

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

COORDINATES ORDERING
IN PARALLEL COORDINATES VIEWS



UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

KATEDRA INFORMATIKY

COORDINATES ORDERING IN PARALLEL COORDINATES VIEWS

(Bakalárska práca)

Študijný program: Informatika
Študijný odbor: 9.2.1 Informatika
Školiace pracovisko: Katedra aplikovanej informatiky
Vedúci: Mgr. Martin Florek
Kód: a3486020-14ef-4b6d-91a8-487b113120d7

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Lukáš Chripko
Študijný program: informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: anglický

Názov: Coordinates ordering in parallel coordinates views


Cieľ: urobiť prehľad technik triedenia suradnic pre vizualizáciu paralelných suradnic. implementovať tieto metódy a porovnať výstupy na reálnych dátach. preskúmať možnosti viacsobného zobrazenia rovnakej suradnice.

Vedúci: Mgr. Martin Florek

Katedra: FMFI.KAI - Katedra aplikovanej informatiky


Dátum zadania: 11.10.2010

Dátum schválenia: 02.11.2010


doc. RNDr. Daniel Olejár, PhD.
garant študijného programu



.....
študent



.....
Vedúci

Čestne vyhlasujem, že bakalársku prácu som vypracoval samostatne s použitím uvedenej literatúry a pod dohľadom môjho vedúceho práce.

.....

Chcel by som poďakovať môjmu vedúcemu Mgr. Martinovi Florekovi za jeho pomoc, rady a za dohľad nad mojou činnosťou.

Abstrakt

Množstvo dát, ktoré musíme spracovať, neustále naberá na objeme. Preto sa vyvíjajú nové a lepšie metódy na zobrazenie týchto údajov a takisto nové metódy na ich analýzu. Jednou z takýchto vizualizačných metód je aj využitie paralelných súradníc. V bakalárskej práci sa venujeme využitiu niekoľkých spôsobov ako spracovať a efektívne zobrazíť viac-rozmerné údaje a získať z nich potrebné informácie. Práca obsahuje rôzne grafické techniky, ale aj užitočné algoritmy, ktoré transformujú údaje do výhodnejšieho tvaru.

KĽÚČOVÉ SLOVÁ: Paralelné súradnice, viac - rozmerný, triedenie súradníc, analýza hlavných komponentov

Abstract

Amount of data that must be processed continuously gaining volume. Therefore, developing new and better methods depicted on these data and also new methods for their analysis. One such visualization method is the use of parallel coordinates. In my work I deal with them and use just a few ways to handle and effectively portray multi-dimensional data and obtain the information they need. Here you will find various graphic techniques but also useful algorithms transforming data-set in a more favorable shape.

KEYWORDS: Parallel Coordinates, Multi-dimensional, Principal Component Analysis, PCA, Clutter based analysis, coordinates ordering

List of Figures

1.1	Venn diagram[10].	5
1.2	Map [11].	6
1.3	Charles Minard’s diagram[9].	7
1.4	Dr. John Snow’s map[12].	8
2.1	Example of data-set in parallel coordinates.	9
2.2	One multi-dimensional point in parallel coordinates[1] Figure 1.1 . . .	10
2.3	NYC subway ridership from 1904 to 2006. Each dimension coresponds to year and each record represents one subway route.	11
2.4	Information about coal disasters.	12
3.1	Data-set with zero percent blending value. Similar lines overlap. Can- not regonize data concetration.	13
3.2	Data-set with eighty percent blending value. Similar data shows up and we can see data concetration.	14
3.3	Several green colored records that we’ve outlined the course.	15
3.4	Several green colored records that we’ve outlined the course.	16
4.1	Data-set with colored outliers	19
4.2	Data-set with colored outliers after reordering	19
4.3	Data-set before using PCA alogrithm.	21
4.4	Data-set after reoredring by significance and coverted records.	21
4.5	Data with no analysis yet.	22
4.6	After PCA analysis we see reorder dimensions by significance.	22
4.7	Coloured outliers between each dimension.	23
4.8	Outlier reduction using clutter based analysis.	23

5.1	Our application for data analysis using parallel coordinates.	25
-----	---	----

Contents

List of figures	viii
Introduction	1
1 State of the art report	3
1.1 Information	3
1.2 Visualization	5
1.3 History of visualization	7
2 Parallel coordinates	9
2.1 Introduction	9
2.2 How does parallel coordinates works	10
2.3 Usability of parallel coordinates	11
3 Methods for increasing legibility	13
3.1 Blending	13
3.2 Coloring	15
3.3 Combination	16
4 Dimensional ordering	17
4.1 Ordering purpose	17
4.2 Clutter based analysis	18
4.3 PCA	20
4.4 Combination of different analysis	22
5 Implementation	24

5.1	Technologies used	24
5.2	Application	25
	Conclusion	26
	Bibliography	28

Introduction

People have always been collecting and storing large amount of data. But how to effectively read and display those data? During the eighties in computer research created a new direction called visualization. It has become very important discipline dealing with displaying information. By now its biggest contribution has been not only to entertainment, science or city planning but also in data analysis.

In fact, visualization provides wide range of processes helping to solve various problems. We can transform problem into visual model, which provide better image about the problem. On the other hand while progress in visualization application in physical science, visualization in geography or graph visualization of physical processes, has been increasing, the progress of visualization techniques in fields such as finance, communications or control has been much slower because of presence of large variables number, causes problems with the appearance and legibility of information.

