a trend corresponds to a large number of variables which leads to long label lists that dominate the visual representation (PC1 label in figure 14) or stay hidden because of space restrictions (see figure 7). Therefore, it is useful to let the user decide and to provide adequate settings for this purpose. For tuning the labeling parameters he is supported by our proposed visual representations for PCA results and original data.

Besides understanding the PCA axes, a further challenge is to annotate the meaning of the score values in the visualization to enable users to estimate which score value corresponds to which original variable value.

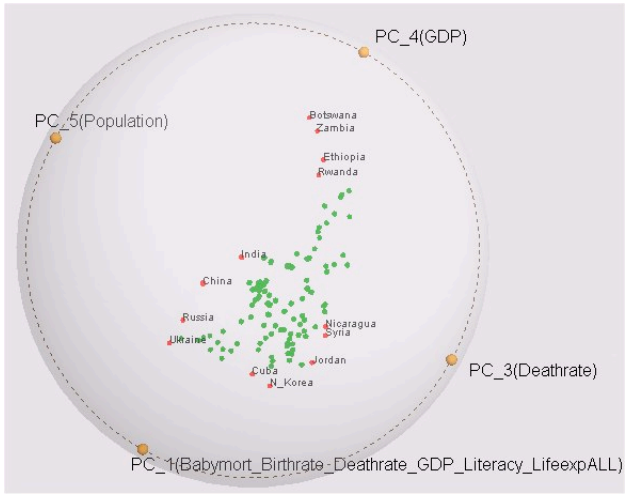


**Figure 14: ShapeVis visualization of data objects in PC space using enriched labeling with data variable names (mapping objects to 2D locations using a spring model, see Theisel and Kreuseler, 2002)**

As a first solution, we tested annotating minimum and maximum score values at the axes with coresponding original values. Unfortunately, this kind of annotation even extends the long list of original variable names. An alternative approach is to render a brief and abbreviated labeling and to provide additional information on request via brushing, e.g. depicting on PCA axes resp. data objects using the mouse. However, abbreviated labels are usually application dependent and therefore difficult to generate automatically. To conclude, further research needs to be done to adequately enrich visualizations in PC space by appropriate annotations.

## 4 Conclusions

In this paper, we systematically discussed how to combine information visualization and PCA in the different stages of the visualization process. As a result, we conclude that the enrichment of the original data by PCA generated data and the application of these PC values in the visualization process is a strong support for users studying trends. A variety of representations explicitly and/or implicitly visualizing PC values (scores, loadings eigenvalues) and original data values have been developed and presented. Table 1 demonstrates the tasks performable with these representations on the basis of the kind of PC values applied.

However, there are still challenges for future work. First of all, we want to supply the proposed PCA based visualizations to non-statistic expert users and evaluate them. Finally, further research needs to be done to improve the understanding of trends generated by the PC, for instance by improved annotation strategies for PC axes.

## References

Almar, R.; and Stasko, J. (2004): A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. Proc. IEEE InfoVis'04, Austin, USA, 143-149, IEEE Computer Society Press

Fujishiro, I.; Ichikawa, Y.; Furuhata, R.; and Takeshima, Y. (2000): GADGET / IV : A Taxonomic Approach to Semi-Automatic Design of Information Visualization Applications Using Modular Visualization Environments. Proc. IEEE InfoVis'00, Salt Lake City, USA, 77-83, IEEE Computer Society Press

Joliffe, I. T. (1986): Principal Component Analysis. Series in Statistics. Springer.

Komura, D., Nakamura, H., Tsutsumi, S., Aburatani, H. and Ihara, S. (2004): Multidimensional support vector machines for visualization of gene expression data. Proc. ACM symposium on Applied computing, Nicosia, Cyprus, 175 – 179.

Kreuseler, M.; Schumann, H. (2002): A Flexible Approach for Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics **8**, S: 39-5, IEEE Computer Society Press.

Landgrebe, J., Wurst, W. and Welzl, G. (2002): Permutation-validated principal components analysis of microarray data. Genome Biology, **3**: research 0019.1-0019.11.

Mackinlay, J. (1986): Automating the Design of Graphical Presentations of Relational Information. Transactions on Graphics, **5**: 110-141, ACM Press.

Müller, W. and Alexa, M. (2004): Visual Component Analysis. Proc. of Joint IEEE/EG Symposium on Visualization, VisSym'04, Konstanz, Germany, 129-136

Roth, S.F. et al (1996): Visage: A User Interface Environment for Exploring Information. Proc. IEEE InfoVis'96, San Francisco, USA, 3-12, IEEE Computer Society Press

dos Santos, S., and Brodlie, K. (2004): Gaining understanding of multivariate and multidimensional Data through Visualization. Computers & Graphics, **28**: 311-325, Elsevier.

Senay, H.; and Ignatius, E.(1994): A Knowledge-based System for Visualization Design. IEEE CG&A; **14**: 36-47.

Shneiderman, B. (1996): The Eyes have it: A task by data type taxonomy of information visualizations. Proc. IEEE Symposium on Visual Languages'96, 336-343 IEEE Computer Society Press.

Tang, D.; Stolte, C.; and Bosch, R. (2004): Design Choices when Architecting Visualizations. Information Visualization, **3**: 65 – 80, Palgrave Macmillan Ltd.

Theisel, H.; and Kreuseler, M. (1998): An Enhanced Spring Model for Information Visualization. Computer Graphics Forum, vol. 17, no. 3, (Proceedings Eurographics '98)

Yang J., Ward M. O., Rundensteiner, E. A., and Huang, S. (2003): Visual hierarchical dimension reduction for exploration of high dimensional datasets. Proc. of Joint IEEE/EG Symposium on Visualization, VisSym'03, Grenoble, France, 19-28.

Zhou, M. X.; Chen, M.; and Feng, Y. (2002): Building a Visual Database for Example-based Graphics Generation. Proc. IEEE InfoVis'02, Boston, USA, 23-32, IEEE Computer Society Press.

| | **PC scores** | **PC loadings** | **PC eigenvalues** | **Original data** |
|---|---|---|---|---|
| **PC scores** | Analyze trends in the PC space and the relationships of main trends (figues 6, 11 left, 12, 14)<br>Interactively reorder PCs (7 and 11 left)<br>Focus on data objects strongly representing trends (figures 7, 11 left, 12, 14) | Annotate PC axes (figures 7 and 14) | Annotate PC axes (figures 7, 14)<br>Focus on major trends (figure 7) | Get deeper insight into meaning of PC axes (figures 8, 9, 11 right)<br>Analyze trends in observation space (figure 9, 11 right)<br>Focus and/or order on data objects in original space (figure 13) |
| **PC loadings** | / | Get a compact overview about strength and relationship of the main trends (figures 2, 4 ) | Focus on major trends (figure 4) | Get overview and deeper insight into meaning of PC axes (figures 2, 4)<br>Compare dependencies in original data with PC axes directions (figure 10)<br>Focus and/or order original data variables |
| **PC eigenvalues** | / | / | **?** | **?** |
| **Original data** | / | / | / | Typical InfoVis tasks (figures 3, 5, 8, 10) |

**Table 1: Possible combinations of PCA results and data for visualization purposes and the possible tasks; each cell represents a specific combination of certain kinds of input data to be used explicitly or implicitly in the visualization process; the lower triangle matrix is left empty for symmetry reasons**