Immersive Visual Data Mining: The 3DVDM Approach*

Henrik R. Nagel[1], Erik Granum[2], Søren Bovbjerg[3], and Michael Vittrup[3]

[1] NTNU, Trondheim, Norway
[2] Aalborg University, Aalborg, Denmark
[3] ETI, Nørresundby, Denmark

Abstract. A software system has been developed for the study of static and dynamic data visualization in the context of Visual Data Mining in Virtual Reality. We use a specific data set to illustrate how the visualization tools of the 3D Visual Data Mining (3DVDM) system can assist in detecting potentially interesting non-linear data relationships that are hard to discover using traditional statistical methods of analysis. These detected data structures can form a basis for specification of further explanatory statistical analysis. The visualization tools are shown to reveal many interesting patterns and in particular the dynamic data visualization appears to have a very promising potential.

To further explore the human faculties, sound has also been used to represent statistical data. Current technology enables us to create advanced real-time 3D soundscapes which may prove useful since the human ears' field of hearing is larger than the eyes' field of view, and thus is able to inform us on events happening in areas that we cannot see. The audio-visual tools in the 3DVDM system are tested and the effectiveness of them is discussed for situations where sound acts as support for visual exploration, as well as use of sound as a sole cue for analyzing data in VR.

1 Introduction

Most Visual Data Mining (VDM) methods have been designed for PC hardware using 2D graphics or 3D graphics on a monitor, but with the continuous improvements in computer technology, it is now possible with standard high-end PCs to drive Virtual Reality (VR) systems. Such PCs can visualize many objects while simultaneously allow navigation among them and perform intensive background calculations, all in real-time. The aim of the project described in this chapter was therefore to explore how VDM could be enhanced with VR.

The basic visualization technique that was used was the 3D Scatter Plot, where 3D objects with attributes (glyphs) were used as markers. The approach used for designing the glyphs was to make use of the multi-modal audio-visual aspects of VR in order to create attributes that perceptually were distinguishable from each other. While there are many possibilities available for extending the attributes of ordinary glyphs, this project focused in particular on the

* This work was done while the authors were working at the CVMT laboratory at Aalborg University.

S.J. Simoff et al. (Eds.): Visual Data Mining, LNCS 4404, pp. 281–311, 2008.

visualisation of arbitrary temporal developments, as well as on giving the glyphs sound attributes.

1.1 Virtual Reality

VR is a combination of technologies for providing the human senses with sensory input that change as a function of time. The two most important senses are the visual and the auditive senses, but other senses are also regularly used. A standard virtual environment visualises a Virtual World (VW) as though the user is inside of it and therefore monitors the user, so that the sensory input to him/her can change continuously, according to the user's movements inside the visualised VW. VR technology must therefore continuously recalculate the input being sent to the human senses. To exploit VR in the field of VDM, it is therefore natural to focus on the three major strengths of VR:

- Visual feedback
- Audio feedback
- Continuous recalculation of input to human senses

However, to the last point one must notice, that it need not be user monitoring alone that cause input to the human senses to be recalculated. VW's can also be temporal, so that they e.g. change according to an internal rule, a temporal signal or to input from a database.

1.2 Immersive Visual Data Mining

The main principle behind the design of traditional VDM techniques, such as The Scatter Plot, The Scatter Plot Matrix, The Grand Tour [1], The Parallel Coordinate Plot [2], etc. is that they are viewed from the "outside-in." In contrast to this, VR lets users explore VWs from the "inside-out" by allowing the users to continuously navigate to new positions inside of these, in order to obtain more information from them. VR applications provide more comprehensive input to the human senses and can therefore make more efficient use of the human perceptual skills, when exploring large datasets.

The potential benefit of immersive VDM has been debated [3]. We discuss the existing views in section 2.3. Our hypothesis is that a complementary and valuable benefit is achievable concerning the detection of e.g. non-linear relationships in data, which are most likely to escape traditional methods of data analysis. To support our hypothesis, this chapter presents the immersive VDM methodology and the corresponding tools, available in the latest version of the 3D Visual Data Mining (3DVDM) system [4,5].

We study an example case with the visualizations tools and show how they can help discovering special relationships in data that may require further statistical analysis, by e.g. domain experts. The aim of the chapter is to support the argument in favour of immersive visual data mining, but we would like to stress the limitation of the paper media in terms of illustrating the full scale of the process and the methodology. We also do not focus on the statistical and practical importance of the phenomena in the data, since this is beyond the scope of the chapter.

1.3 Exploiting Sound

The 3DVDM system provides a flexible system for studying immersive VDM. The optimal VR system for this purpose (such as the six-sided Cave) provides technology allowing us to create an immersive environment of surrounding 3D visual cues, as well as the possibility of creating immersive 3D soundscapes.

The intention of using sound in data mining is derived from the underlying idea of the 3DVDM project, which is to encode as much information as the human perceptual system can cope with [6]. The idea is to create a 3D sound interpretation of the data that can serve either as a support for a visualization parameter such as color or as a stand-alone audio representation of one or more data dimensions. A main motivation for this approach is that the human ears' field of hearing is larger than the eyes' field of view, and thus is able to inform us on events happening in areas that we cannot see. It may then be possible to create a sound cue that will draw the listener's attention towards a part of the visualization that is out of visual range. In VR, the user will then be able navigate using sound cues and perform a more detailed investigation of areas of interest.

Our aim is to create a soundscape that in some way represent the surrounding visual world and may be of assistance to the user in several ways. This chapter will therefore also describe how sound can be helpful for performing visual data mining and in addition support orientation and navigation in the artificial worlds that we deal with.

2 Previous Work

2.1 Visual Data Exploration

VDM methods, such as "The Grand Tour", have been implemented in VR in several occasions [7,8], as well as the traditional method for data exploration called "Brushing and Linking" [3].

A well known approach for visualizing multivariate data is to map data variables to visual object properties, such as position, size, shape, color, orientation, etc. Such visual objects are called glyphs [9,10,11], and can be in both 2D and 3D. Glyphs are efficient due to the ability of the human brain to discover patterns within and among objects, as well as to recognize objects that do not "fit" into a discovered pattern. Methodologically, glyphs address the problem of visualizing multivariate data (multidimensional data) with a higher number of dimensions than 2–3 into 2D or 3D space.

An example of the use of glyphs is "Chernoff Faces" [12], in which each observation in a dataset is represented by a cartoon face of which features, such as length of nose, curvature of mouth, size of eyes, and even the shape of the face itself, correspond to variables of an observation. Colored texture has also been used to visualize multivariate data elements arranged on an underlying height field [13]. Using volume visualizations of 3D scatter plots with glyphs representing data point [14], it is possible to use procedural shape generation techniques.

These techniques allow from 1 to 14 additional data dimensions to be visualized using glyph shape.

It is important to note that the success of a particular visualization technique depends upon the analysts' *comprehension* of multivariate data. It is therefore not a goal in itself to visualize as many variables as possible simultaneously, but rather to make it possible for analysts and domain experts to get a useful impression of relationships between multiple variables in a dataset.

2.2 Auditory Data Exploration

It is well known that the visual modality of the human senses provide a powerful means of categorizing objects, e.g. when performing visual data mining. However, some research has been focused on using auditory patterns to present complex information and combining visual and auditory information to facilitate the processing of information [15]. It is concluded that for some types of information, the auditory modality works just as well as the visual modality.

In particular, auditory information can be used as effectively as visual information for a visual search task when speed is not crucial. Furthermore, it is suggested that humans rapidly can extract more than one piece of information from a sound, and then act on the information. These results indirectly suggest that it will be possible to create a useful 3D soundscape, that in some way represents the visual world and yields further information about this.

Much research has been focused on creating auditory displays using different artificial cues for simulating direction and distance. [16] provides a useful overview of things to consider when implementing a 3D sound system, especially regarding distance perception and implementation. Also [17] is a useful source when dealing with the human auditory system.

2.3 Immersive vs. Traditional VDM

In a project from 1999 by Nelson, Cook & Cruz-Neira [3], a thorough comparison between a 2D and a 3D VR version of the VDM software XGobi was performed. The 3D VR version of XGobi was called VRGobi. While the 2D environment featured a traditional graphical user interface, with menus, data windows and dialog boxes, the 3D VR version was essentially a recreation of the 2D environment, but with an added dimension. There was therefore a virtual floor, on which the subjects stood, a virtual, person sized cube, in which the data was visualized and a flat color palette positioned to the right of the user.

In the project, fifteen subjects, mostly statisticians, were asked to complete four tests in each environment. The tests were:

1. Visualization Tests
   (a) Accurately Detecting the Number of Clusters
   (b) Detecting Intrinsic Dimensionality
   (c) Radial Sparseness – Hollow vs Solid Sphere
2. Ease of Computer–Human Interaction
   (a) Brushing Data Points with Color

Concerning the visualization tests, the paper states: "The C2 provides a slight advantage over XGobi in the sphere test, and a large advantage in the cluster test. Subjects performed equally well in both environments on the dimensions test, to such an extent that it indicates that the test was too easy."