Using theory about parallel coordinates our goal was to create application, written in C++ with connection of OpenGL, displaying multidimensional data-sets in readable and informative form. Application shows how to display thousands of multidimensional data in “simple” 2D geometry and how to analyze this data using various methods of dimensional reordering.

Our thesis is divided into four parts. In the first part of our thesis we will introduce issue of information and visualization.

In second part we write about parallel coordinates, its history and usage in displaying large amount of data and about parallel coordinate contribution to data analysis.

Third one deals with possibilities of improving the readability of big amount of data

using parallel coordinates.

The last part become acquainted with ways to sort dimension of individual data-sets and how dimensional reordering works.

Chapter 1

State of the art report

1.1 Information

Information can be described as a message which increases the level of our knowledge. Usual definition of information is an ordered sequence of symbols or symbol that record message, but most often is associated with concepts such as meaning, knowledge, representation, data or entropy.

Entropy, in our context refers to Shannon entropy, which represents the expected value of an information extracted from message.

Information always carries a message for a receiver and this message becomes information for receiver because human interpretation. The value of information obtained from message can be quantified by entropy in units called bits. Accordingly, it follows that the information can be measured and also value of the information can be enumerated, that is why information value of message differs from observer. In this way information can be divided into information without meaning and relevant information, that is why the form of message and signs used for writing message are important. One message can be relevant for one person, but also meaningless for another. One information can be different for different persons, like numbers in binary code do not have same information value for common human and for somebody who study computers or mathematics.

Thanks to ambiguity every information can be interpreted in various ways and we can make improvements to entropy of this information. Thus different information's coding or different displaying can increase entropy. Thanks to this, it has begun to use various techniques displaying the information stored. This techniques can increase value of information received from obtained message. In bulk amount, the information must be filtered and sorted, there is a place where once again different techniques of data transformation are applied. Various transform helps us to streamline amount of information received and get what we want to know.

1.2 Visualization

Simply put, visualization is a way to graphically express a message. Visualization uses various techniques such as images, diagrams or animations. The use of visualization is to present information. Visualization is getting ahead because of modern computer graphics. Most people use visualization in everyday lives without realizing it. Visualization of weather using meteorological data, large range of traffic visualization, all time-line visualization showing changes over time, these are basic data visualization techniques.

Visualization methods often differ according to provided information. The most basic method visualizing data is simple diagrams and graphs.

Visualization is also different according to branch used in. Visualization applications



Figure 1.1: Venn diagram[10].

can be divided into several branches like Scientific visualization, Education visualization, Knowledge visualization, Information visualization, Product visualization, Visual communication and Visual analysis[2]. Each of these branches use various type of visualization techniques and interaction as needed. There are many types of

visualization techniques but in common there are charts (pie chart, histogram, scatter plot), graphs (various types of diagrams, flowchart), maps, time-lines and also some more advanced one as parallel coordinates, cluster diagram and many others[8].

Map below is simple example of data visualization, which provides population density data in world.

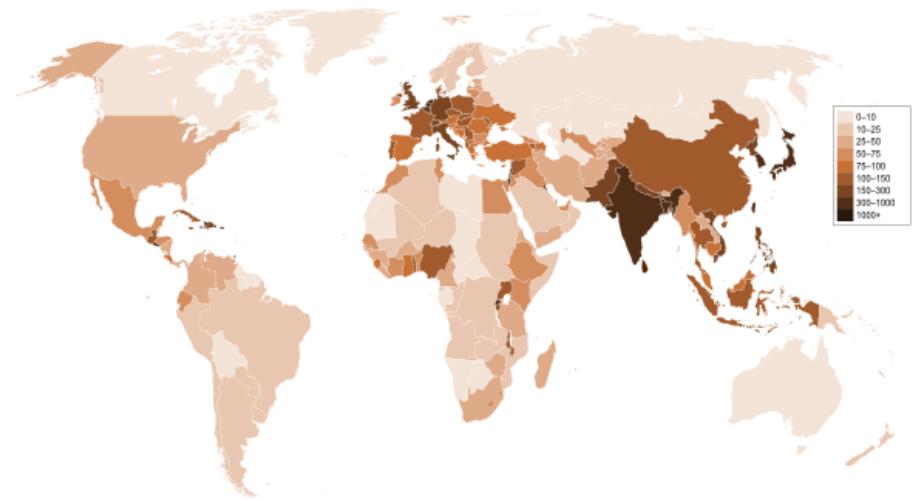


Figure 1.2: Map [11].

1.3 History of visualization

We can say humanity is using visualization from its beginnings. First known map comes from ancient Babylon dating to 600BC, but as the years passed visualization has been improving and changing[2]. Very well known is Charles Minard's flow map of Napoleon's March. It is a simple two-dimensional graph, which displays several variables. Graph shows army size during Napoleon's march. This map displays decreasing army size during march, but also provides location of army.