However, seen from the point of view of authors of this chapter, it is problematic that the design of the VR version of the environment forced the test subjects to view the data from an "outside-in" perspective, rather than allow them to freely navigate around inside of the visual representations of the data. If a 3D shape, such as an S-shape, for instance at a time would emerge, when the data cube was viewed from the top, it is highly unlikely that the test subjects would detect it, since they would view the S-shape from the side and thus see it as a simple line. Allowing the test subjects to navigate freely around inside of the data would also be a great advantage in the "Hollow vs Solid Sphere" test, since the test subjects in VR simply would navigate inside of the sphere and check whether it is hollow or not. However, it would also mean that they would require additional time for navigation, but the potential benefit is that their possibility for detecting shapes that are far more complex than simple spheres would increase dramatically.

Concerning the interaction test, the paper states: "The brushing times were significantly lower when XGobi was used." and later "The problem may be in the user interface design that we developed so re-thinking the design may improve the situation."

We agree with the last quotation. Being inside a VR environment and choosing colors by pointing at them from a distance using a VR interaction device is difficult compared with moving a precision mouse that is resting on a mouse pad. A much more efficient VR interaction technique for selecting colors would be to map the three fundamental colors, red, green and blue, to the three axes in a small coordinate system that is visualized in front of the test subjects, when they press a button on an interaction device. A small sphere in front of the interaction device would then change color, as the interaction device moves around inside of this small coordinate system. In this way, any color can be chosen easily and when the test subjects let go of the special button, the color for brushing is fixed and the color choosing coordinate system disappears. This technique is much faster to use and far more flexible than the technique described in the paper.

It is important to consider that the authors of the paper discussed in this section were early pioneers in the use of VR in a VDM context. The point is therefore not to criticize their work, but rather to point out that, when using VR in a VDM context, there is no implicit guarantee, that one, with success, can make immersive VDM versions of traditional 2D VDM techniques. It is, instead, necessary to re-think the basic principles that one brings into a VR environemnt from work done with traditional VDM techniques and this is what has been attempted in the work described in this chapter.

3 The 3DVDM System

The 3DVDM system that is presented here is the second generation of a software system for exploring data in VR. While the first generation of the software [4] pre-rendered visualizations of data in a time-consuming process, this second generation of the software renders visualizations of data in real-time, while users are interacting. This, potentially, opens up for many new possibilities, of which some are mentioned here.

3.1 VR++ and 3DVDM

To perform experiments in VR, a general-purpose VR software framework called VR++ has been developed [18,19], which supports VR visualisation, interaction, encapsulation, modularity, inter-disciplinarity and distributed computing. VR++ runs on UNIX based computers and current PCs are sufficiently powerful to run the system for display on standard monitors, or if networked, drive CAVE or Panorama visualization arenas.

VR++ is specialized for creating VR applications with real-time computed animations with many frames per second. This animation form requires data visualizations to change so frequently, in response to changes in both data and user input, that users experience a smooth animation, which also reacts appropriately to real-time interaction, such as head movement.

The 3DVDM system has been developed on top of VR++. While VR++ provides functionality for e.g. parallel-processing, communication, parameter control, and visualization of geometric objects, the 3DVDM system provides support for specialized VDM tasks. VDM tools are created by connecting suitable modules from both VR++ and the 3DVDM system.

In 3DVDM, some of the new temporal data exploration methods, based on real-time calculation of arbitrary temporal developments, have been implemented. One of the explored approaches is to enhance glyphs with motion attributes. Another is to visualise continuous streams of data by continuously deleting and recreating the graphics in 3D scatter plots, as new data arrive. To create a clear temporal development, the visualised data is sorted according to one of its statistical variables, so that the sort variable, in effect, is mapped to user time.

3.2 Data Pipeline and Interaction

Figure 1 shows the general approach adopted in this research for visualizing representations of data from databases.

The system contains different data processing modules in a pipeline with the possibility of feedback from users to each module. The bottom four arrows in Figure 1 correspond to four different kinds of interactive, feedback-loops. According to the methodology, the shorter the loop is, the faster and easier to use must the corresponding interaction technique be. The “Visualization Control” feedback loop has therefore been mapped directly to a VR interaction device that

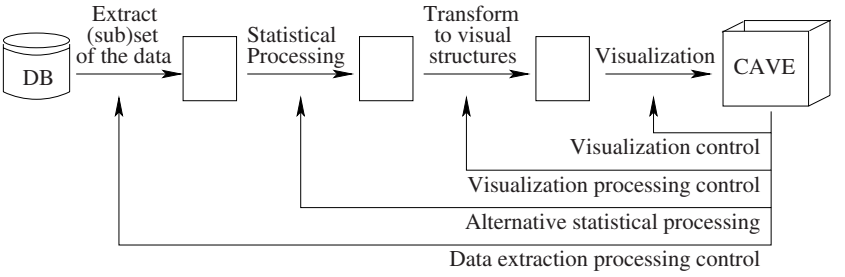


Fig. 1. The 3DVDM Data flow and interaction patterns

analysts hold in their hands at all time during a data exploration session. The other three, longer, feedback loops, are, however, handled by a control panel, which in real-time and with visual feedback allows for modification of the parameters on which a visualization is based.

Accessing Databases. First, a relevant subset of the data in a database is extracted and stored as an easy accessible, internal database, which is passed on for processing to later stages of the pipeline. The 3DVDM system has specially designed software for loading "comma-separated-values" (CSV) files, in an efficient way. This is accomplished by performing a complete analysis of the data the first time they are loaded. The result of this analysis is stored along with the data in binary format, beside the original data file. When the dataset is loaded again, the result of the analysis phase is loaded instead of the original data. This allows fast loading of large datasets, once initially analyzed.

The data handling part of the 3DVDM system is designed with real-time rendering in mind, rather than with handling of large databases in mind. This means that emphasis has been put on efficient storage of data in memory (RAM), and efficient extraction of data from memory, rather than on, e.g. on-line access to data from harddisks.

Preprocessing. The next stage in the pipeline is statistical processing. The current version of the 3DVDM system provides a set of facilities for statistical computations and preprocessing. However, while this topic in general is very important, the research reported here has emphasis on visualization facilities.

Transformation to Visual Information. In this step, immersive VDM algorithms process the data and transform it into an equivalent symbolic graphical representation. This symbolic data is independent of specific hardware and software requirements.

Visualization. Last step is to transform the symbolic data to polygons, which subsequently are rendered using OpenGL in a 3D space for visualization. An important part of the methodology is that it at this step must be possible to

replace all polygons being displayed many times per second. This is required in order to make it possible to follow arbitrary temporal developments.

3.3 Principles of Data Visualization in 3DVDM

The 3DVDM approach to data visualization is built around the possibility to navigate around in a Virtual World (VW) - an "Extended 3D Scatter Plot" - to explore arbitrary view directions from arbitrary view points. "Extended" means that data points are visualized as objects that may have different visual properties.

The ease, with which navigation occurs, highly depends on the frame-rate and therefore the number of visual objects in the VW and on the complexity of each visual object. It is an advantage to be able to visualize as many objects as possible, and this means that the complexity of each visual object must be reduced as discussed in [4].

Another consequence of exploring data in the above VW is that attention must be paid to the distance between VR users and 3D visual objects currently observed. This is discussed in [6], where it is suggested that object size is kept constant to allow the perceived size to support depth perception. Object size is also the reference for a spatial distance perception in general, and for design and evaluation of perceptual conditions. Some properties like surface texture should be observed at close range (measured in object size units) to support a perceptual grouping, object shape still works at "medium" range, while color is much more dominant perceptually and can be seen to define cluster structures also at long range distances of observation.

Dynamic Visualization (DV) of data is a new feature of the 3DVDM system. It can be implemented in several ways, e.g. precomputed animations and real-time computed animations. For each of these two cases, there are two possibilities of interest: few frames shown for several seconds each, or many frames shown for a fraction of a second each.

3.4 Rendering Sound

The 3DVDM system also provides tools for using 3D sound to perform general and detailed investigation of data visualized in a scatter plot. A sound engine is designed and integrated into the VR++ system. With this, it is possible to place sampled sounds or synthesized sounds at any point in the virtual world. It is thus possible to attach sounds representing statistical values to selected objects, or add sound representation to groups of objects, such as density clusters or measurements of density in solid angles around the users current position. Adding sound to this infinite visual space with thousands of generators and with no major solid structures to influence with reverberation, as in the natural world, has appeared to be a rather complicated task with a large number of possibilities and constraints.

Databases of concern typically have many thousands of observations, and our current software can handle 32 simultaneous sound generators. Even if this

may seem insufficient for creating an interpretation of thousands of observations, we need to consider that a soundscape created from thousands of simultaneous sounds will not necessarily yield useful information (it is more likely that it will not). The solution to this problem is to create a soundscape that changes during time, allowing the listener to be able to pinpoint locations in space, which may be interesting to explore even further. For this purpose the number of voices is sufficient. The change in time can either be an automated process, or controlled by the user in ways that will be described in more detail later.

The sound system can produce a sound field generated with 2, 4, 5 + 1 subwoofer (5.1) and 8 speakers in different configurations. The optimal 3D sound field is generated with 8 speakers placed in each corner of a cube, because this creates an even placement of the speakers in all directions of the listener, and represents all three directions in the 3D world. This configuration is widely used with CAVEs.

4 Basic Tools

4.1 Visual Data Exploration

Overview. The initial problem is to find and select one or more sets of three suitable variables (triplets) to define the coordinate system and hence the spatial layout of the objects in the 3D space.

One of the most basic 2D visualisation techniques, for this purpose, simply shows all possible, different combinations of two variables of a dataset. For each combination, a small histogram is shown, where the colour of each cell is mapped to the number of records that fall into the cell's value range for each of the two variables. An example of this is shown in Figure 2.

Selection. When a set of candidates is selected, a "scatter plot tour" facility of the 3DVDM system allows systematic inspection of all unique combinations of triplets. Using a control variable encoded as object color, each candidate triplet may be investigated in as much detail as desired, but given the typical number of candidate variables, a first run through may be used to eliminate the least interesting variables/triplets.

Analysis. Having selected one or some triplets of interest, a more detailed 3D scatter plot analysis and visual exploration can take place. Alternative sets of the other variables can be assigned to object properties, and the user can navigate around and observe the visual world from "inside-out" or "outside-in" as he or she pleases. Being observant throughout the above process, it is very likely that interesting (sub)structures in the data will be observed.

While the above relies on a static VW within which the user can navigate around, the system now also allows dynamic visualizations of data. Color scales can cycle with real-time feedback, and a "Macro Dynamics" facility allows a window sliding through a sorted data base, such that a sort-variable in a sense is

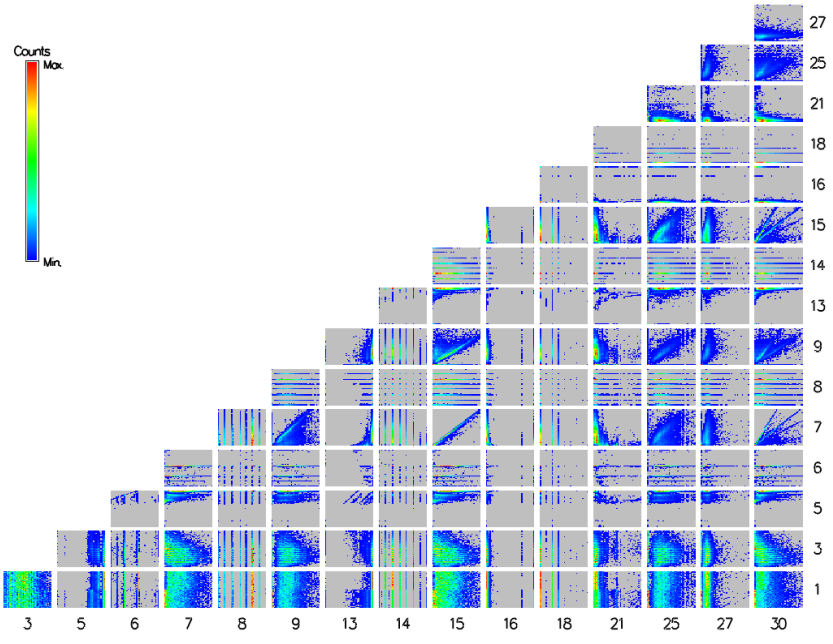


Fig. 2. An auto-scaled flat histogram for each of the interesting combinations of 2 variables in a data dataset with 30 columns

mapped to a time scale. A "Micro Dynamics" facility allows statistical variables to control movement of all objects individually.

Exploration. Navigation in the virtual 3D space means here to control both the viewpoint and view direction as one pleases. Hence one may "fly" around and within the visualized coordinate system and observe the objects.

The navigation interface and display of the current view are all dependent on the visualization system used. When using the CAVE stereo visualizations are provided and all view directions are available for the user.

The Panorama is a popular visualization system for our VDM, and navigation is controlled according to the direction of a "Wanda", i.e. a device that is tracked with 6 degrees of freedom (position and orientation). Here view direction is fixed to the (forward) "motion" of the navigation.

The most important feature concerning immersive VDM is the real-time response to user movement while intensive background processing is being performed, - something that requires the VDM software to make use of parallel processing.

4.2 Auditory Data Exploration

The human auditory system is capable of pinpointing a single sound source in an otherwise complicated soundscape. This is referred to as the "cocktail party

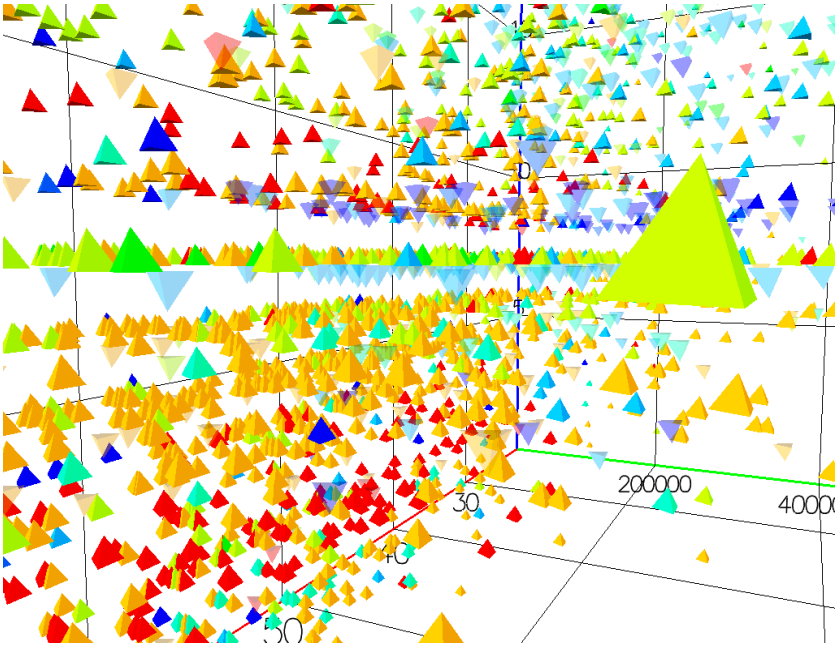


Fig. 3. A 3D Scatter Plot of mortage loans, mapping "interest rate" to color, "leaving customer" to shape, and "contribution rate" to opacity

effect"[17]. This means that we can encode much information into the soundscape and expect the listener to be able to decode a single source, as long as it is distinguishable from the rest of the sound sources. This may be achieved by constantly changing the soundscape - a side effect that will happen automaticly when the data base is sampled in time (only requiring that the database contains different observations).

Likewise [17] speaks of the "ventriloquism effect". This phenomenon indicates that if there is a visual cue and a sound is located close to this object, the sound is automatically perceived as being where the visual object is. This means that if we visualize the currently sampled objects the listener should perceive the sound as originating from this object. This may compensate for possible shortcomings in the 3D sound rendering system.

When placing sound cues at different distances one should consider that our auditory system perceives distance far more accurate when using familiar sounds than for unfamiliar sounds [20].

Based on this, we propose using two types of sounds: sound samples and synthesized sounds.

Sampled Sounds. Using sampled sounds where an exchangeable sound bank consists of selected samples that may (or may not) be related in a way that makes them yield some kind of numerical information. If the samples have numerical

information, such as the sound of the numbers from 1 to 10 spoken out loud, the samples can be used to represent variables with a few, ordered values, while samples without any inherent numerical information, like e.g. animal sounds or the recordings of different musical instruments, would be best suited for representing categorical variables. The samples should, however, be short (i.e. less than 500ms) to avoid a clouded soundscape.

Synthesized Sounds. Using synthesized sounds generated with a simple sawtooth waveform with musically tuned pitch. Low value can correspond to a low pitch and visa versa. The sounds are created with a relatively short envelope time consisting of a short attack around 30ms and decay around 100ms. The resulting sounds are short musical sounds with a noticeable onset, that should create sufficient attention for the listener in order to locate it's position and perceive its encoded value. In the non-categorical case it may be more difficult to define the ideal sound sample bank. For this purpose it may be more suitable to use these synthesized sounds and achieve value information by mapping the values to musical pitch. The pitches are chosen in predefined scales, which have different spacing between the pitches and different tonal content. The 3DVDM system supports a few of these scales to which values can be mapped. The scales are: Pentatonic, Aeolic, Major Chord, Chromatic, Melodic (Ionic) and Octave (Boolean). In some scales the tones are close to each other (Melodic/Chromatic) while the others have more or less spread (3-5 semi-tones between them). The Octave scale only has two tones placed in the distance of an octave. When using the synthesized scales, the soundscape will most likely have some kind of musical content.

We choose to map values to musical pitch, because most people are familiar with scales and consequently should be able to distinguish between different tones and interpret them in the context of an exploration process. Musical pitch is, however, an entire field of research and there exist theories about which tones that, perceptually, are considered to be related and vice versa. Making experiments with the perceptual aspects of musical pitches is, however, outside of the scope of this project and we will therefore here limit ourselves to making use of the concept of musical pitch.

Our wish to encode easy understandable numerical information into the soundscape introduces a problem that one must consider. Under normal conditions, the auditory system is treated with many *different* sounds. In our case, the soundscape will be created from many *similar* sources. This means that events are more sensible to masking, which means that sounds with higher intensity "absorbs" the weaker ones. We may overcome this effect to some degree by choosing many different sounds or by lowering the number of simultaneous sounds.