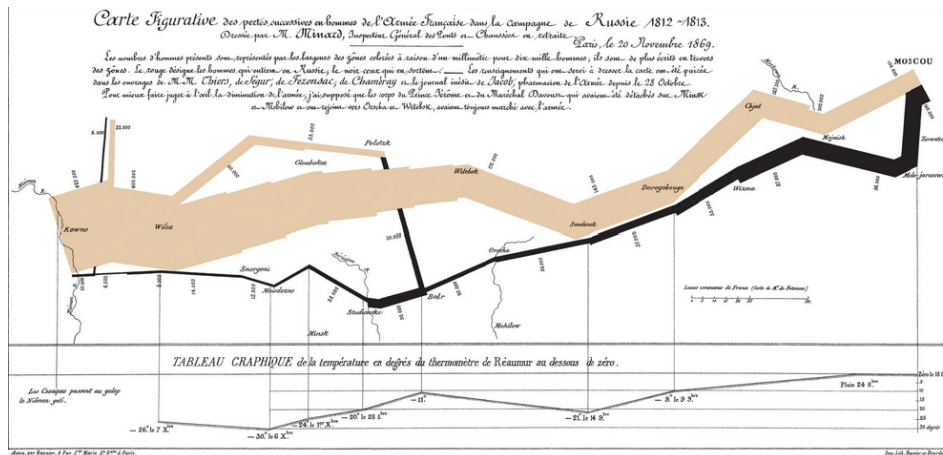


Figure 1.3: Charles Minard's diagram[9].

Another very well known is visualization if from 1854 when cholera epidemic stroke in London. Dr. John Snow for any infection or a dead man drew a dot on the map of the city. He realized that concentration of dots on map is around one city water pump. After closing this water pump the epidemic stopped.

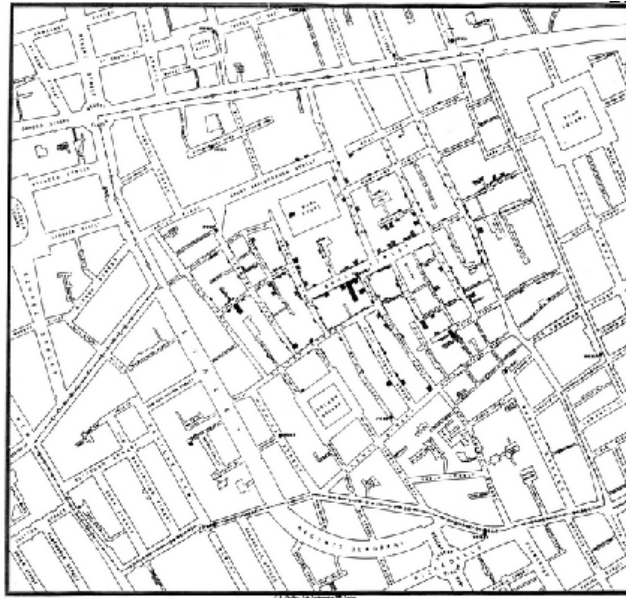


Figure 1.4: Dr. John Snow's map[12].

Chapter 2

Parallel coordinates

2.1 Introduction

Parallel Coordinates is one of the techniques using for visualizing data. Inventor of parallel coordinates is Maurice d'Ocagne in 1885 but real popularization came in 1959 when Alfred Inselberg re-discovered them. They have wide usage in computer visualization, data mining, optimization and many others. It is worth to mention Air Traffic Control where parallel coordinates are used in Collision Avoidance Algorithms. The newest improvement of parallel coordinates came, when Moustafa and Wegman generalized a transformation of Cartesian coordinates system into a new parametric space with coordinate system, which forms parallel coordinates.[1]

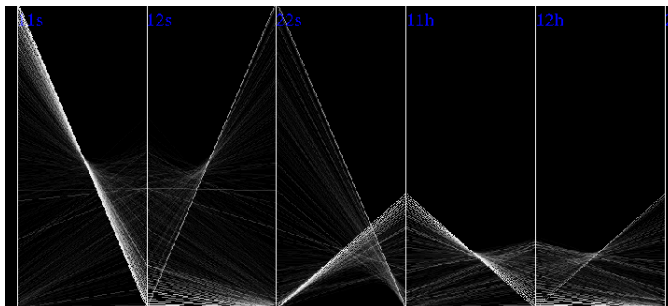


Figure 2.1: Example of data-set in parallel coordinates.

2.2 How does parallel coordinates works

Parallel coordinate system is created from Euclidean plane $\mathbb{R} \times \mathbb{R}$ with Cartesian coordinates $(x,y - \text{axis})$ where N lines, labeled X_1, \dots, X_N , are placed equidistant and perpendicular to the $x - \text{axis}$. These lines form parallel coordinates system for Euclidean N -dimensional space, then point $C(c_1, c_2, \dots, c_n)$ continuous poly-line C . Every poly-line is form of continuous line segments between axes where each point corresponds to one line. In this way large amount of points can be easily displayed and other dimensions can be added. It is easy transformation from $N - \text{dimensional}$ object to $(N - 1)$ lines. For example point in Cartesian coordinates is represented by line in parallel coordinates and the line in Cartesian coordinates is represented by points on this line which can be transformed to parallel coordinates.[1]

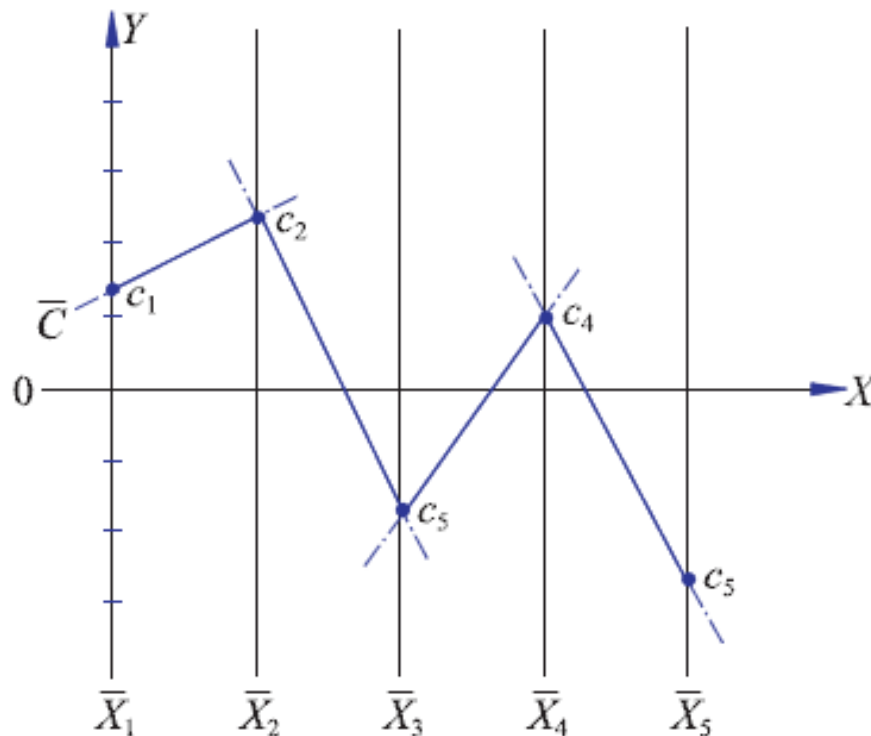


Figure 2.2: One multi-dimensional point in parallel coordinates[1] Figure 1.1

2.3 Usability of parallel coordinates

Broade usage of parallel coordinates makes them suitable for use in large number of sectors. Mainly they are used for analysis of bigger multidimensional data-sets. It is not intended for analysis of records but for analysis of data behavior as a whole. They are used for data analysis to discover different relations in data-sets and link between dimensions.

Density of urban transport can be valuable example for parallel coordinates. Figure 2.3 shows us information about subway in New York underground system. Each dimension represents one year period from 1904 to 2006 and each of 423 lines represents one underground route. The line values itself represents ridership of the route in each year. We can see the biggest concentration of lines in down part what means these routes do not need so many persons to work, on the other hand it may warn of over-employment and more efficient system. This is example of fewer records with high number of dimensions, usually fewer dimensions is used but large amount of records need to be displayed.

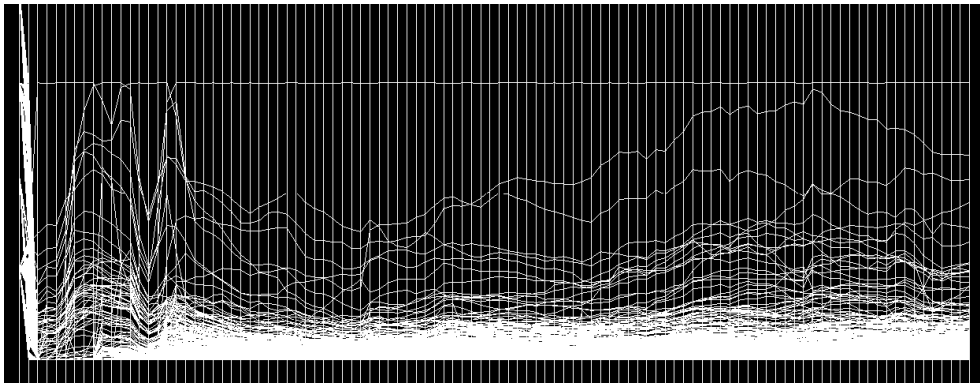


Figure 2.3: NYC subway ridership from 1904 to 2006. Each dimension coresponds to year and each record represents one subway route.

To see better connection between records, see figure 2.4, which describes coal disasters. Each dimension represents information from records. We can see more interesting

facts at once for example, density of deaths in periods, connection between interval and number of deaths.

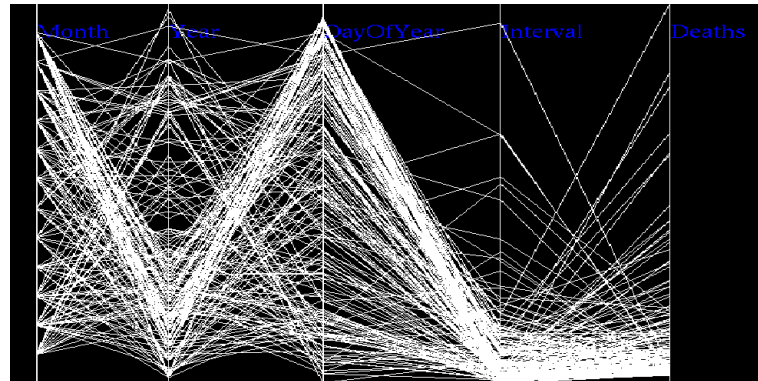


Figure 2.4: Information about coal disasters.