In practice, the 3DVDM system has control options to adjust such parameters to individual perceptions. These settings can be applied in real-time depending on what the listener may find useful in a given situation.

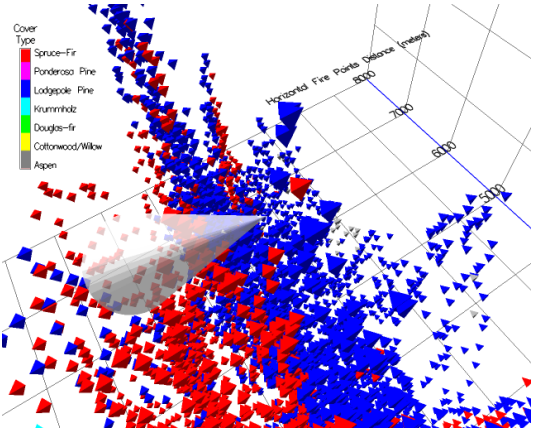


Fig. 4. Data mining with a virtual torch

Tools for Rendering a Soundscape. There are two major methods which both operate on visual 3D scatter plots:

- A render tool, which can be used for rendering a 3D soundscape around the user representing either statistical values or density measured in a user definable cubic grid. The soundscape rendering can be of the whole data set with random sampling, sliding windowed data sampling or rendering a sweep of a specified axis in time. The database is sampled several times each predefined time interval, e.g. each 10ms, and if the threshold conditions for a sample are met a sound will play. At each sampling period, a specified number of voices (1-32) will be triggered on this basis.
- A torch tool, which is placed in the hand of the user. In VR the user can point this virtual torch towards objects in space and get sound responses from these representing statistical values or local area densities (see Figure 4).

When a data sample is set to play it is rendered at its actual position in 3D space. Ideally the listener will be able to locate the origin of the sound (position and distance). It is possible to apply density thresholds in order to concentrate the attention to higher populated areas. Likewise the user may define a distance threshold and discard records placed in virtual space beyond this threshold.

5 Methodology for Visual Data Exploration in 3D Worlds

To demonstrate the facilities and the potential of the 3DVDM system a specific dataset is here described and is followed by a pre-analysis.

Table 1. Basic information about the variables in the Forest Cover dataset

Name	Index	Values	Range
Elevation	1	1978	1859 – 3858 meters
Aspect	2	361	0 – 360 azimuth
Slope	3	67	0 – 66 degrees
Horizontal Hydrology Distance	4	551	0 – 1397 meters
Vertical Hydrology Distance	5	700	-173 – 601 meters
Horizontal Roadways Distance	6	5785	0 – 7117 meters
9am Hill-shade	7	207	0 – 255
Noon Hill-shade	8	185	0 – 255
3pm Hill-shade	9	255	0 – 255
Horizontal Fire Points Distance	10	5827	0 – 7173 meters
Wilderness Area	11	4	Cache la Poudre, Comanche Peak, Neota, Rawah
Soil Type	12	40	1 – 40
Forest Cover Type	13	7	Aspen, Cottonwood/Willow, Douglas-fir, Krummholz, Lodgepole Pine, Ponderosa Pine, Spruce-Fir

5.1 Data Preparation

A publicly available dataset called "Forest Cover Data" is used[1]. The dataset was used to predict the forest cover type for 30×30 meter cells on the basis of cartographic variables obtained from the US Forest Service (USFS) Region 2 Resource Information System (RIS) data [21].

Data Summary. The following information is from the dataset documentation:

Number of observations:	581012
Attribute breakdown:	10 quantitative and 2 categorical variables (Wilderness Area and Soil Type)
Independent variable:	Forest Cover Type
Missing Values:	None

The dataset is not balanced with respect to the dependent variable, where the number of observations ranges from 2747 to 283301 for the individual forest cover types; refer to the web-site for more information and description of the variables and basic statistics.

Data Conversion. There are 13 variables of which the last three are categorical (qualitative), that is, without any meaningful ordinal order. The categorical variable Forest Cover Type, index 13, is encoded numerically as 1 to 7 and can directly be used. Index 11 and 12 are encoded as 44 binary variables (4 mutually exclusive values from Wilderness Area and 40 from Soil Type) were converted to two numerical variables, as seen for index 11 and 12 in table 1.

[1] http://kdd.ics.uci.edu/databases/covertype/covertype.html

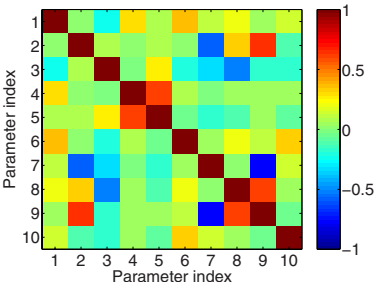


Fig. 5. Correlation coefficients σ_{xy} for the variables, index 1-10

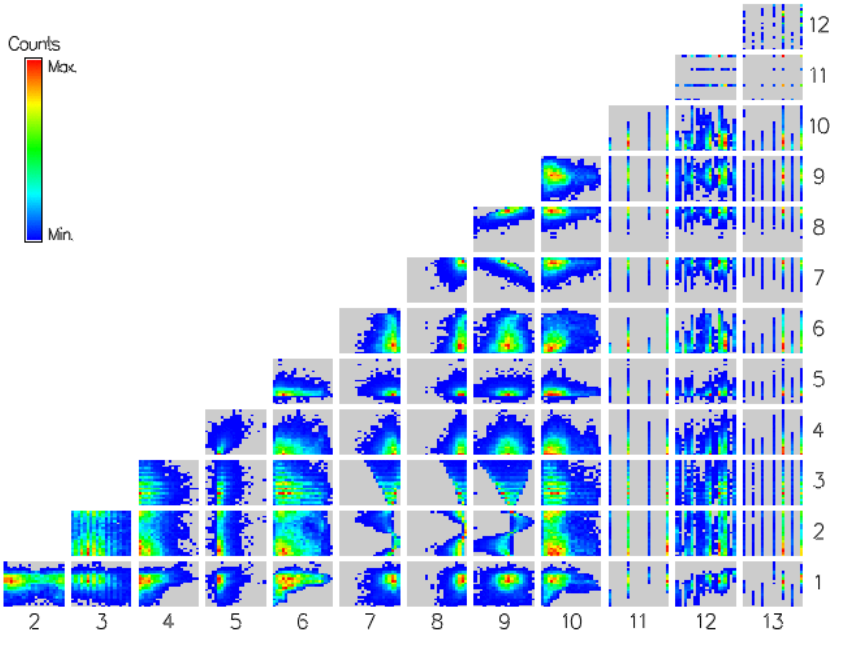


Fig. 6. An auto-scaled 2D histogram for all combinations of two variables

The variable "Soil Type" is generally excluded from further consideration in this study due to interpretation ambiguities.

5.2 Basic Statistical Analysis

A standard statistical package is used to provide basic information of the data set. All histograms are uni-modal and most have skew distribution. Index 5 and 7 have very narrow distributions compared to the range, and may not have a lot to offer. Index 2 (Azimuth, 0^0-360^0) has a "circular" distribution peaking at 50^0 and minimum at 230^0.

Correlation Analysis. Figure 5 shows a color encoding of the correlation coefficients σ_{xy} for the 10 quantitative variables of the dataset. Note that a correlation coefficient σ_{xy} of -1 is just as strong as a correlation coefficient of 1.

Strongest (positive or negative) correlation coefficients are found between Hill-shades, $\sigma_{79} = -0.78$ (9am/3pm Hill-shade), and $\sigma_{89} = 0.59$ (Noon/3pm Hill-shade), while $\sigma_{78} = 0.01$ (9am/Noon Hill-shade) is surprisingly weak. Other strong correlations are $\sigma_{45} = 0.60$ (Horizontal/Vertical Hydrology Distance) and $\sigma_{29} = 0.65$ (Aspect/3pm Hill-shade).

If the estimation of the linear correlation is too strong, one of the involved variables could be discarded in order to avoid redundancy. However, non-linear relationships may underly the computed correlations.

2D Histograms. The 3DVDM system's 2D histogram facility is used for showing all unique combinations of two variables of the data set. This is shown in Figure 6, with color mapped to the number of records in each cell.

The 2D histograms reveal both linear and non-linear relationships. As the number of variables is relatively low, we retain all 10 quantitative variables as potential candidates for mapping to the coordinate axes in "interesting" triplets. Soil Type, index 12 is discarded from further investigation here, while index 11 and 13 may serve as "dependent" variables in our investigations below.

6 Visual Exploration of Static Worlds

We will here discuss a "tour" facility of 3DVDM, and how to use the results from this "tour" to the best of its account.

6.1 3D Scatter Plot Tour

We have retained all 10 quantitative variables as candidates for "spatial" triplets. There are, in general, many possible combinations of mappings of variables to the axes of a 3D scatter plot. In our example, the combinatorics leaves us with the following number of unique triplets:

$$\frac{10 \times 9 \times 8}{1 \times 2 \times 3} \quad = \quad 120 \tag{1}$$

The tool displays one 3D scatter plot visualization for each triplet with the mappings to the three axes of the coordinate system changing at regular intervals, while the color represents the dependent variable. One can, however, also assign fixed mappings to the variables for the three axes of the coordinate system, and let the mappings to the other object properties change regularly.