Chapter 3

Methods for increasing legibility

3.1 Blending

Although usually retrieve information from data in parallel coordinates can be useful when more data it starts be illegible. Several methods can be use to increase readability of displaying data. First worth is to mention blending. Blending not only helps to reduce number of visible lines and "clears" the screen but also provides some new information value. Lets compare figures 3.1 and 3.2 .

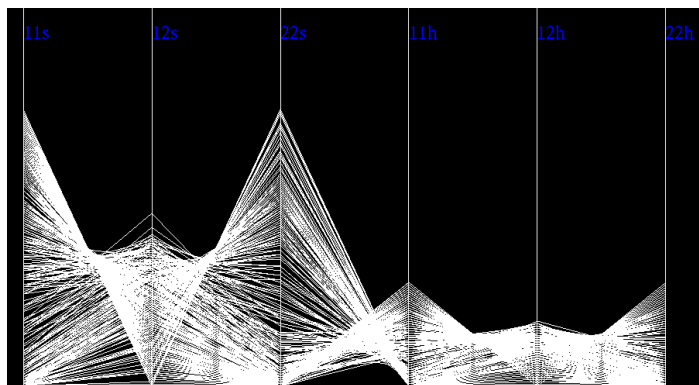


Figure 3.1: Data-set with zero percent blending value. Similar lines overlap. Cannot recognize data concentration.

We can see data are more legible. Note that paler place means multiple occurrence

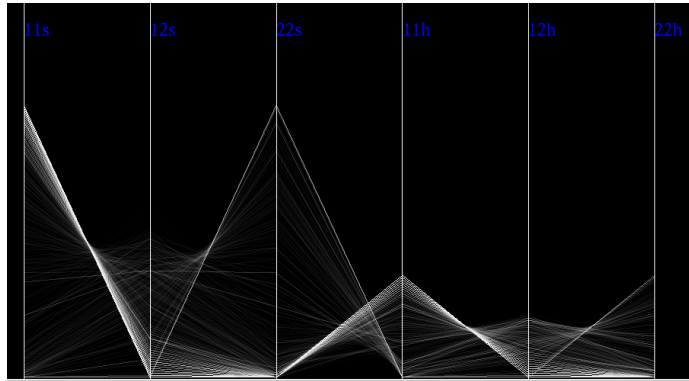


Figure 3.2: Data-set with eighty percent blending value. Similar data shows up and we can see data concentration.

records, what we can use for determination of data density around important points. In this way less significant data disappears and more significant data shows up.

3.2 Coloring

Another useful way of improving parallel coordinates legibility is record coloring. The point is to paint some records to see their progress and to distinguish two groups of records with the similar behavior.

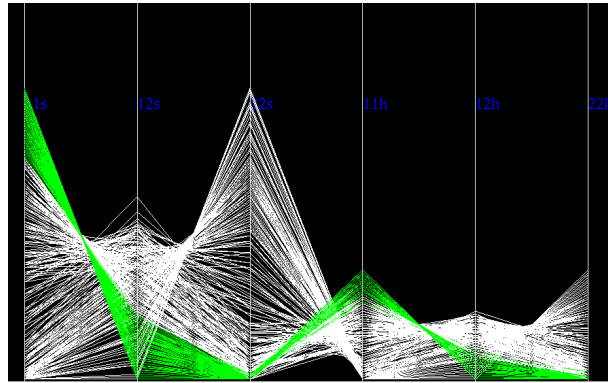


Figure 3.3: Several green colored records that we've outlined the course.

3.3 Combination

These methods can be combined in order to achieve better results in supplying the necessary information. We can highly improve our results and get more legible data-set by shifting dimensions position.

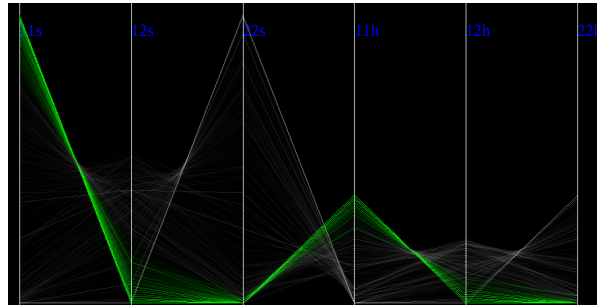


Figure 3.4: Several green colored records that we've outlined the course.

Chapter 4

Dimensional ordering

4.1 Ordering purpose

We showed how to improve the readability of read information, thereby improving the information value which can be received. But as you might have noticed in parallel coordinates one dimension connects a maximum of two other. Now the question is just as determined and rearrange the dimensions in the individual data so that we can get the inner data behavior and thus obtain the best possible information value.

There exists several methods to get dimension ordering, but in our application we implemented two of them. One is called Clutter based analysis and the second one is based on method called principal component analysis (PCA).