The tool allows the user to rank the visualizations according to how "interesting" they look, by manually assigning a score from 0 to 9 to the individual combinations. These combinations, together with their score and basic statistical information, are stored in a log-file for easy access and analysis.

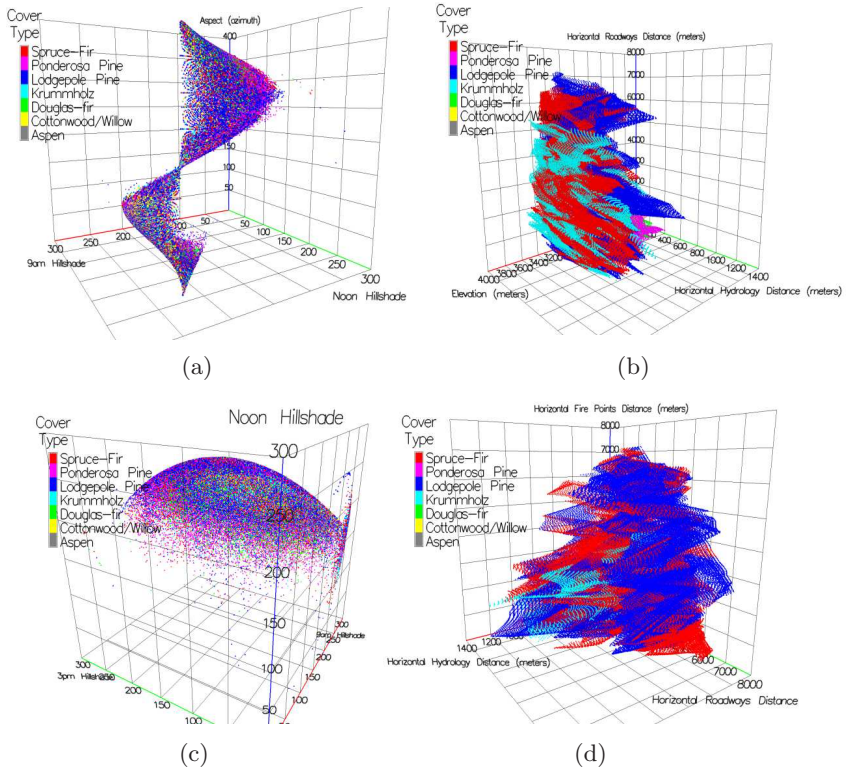


Fig. 7. Examples from 3D scatter plot tour with color representing Forest Cover Type

All of the 10 variables were involved in "interesting" visualizations, but variables 1, 3, 4, 6, 7, 8, 9, and 10 were dominating. Conclusion of the tour is to continue with closer analysis of the triplets 1, 4, 6 (Elevation, Horizontal Hydrology Distance, and Horizontal Roadways Distance) and 7, 8, 9 (9am, Noon, and 3pm Hill-shade).

6.2 3D Scatter Plots and Object Properties

Once interesting triplets have been found, further analysis can be performed in VR. This section briefly describes how to use some of the scatter plot features of 3DVDM, e.g. the programmable dynamic color scales.

Extended 3D Scatter Plot Analysis. We choose two triplets after the tour, and starting with Figure 7(b) we can now experiment with different color mappings. It may be of interest to see how the Wilderness Area variable is distributed when using this triplet, and we therefore choose to map this variable to color instead of Forest Cover Type, as shown in Figure 8.

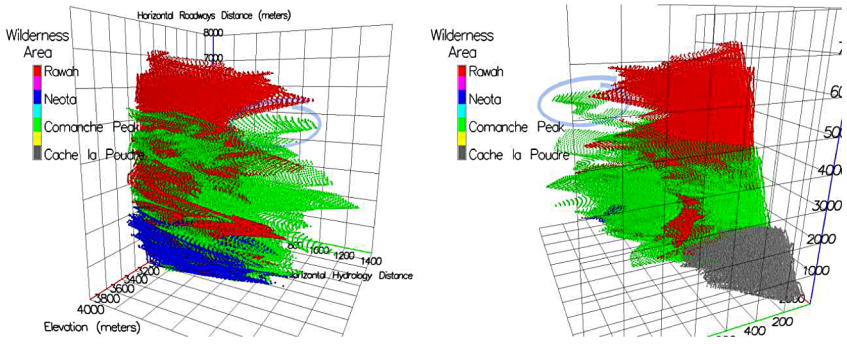


Fig. 8. Same mapping to axes as in Figure 7(b), but with Wilderness Area mapped to color. The figure is seen from two different viewpoints.

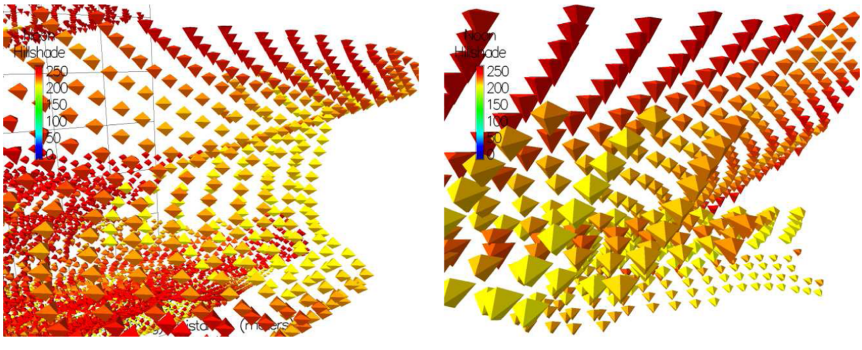


Fig. 9. 3D scatter plots with Noon Hill-shade mapped to color and slope mapped to shape. This figure is a close-up on the marked substructure in Figure 8.

For this triplet, two of them (elevation, horizontal roadways distance) show relevance for the 2D distribution of Wilderness Area, while the third (horizontal hydrology distance) drags out some specific substructures.

Taking the visualization further, we can explore relationships between more variables, e.g. by mapping other variables to object properties like object shape, orientation, brightness, opacity, etc. Figure 9 shows a closeup on the marked area in Figure 8, again seen from two different viewpoints.

These sub-figures have same spatial mapping as in Figure 8, but this time with Noon Hill-shade mapped to color and Slope mapped to shape. The snapshots are seen from "inside-out", in contrast to earlier figures which were viewed from an "outside-in" point of view.

Exploring Color Mappings. Now we will use Figure 7(c) for further visual data exploration. Figure 10 is a scatter plot of the variables 9am, Noon and 3pm Hill-shade (index 7, 8 and 9) mapped to the axes, now with slope (index 3) mapped to color. It is now very clear that a correlation *does* exists, although σ_{78} was computed to be only 0.01 (Figure 5 in section 5.2).

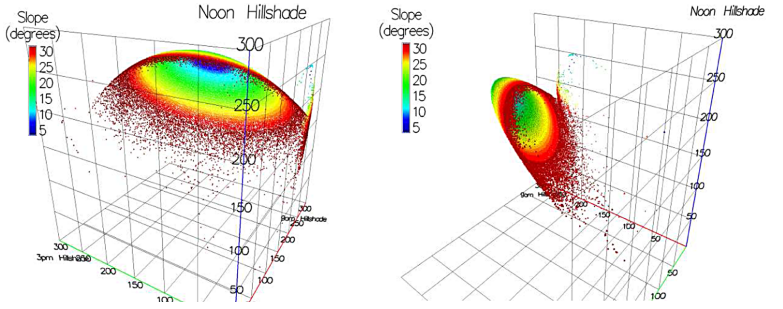
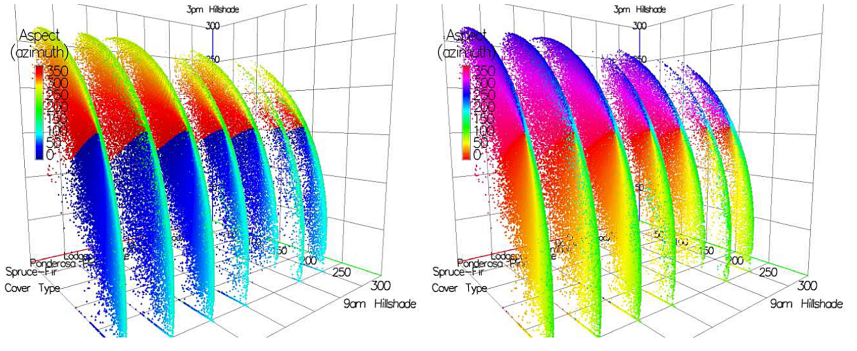


Fig. 10. Studying a selected situation using a different color scale



(a) Inappropriate color scale (b) Appropriate color scale

Fig. 11. Same figure shown using two different color scales

The scatter plots in Figure 10 gives us knowledge of the dataset, which in turn can be used in a partial correlation analysis to compute the partial correlation coefficient $\sigma_{78|9}$ (that is, σ_{78} given σ_9), which yields

$$\sigma_{78|9} = \frac{\sigma_{78} - \sigma_{79}\sigma_{89}}{\sqrt{(1-\sigma_{79}^2)(1-\sigma_{89}^2)}} = 0.94 \tag{2}$$

Thus, according to equation 2 there indeed exists a strong relationship between variables with index 7 and 8 (9am and Noon Hill-shade) - which is intuitively expected given their names, and seen in the plot. This clearly suggests that 3pm Hill-shade (index 9) should be taken into account.