4.2 Clutter based analysis

We can see the changing dimension in parallel coordinates may apparently alter the behavior of the records. It is because of parallel coordinates provides easy view only on neighboring dimensions but does not provide connection between foreign dimensions. Normally, the line between the two dimensions of possibility to connect to a clusters, helping to understand connection between dimensions. However, if the dimensions have too many of lines that do not belong to any cluster, then the space between those dimensions is cluttered. The points which do not belong to any cluster are called outliers. Really the point is, if there are too many outliers between dimensions, it means dimensions do not relate to each other too much. So the Clutter based analysis is based on minimizing the number of outliers in our dimension order.[3]

The only problem is to determined if the line is or is not outlier. Outlier is every point which has no neighbor in the desired neighborhood. In my application this neighborhood is not fixed and can be resized. This force us to compare every data-point to each other which need at least $O(r^2)$ steps, where r is number of records. We also wants to treat every combination of two dimensions, that means we need to produce matrix with size n^2 where n is number of dimensions. This matrix is done in $O(n^2r^2)$ time. Unfortunately, as we want to test all possible dimensions order, we need to permute all dimensions what is done in $O(n!)$ time. After removal of permutations we need n steps to evaluate means for every permutation what is totaly $O(nn!)$ time.[3]

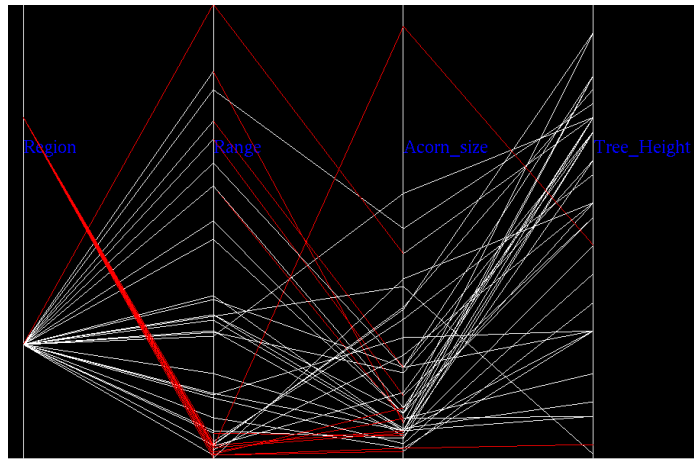


Figure 4.1: Data-set with colored outliers

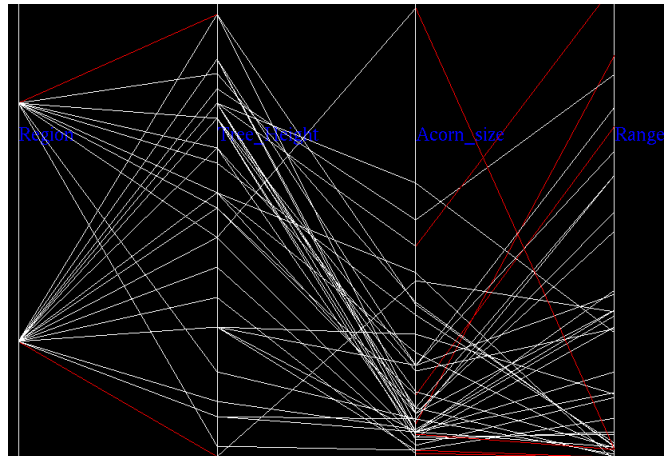


Figure 4.2: Data-set with colored outliers after reordering

We can see rearranged dimensions in order to minimize outlier number.

4.3 PCA

PCA or principal component analysis is a method of processing data in order to highlight similarities and differences of dimensions. This analysis provides advantages such as possible reduction of dimension count and easy to find patterns in multi-dimensional data-sets.[4]

PCA is composed of several steps. First of all we have to subtract mean of dimension from each value of this dimension. Mean is sum of all points divided by number of points.

Next step is to calculate covariance matrix, but first of all we need to know what covariance means and what does covariance matrix mean. Simply put covariance is the number of association between two dimensions. This number inform if the values from one dimension increases or decreases with values from another one and how fast they do so. Covariance matrix is then a matrix of all covariances in data-set, there are covariances from all combination of dimensions.

After we calculated covariance matrix for data-set we have to calculate eigenvalues and eigenvectors. These eigenvectors simply represent lines which characterise data-set and the eigenvalues represents significance of each dimension. The highest eigenvalue is the principle component.[4]

With all data we need to create new data set, which corresponds to transformed data removing less significant information. First of all we order eigenvectors by eigenvalues and form matrix from then and transpose this matrix. We also transpose matrix with data-set loaded and then we multiply transposed vectors with transposed data.[4]

This can be used to order dimension by their significance, and help better understand data-set patterns and main stream of information.

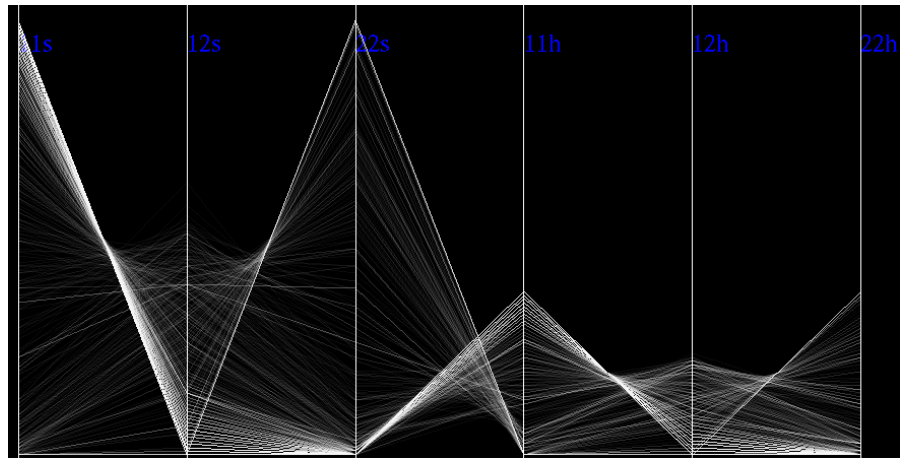


Figure 4.3: Data-set before using PCA algorithm.

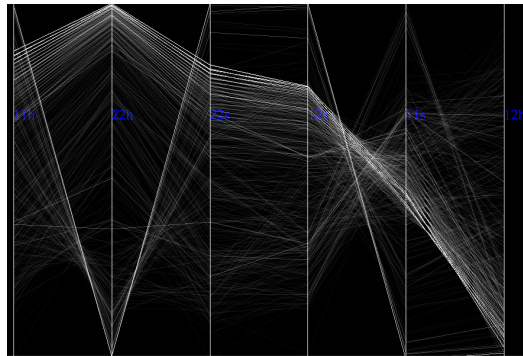


Figure 4.4: Data-set after reordering by significance and covered records.

4.4 Combination of different analysis

These analysis methods can be applied together. It's usual to run PCA analysis and then remove outliers with clutter based one. It leads to better results.

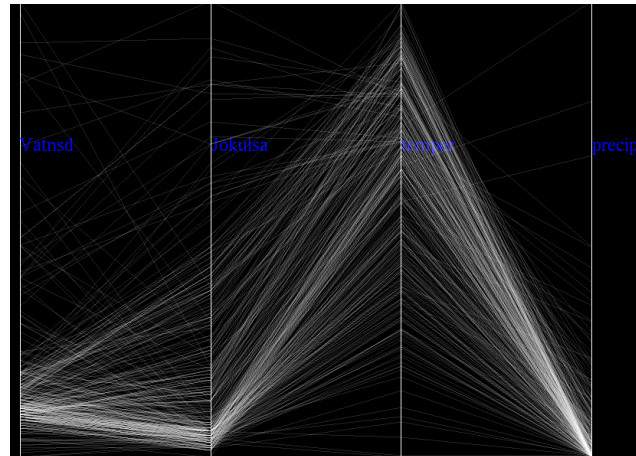


Figure 4.5: Data with no analysis yet.

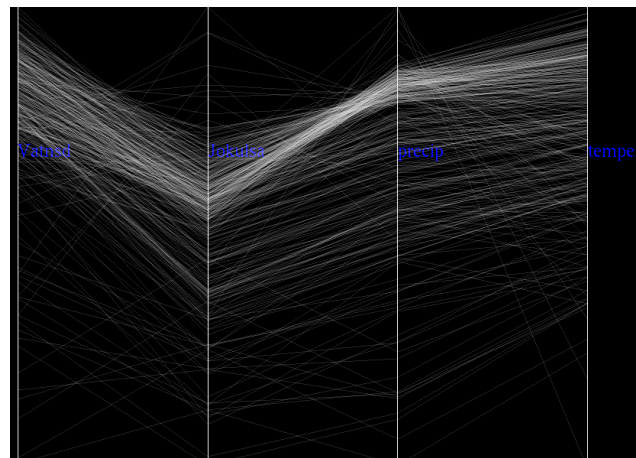


Figure 4.6: After PCA analysis we see reorder dimensions by significance.

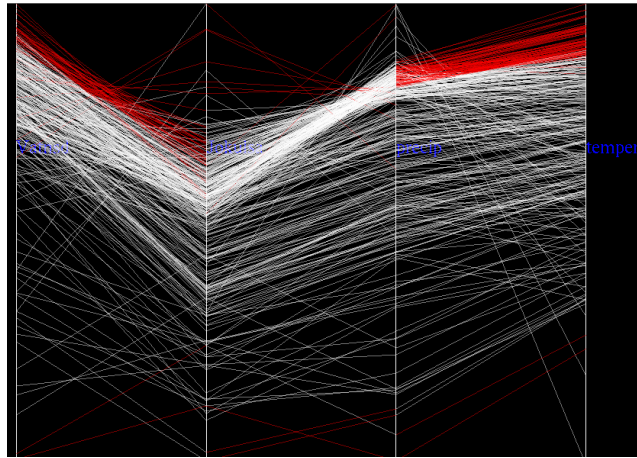


Figure 4.7: Coloured outliers between each dimension.

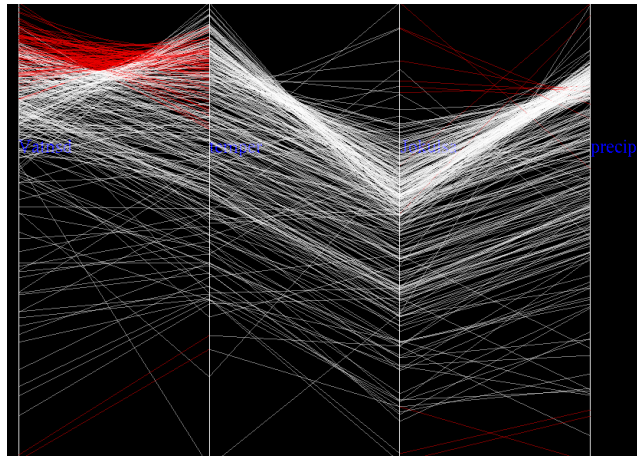


Figure 4.8: Outlier reduction using clutter based analysis.