Programmable Color Scales. Color is a very important object property, and until now we have used the "linear rainbow" color scale in two different versions - a discrete color scale (Figure 9) and a continuous scale (Figure 10). 3DVDM allows the user to define new color scales, or to select from a set of predefined scales - and also inversing, reversing or cycling them in order to enhance spatial structures.

The importance of using an appropriate color scale is illustrated. As an example *azimuth* (in degrees, from 0^0 to 360^0) is mapped to color, in Figure 11(a). Notice that the cathegorical variable Forest Cover Type is mapped to one of the axes, which results in seven planes along this axis.

Using any of the so far presented color scales is not feasible as 0^0 and 360^0 will be mapped to widely different colors, in spite of the fact that they describe the same direction. A "wrap-around" color scale as presented in Figure 11(b) solves this problem.

7 Visual Exploration of Dynamic Worlds

We consider two kinds of Dynamic Visualizations (DV), which one might distinguish between by using the terms "macro" and "micro" DV.

7.1 Macro Dynamic Visualization

In macro DV, data is sorted according to one of the variables of the database. The system then visualizes data from a "data window" which is sliding through the database. These kinds of animated visualizations therefore facilitate an alternative understanding of global trends in data, by using the time scale.

Figure 12 shows snapshots of a dynamic visualization made with the *macro* tool, provided with the 3DVDM system.

The same triplet as used in Figure 7(d) is used for this illustration, and elevation is mapped to the time axis. The sub-figures show not only how Forest Cover Type (mapped to color) changes as we "walk up the mountains", but also how Fire Points Distance, Hydrology Distance and Roadways Distance change according to altitude (all horizontal). The sub-figures cannot show the smooth, continuous behavior of the "snakes", which we see in VR and real-time animation in the "real" virtual world. However, we hope the point is made that potentially interesting substructures may be detected with this facility.

If we instead choose to map azimuth to the time axis, we can walk around the mountains, giving us an impression of how Forest Cover Type is influenced by this variable given the triplet. Also, some cover types may prefer very steep slopes; this may be revealed by mapping the slope variable to time.

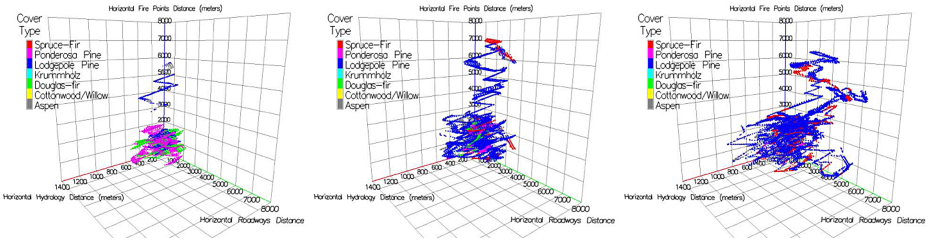


Fig. 12. Three snapshots of a macro dynamic visualization

7.2 Micro Dynamic Visualization

In micro DV, the visual objects in a data visualization may change visual properties and/or position according to a rule that may be unique to each visual object and controlled by its statistical variable. The visual objects can e.g. change their visual properties arbitrarily with respect to position, color and shape. This makes it theoretically possible to distinguish between visual objects by their dynamic behavior. It attracts attention when a subgroup of visual objects behave similarly. Microscopic DV can be used also for detecting clusters in datasets.

In an attempt to obtain a better understanding of vibrations, a "tail" was added to the vibrating glyphs. It was created by, at a given moment in time, simultaneously visualising all glyphs in their most recent position and in a number of previous positions. The number of previous positions determine the length of the tail. When sufficiently many positions are visualised, the entire shape formed by the movement becomes clearly visible. This converts moving patterns to static

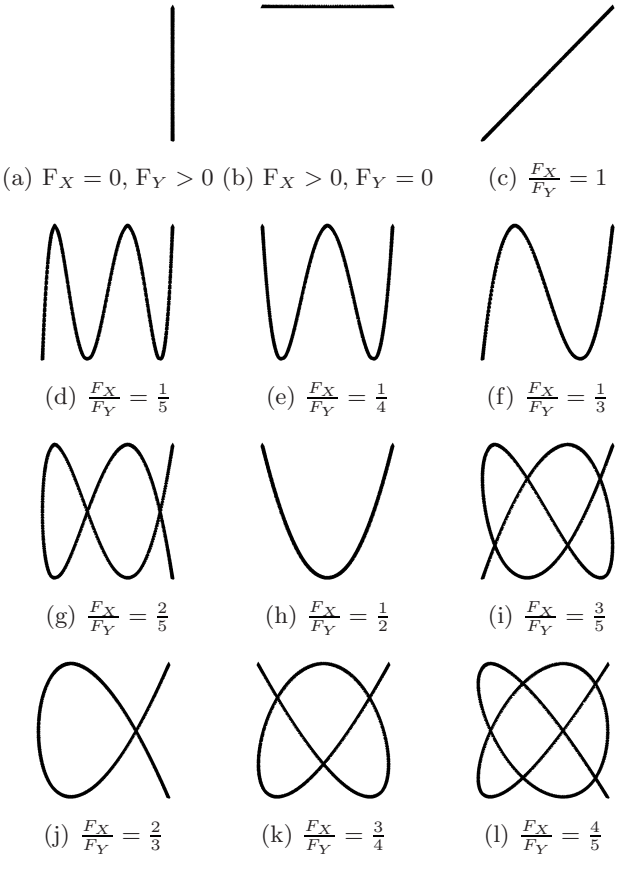


Fig. 13. Mapping categorical variables to the frequency of vibrating glyphs

patterns, and can thus not be said to be a defence for dynamic glyph attributes, but it does allow a more detailed study of vibrational modes.

In contrast to the previous visualisation methods, one must with this method only use categorical variables. Alternatively, one can round-off continuous variables to a few discrete steps. When mapping two variables to the frequency attributes in a 2D visualisation, the shape of the movement patterns depends upon the ratio between the two frequencies. This means that the ratio between frequencies is clearly displayed, but that the actual values cannot readily be deduced. In Figure 13, some of the basic movement patterns in two dimensions are shown.

These curves are well-known by mathematicians and are e.g. described in [22], [23] and [24]. The curves are usually called Lissajous curves after the French mathematician *Jules-Antoine Lissajous* (1822-1880) who discovered them in 1857, while studying wave patterns. However, it is also said that the American astronomer and mathematician *Nathaniel Bowditch* (1773-1838) discovered the curves already in 1815. The curves are therefore also sometimes called Bowditch curves. Plots that make use of these kinds of curves can therefore, perhaps, be called "Lissajous Plots".

8 Auditory Exploration of Static Worlds

In this section, three cases will be presented where the sound tools of 3DVDM are used. The first case will investigate usage of soundscapes in a situation where sound acts as a support for color, which represents Forest Cover Type. The second will investigate the same dataset but with sound used to represent Wilderness Area. In the third test we will attempt to map Vertical Hydrology Distance to sound.

For all cases the axes indicate Elevation, Horizontal Roadways Distance and Horizontal Hydrology Distance. This combination does not receive particular high score when we calculate the partial correlation coefficient. Still, interesting shapes appear as "tongues" stretching towards two of the axes' higher ends, Horizontal Hydrology Distance in particular. Figure 14 shows a 3D scatter plot

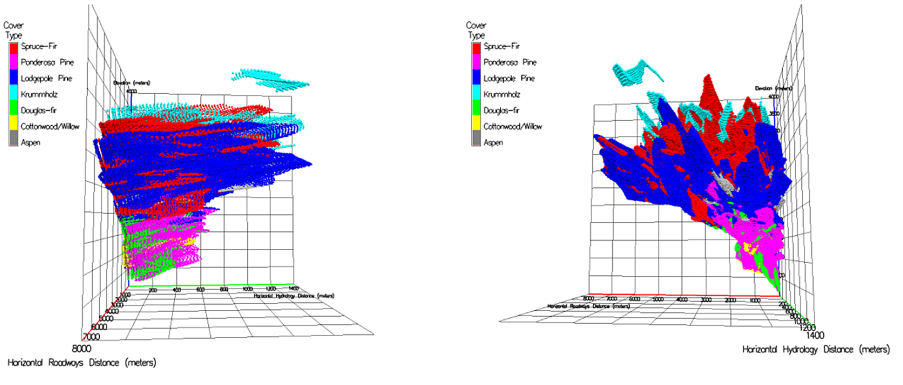


Fig. 14. 3D scatter plot from different angels

of the dataset used for these examples. The two pictures show the same scatter plot from different angles.

As it is difficult to present a soundscape in time in a document of this type, we will try to make a description of the soundscapes that occur.

During the tests the distance threshold is adjusted to investigate how this function affects the soundscape and the listeners perception of the content. We also try different rendering methods: Rendering the whole set and rendering a sweep along the three axes. When applicable we bring out the virtual torch to test this on areas on high interest.