Chapter 5

Implementation

5.1 Technologies used

The application is written in C++ using OpenGL libraries. The main reason for using C++ was ease of use classes from the program in other applications. C++ is fast and object oriented programming(OOP). At the same time C++ works very well with OpenGL. We also used threads and pre-programmed library for dealing with matrices and statistics. Moreover OpenGL provides effective rendering. Blending is also supported in OpenGL.

We used architecture Model-View-Controller (MVC). MVC separates the processing, display and control data. MVC is often in visual applications, thanks to the separation of the logical part from the interface. This separation helps to use computational part with few changes in various applications.

This application works in multiple threads due to separation of rendering and control section. Multi threads also improves application performance on multi-core PC that are common nowadays.

5.2 Application

Application implements several features including reading file in okc file format[13]. After reading data from file data analysis can begin. Two algorithms are supported, principle component and clutter based analysis. In clutter based analysis the level distance, used to determine the outlier, is set by user using trackbar.

Additional feature is manual moving with dimensions using drag and drop action by mouse. If user moves one dimension to the other they change positions and reconstruction on data-set is done.

Coloring records is done by selecting points of this record on any dimension. After point or points are selected entire record is colored and color value of this record can be changed. Blending values can be set for entire data-set.

Each of these actions (algorithms or visual actions) is working together but independently from each other , so new functionality can be add in future, without interference the others.

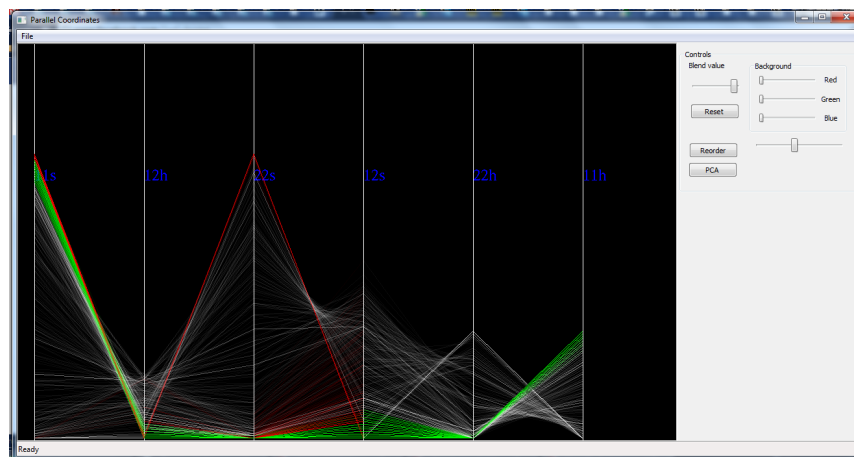


Figure 5.1: Our application for data analysis using parallel coordinates.

Conclusion

Parallel coordinates are very simple and useful way of data visualization. We tried to explain basics of these technique and also explain the necessity of data visualization. There are several ways of increasing information value of visualized data in parallel coordinates. We showed you how to increase legibility of data-sets. We also implemented two data analyzing algorithms in my application, explained how they work and also described purpose of usage these techniques. Even though the use of parallel coordinates is useful, without further sorting coordinates and reordering has information value below for us sufficient number. Therefore we consider coordinates ordering as essential part of data-set analysis.

Bibliography

- [1] Alfred Inselberg. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. 2009.
- [2] Michael Friendly. *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. 2009.
- [3] Wei Peng, Matthew O. Ward, Elke A. Rundensteiner. *Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering*. In *INFOVIS 2004*, pages 89–96, 2004.
- [4] Lindsay I Smith. *A tutorial on Principal Components Analysis*. 2002.
- [5] Parallel coordinate graphics using MATLAB. In *Cornell University*, 2002.
- [6] Hamza Albazzaz, Xue Z. Wang. *Historical data analysis based on plots of independent and parallel coordinates and statistical control limits*. *Journal of Process Control*, pages 103–114 The University of Leeds, Leeds, UK, 2005.
- [7] Yong Ge, Sanping Li, V. Chris Lakhan, Arko Luciee. Exploring uncertainty in remotely sensed data with parallel coordinate plots. In *International Journal of Applied Earth Observation and Geoinformation*, 413–422, 2009.
- [8] Visualization types. 2010 TIBCO Software Inc. <http://stn.spotfire.com/stn/Configure/VisualizationTypes.aspx>.

- [9] Charles Minard's Map. Edward Tufte.
<http://www.edwardtufte.com/tufte/minard> (29.5.2011).
- [10] Venn diagram. Harvest Capital – countries comparison.
<http://www.harvest-capital.com/globalgrowth.htm> (28.5.2011).
- [11] World Density Map. http://en.wikipedia.org/wiki/Population_density
(1.6.2011).
- [12] Dr. John Snow's map. Snow's Cholera Map London.
http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm.
- [13] File Formats of multi-dimensional data-sets. XmdvTool Page.
<http://davis.wpi.edu/xmdv/fileformats.html>.