8.1 Sound Supporting Color

The main objective in this test is to investigate how the use of dynamic 3D soundscapes works as support for a visual parameter, in this case: color. We will also investigate if it is possible to navigate the soundscape, so that we can locate areas of high concentration. Finally we will will investigate the possibility of locating interesting areas that will not be visible to the eye.

The database is sampled each 10ms, and 16 entries are randomly picked and mapped to sound samples of spoken numbers 1 to 7 representing the seven Cover Types. Initially all thresholds are set to maximum values, so that all observations are potential sources. The listener is placed in the middle of the coordinate system, and starts navigating the soundscape from there. It is allowed to adjust distance threshold.

- The immediate overall impression is a soundscape consisting of the numbers 5 and 7, which are the two dominant types of cover type: Lodgepole Pine and Spruce Fir. It is difficult to hear other types.

Distance threshold is lowered to 10 (graphical) units[2]. This allows the listener to investigate close range areas further by navigating through the data.

- It doesn't reveal anything straight away but closing in on the area around the middle of the elevation axis increases the number of Aspen (Type 1) to a noticeable level. It reveals what can be seen from the color: that the number of Aspen are few compared to Lodgepole Pine and Spruce Fir at that Elevation point (and they are close to roads in general).

Next, the Elevation Axis is rendered in time. Data sampling is done using a sliding window (Figure 15) and distance threshold reset to maximum. The objective is then to navigate around in the soundscape and listen for interesting things.

- This gives the impression of gradually changing Cover Type as the Elevation increases. After several passes it also reveals that there still are types that we cannot see in the scatter plot when Lodgepole Pine and Spruce Fir become visually dominant (primarily Aspen).

[2] For reference: The coordinate system is $100 \times 100 \times 100$ graphical units.

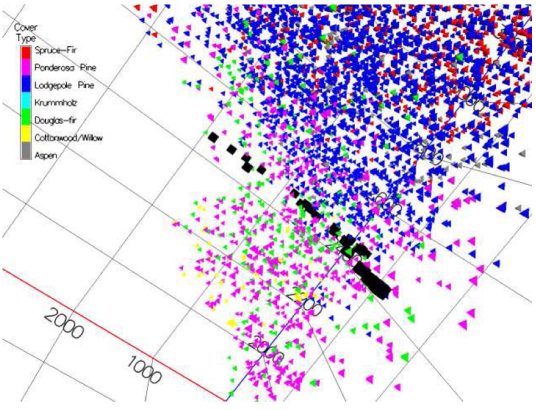


Fig. 15. Sampling along the Elevation axis. Black cubes mark the current samples.

Rendering the other axes does not reveal anything that can not be seen from the colors. Using the torch to investigate statistical values does not really make sense in this case where sound acts as support for color.

8.2 Sound "On Its Own" – Categorical Variables

The main objective of this test is to investigate how the use of dynamic 3D soundscapes works for data mining when there is no visual reference, and the data variable that is used for rendering the soundscape consists of a few (four) categorical values. We will see if it is possible to get an idea of how the selected variable is distributed. We will also investigate if it is possible to navigate though the soundscape, so that we can locate areas of interest.

The database is sampled each 10ms, and 16 entries are randomly picked and mapped to sound samples of spoken numbers 1 to 4 representing the four wilderness areas. Color still shows Cover Type. Initially all thresholds are set to maximum values, so that all observations are potential sources. The listener is placed in the middle of the coordinate system, and starts navigating the soundscape from there. It is allowed to adjust distance threshold. Doing a bit of "cheating" by changing color to represent Wilderness Area shows the layout in Figure 16.

– The initial experience when sampling the whole set randomly is an overweight of area 2 and 4. This is expected from the layout of the database.

We choose to render the three axes one by one while navigating the dataset to get a picture of the Wilderness Area distribution.

– It becomes clear that the large tongue stretching out the Horizontal Roadways Distance axis is area 4. About half the maximum distance area 2 gradually increases. The slim tongues that reach out the Horizontal Hydrology Distance axis are primarily area 2 with some representation of area 4. Data

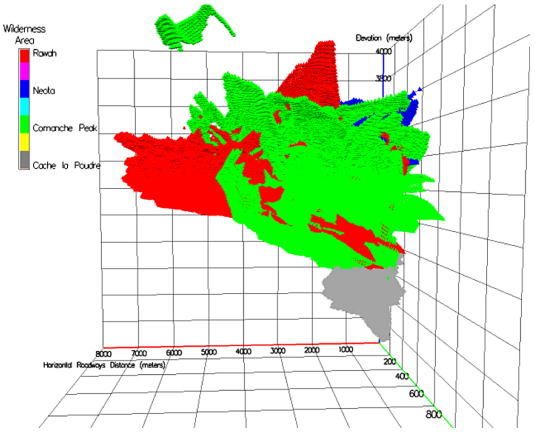


Fig. 16. A "sneak peak" at Wilderness Area distribution using color

from area 3 seems to be located at high Elevation with low Roadway Distance and area 1 is located at low Elevation and low Hydrology and Roadway Distances.

We then choose to try identifying what kind of Forest Cover that is in different Wilderness areas. By looking at the colors one can see that Cottonwood/Willow, Douglas fir and Ponderosa Pine are grouped in one corner of the scatter plot. Distance threshold is set to 10 and the area is investigated further (by navigating into this area).

- This reveals that all these Cover types are in area 1 until the elevation reaches a certain level.
- This area also has a few Lodgepole Pines.

It now seems feasible to bring out the torch and point it where area 1 seems to stop (Figure 17) to see how these Cover Types are distributed in the Wilderness Areas.

- Closer investigation with the torch reveals that area 1 has a few observations around Elevation 2600 where it seems to stop. Areas 2 and 4 take over and have a few Douglas-fir and Ponderosa Pine but no Cottonwood/Willow.
- Moving a bit upwards along the Elevation axis with the torch aimed at Aspen confirms that this only exist in area 4.

8.3 Sound "On Its Own" – Continuous Variables

The main objective of this test is to investigate how the use of dynamic 3D soundscapes works for data mining when there is no visual reference, and the data variable that is selected for rendering the soundscape consists of many

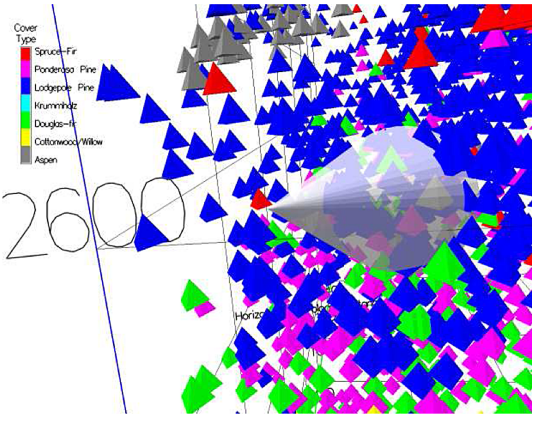


Fig. 17. Investigating area 1, maximum elevation area

different observations. We will see if it is possible to get an idea of how the selected variable is distributed. We will also investigate if it is possible to navigate though the soundscape, so that we can locate areas of interest.

The database is sampled each 10ms, and 16 entries are randomly picked and mapped to synthesized waveforms spaced on a Ionic scale over two octaves. The values in the chosen scale (Vertical Hydrology Distance) is normalized and mapped to this scale so that low values become low pitch and visa versa. Color still shows Cover Type. Initially all thresholds are set to maximum values so that all observations are potential sources. The listener is placed in the middle of the coordinate system starts navigating the soundscape from there. It is allowed to adjust distance threshold.

– The immediate overall impression is a soundscape with many different values but with high concentration of mid range values and very few in the ultimate high and low region.

Distance threshold is again lowered to 10 units allowing closer inspection of local areas by navigation.

– There is a noticeable change in the soundscape along the Horizontal Hydrology Distance axis. At low distance the values seem concentrated on a value in the middle of the lower octave (i.e. around 1/4th of the maximum Vertical Distance, i.e. around 0 meters, since this data variable has both negative and positive values). There doesn't seem to be any very low or high values.
– Moving in the direction of high Horizontal Hydrology Distance creates a more distributed soundscape with many notes of different pitch. Sounds seem concentrated around middle values with a larger spread than initially, but there are definitely some very low and high values in this area.

It seems feasible to try to render along the three axes, especially Horizontal Hydrology Distance should be interesting.

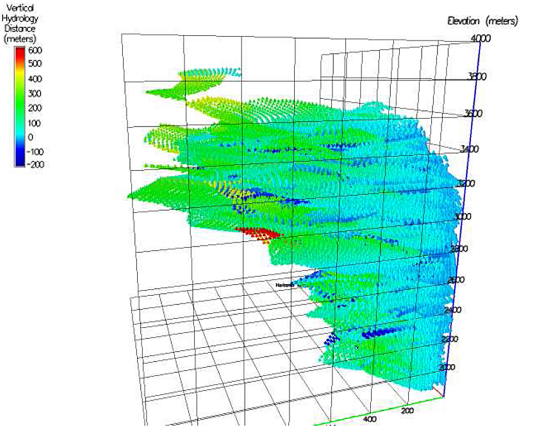


Fig. 18. Mapping Vertical Hydrology Distance to color

- This confirms what was indicated rendering the whole scatter plot. There is a strong representation of a level about 1/4th as mentioned above
- This spreads out as Horizontal Hydrology Distance increases.
- When Horizontal Hydrology Distance is 0 Vertical Hydrology Distance is also 0 provided that the pitch value we hear as about 1/4th of the total tonal range is equal to the value of 0. (This correlation between the two distances would be expected for e.g. trees close to lakes and rivers).

Mapping Vertical Hydrology Distance to color (Figure 18) gives a clearer view of how the distribution of this value changes as Horizontal Hydrology Distance increases. The red tongue was not detected during the test, but was only audible as a few high pitched sounds which were difficult to locate.

9 Discussion

9.1 Visual Data Exploration

We aimed at demonstrating the potential benefit of immersive VR for Visual Data Mining, using a new version of our 3DVDM system and a new version of the VR++ framework upon which it is based. The system is capable of providing real-time user response and navigation as well as showing dynamic visualizations of large amounts of data.

Although we find the real-time performance of the system very adequate, we have not included any tests to support this claim. Instead, we have put emphasis on illustrative support of our hypothesis that a complementary and valuable benefit of using the system's visualization tools is achievable concerning the detection of e.g. non-linear relationships and substructures in data, which are most likely to escape traditional methods of data analysis.

An investigation of a data set starts with simple uni- and bi-variate analyses to become familiar with data and possibly reduce the number of variables to a reasonable number. In the case study presented we had 10 quantitative variables, and they were all forwarded to an evaluation of which combinations of three could be most interesting for defining the three spatial axes in the "Extended Scatter Plots", where more detailed visual inspection could take place.

A "Scatter Plot Tour" presented systematically the 120 unique "triplets" of the 10 variables with the dependent variable (Forest Cover Type) mapped as color of the data points (visualized as small objects). The criteria for "interesting" is entirely based on the user's subjective evaluations when observing data in 3D space. This is potentially a weakness, but the point is to complement the algorithmic methods with the power of human perception.

Several interesting triplets were observed and a few were illustrated and selected for further investigation. One triplet was used to demonstrate that the objects might encode multiple visual properties representing more statistical variables simultaneously, while another demonstrated the use of flexible and controllable color scales. Eventually the Macro Dynamic visualization was demonstrated and revealed most surprising substructures in the data.

The full benefit of the 3DVDM tools assumes 3D VR visualization systems like a CAVE or a Panorama, where the user can navigate around and pursue intriguing views. Such benefit can only be experienced "in situ", and for this paper we are left with the means of mono illustrations as presented on a monitor.

However, we hope that the major points of the approach and its potential do come through via the many illustrations provided and their organization as a successively progressing use of the visualization tools. Several peculiar data structures were detected through the visual inspection, and they could possibly warrant a more detailed statistical analysis to explain the phenomena as valuable information or otherwise.

The visualizations presented are all selected on the basis of their perceptual particularities, without any claims whatsoever about practical and/or statistical significance.

9.2 Auditory Data Exploration

The tools and methods presented in this chapter are only a few of many possibilities of this system. There are virtually unlimited ways to construct a soundscape using different sounds and settings.

In general, sound seems to support a visual parameter like color quite well. It doesn't reveal much new information; but can be useful when some data structure occludes other interesting observations. If certain clusters are out of visual range it will create sound cluster in that direction, provided that there are sufficient data in that cluster.

Mapping statistical values to a few different sounds works when the data values are a few categorical values. If the data values are continuous it is more difficult to suggest a soundscape with the same informational value, though using the synthesized sounds may still yield some information about relative values.

Sampling of the whole 3D scatter plot randomly and especially along the three axes in time gives a soundscape that is similar to the one we will get by choosing color rather than sound. The real strength of this method appears when comparing color and sound information to investigate the correlation between these. The torch is useful for finding the direct statistical value of a given observation as long as the statistical values are few and preferably categorical.

An important parameter to consider is the distance threshold that enables the listener to concentrate on the local area, because this also seems to eliminate most of the potential background noise created from distant objects. However, this also eliminates the ability to navigate towards distant areas on basis of the auditory cue from these. This was especially true for the last test where the synthesizer was used. In the second test it was possible to work with a higher distance threshold, probably because of the few different sounds.

When the user becomes familiar with the current properties of a 3D soundscape it may become possible to navigate supported by sound cues, given that there are clusters that provides sufficient positional cues. In cases where there are no apparent clusters or other patterns in the soundscape it may just confuse the user. In this case it is probably a better solution not to use the sound tools.

In any case, using soundscapes does hold enough information to give a strong indication of the distribution of a given statistical value. This information is also significant enough to trigger a closer investigation in most cases.

10 Conclusions

Concerning the potential benefit of immersive VR for VDM, our hypothesis was that a complementary and valuable benefit is achievable concerning the detection of e.g. non-linear relationships in data, which are most likely to escape traditional methods of data analysis.

We have not presented tests that verify explicitly the benefit of navigation and real-time user response in immersive VR system, but we have illustrated the usefulness of the 3DVDM framework designed for VR through a series of examples. The VDM tools do help in discovering remarkable non-linear data relations and substructures in the dataset used, which it would have been very difficult or impossible to detect using more traditional methods of analysis. In particular the Macro Dynamic Visualization revealed unexpected substructures.

Commenting on the actual practical and statistical significance of the discovered data structures is beyond the scope of the chapter. No statistical nor conclusive analysis is aimed at with the system. The output is specification of phenomena, that may warrant follow-up of proper statistical analysis.

Concerning the use of sound for data exploration, this project intended creating software tools that allowed us to use sound to assist us in performing visual data mining in VR. This chapter presented two basic sound tools developed for this purpose, and our aim was to present and test these tools, in various ways.

Tests have shown that it is possible to use sound for data mining in VR as either a support for visual parameters or as a stand-alone method, and especially using different exchangeable sample banks to represent statistical values proves to be a useful way of locating data values in VR. The success will depend on which type of data we wish to investigate, since keeping a simple soundscape is crucial for precise perception of value. Still it is possible to encode some kind of information about level for data that is more complicated.

Future tests should try to investigate the threshold of complexity for soundscapes that are useful for data mining (i.e. holds some kind of numerical value). This is important in order to avoid listener fatigue and information overload which was a common problem during this work.

Acknowledgments

We gratefully acknowledge the support to the 3DVDM project from the Danish Research Councils, grant no. 9900103. We also acknowledge Jock A. Blackard and Colorado State University with thanks for making the Forest Cover database available.

References

1. Asimov, D.: The grand tour: a tool for viewing multidimensional data. SIAM J. Sci. Stat. Comput. 6(1), 128–143 (1985)
2. Inselberg, A., Dimsdale, B.: Parallel coordinates: a tool for visualizing multidimensional geometry. In: VIS 1990: Proceedings of the 1st conference on Visualization 1990, pp. 361–378. IEEE Computer Society Press, Los Alamitos (1990)
3. Nelson, L., Cook, D., Cruz-Neira, C.: Xgobi vs the c2: Results of an experiment comparing data visualization in a 3-d immersive virtual reality environment with a 2-d workstation display. Computational Statistics: Special Issue on Interactive Graphical Data Analysis 14, 39–51 (1999)
4. Nagel, H.R., Granum, E., Musaeus, P.: Methods for visual mining of data in virtual reality. In: Proceedings of the International Workshop on Visual Data Mining, in conjunction with ECML/PKDD2001, Freiburg, Germany, 2nd European Conference on Machine Learning and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 2001, pp. 13–28 (2001)
5. Nagel, H.R.: Exploratory Visual Data Mining in Spatio-Temporal Virtual Reality. PhD dissertation, Faculty of Engineering and Science. Aalborg University. Denmark (2005)
6. Granum, E., Musaeus, P.: Constructing virtual environments for visual explorers. In: Quotrup, L. (ed.) Virtual Space: The Spatiality of Virtual Inhabited 3D Worlds, Springer, Heidelberg (2002)
7. Symanzik, J., Cook, D., Kohlmeyer, B.D., Lechner, U., Cruz-Neira, C.: Dynamic statistical graphics in the c2 virtual environment. In: Second World Conference of the International Association for Statistical Computing, Pasadena, California, USA, February 1997, vol. 29, pp. 35–40 (1997)
8. Wegman, E.J., Symanzik, J.: Immersive projection technology for visual data mining. Journal of Computational and Graphical Statistics 11(1), 163–188 (2002)

9. Carr, D.B., Nicholson, W.L.: Evaluation of graphical techniques for data in dimensions 3 to 5: Scatterplot matrix, glyph, and stereo examples. In: Proceedings of the Section on Statistical Computing, Alexandria, VA, American Statistical Association, pp. 229–235 (1985)
10. Pickett, R.M., Grinstein, G.: Iconographic displays for visualizing multidimensional data. In: Proceedings of the IEEE Conference on Systems, Beijing and Shenyang, People's Republic of China, Man and Cybernetics, pp. 514–519 (1988)
11. Ribarsky, W., Ayers, E., Eble, J., Mukherjea, S.: Glyphmaker: Creating customized visualizations of complex data, pp. 57–64. IEEE Computer, Los Alamitos (July 1994)
12. Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association 68(342), 361–368 (1973)
13. Healey, C.G., Enns, J.T.: Large datasets at a glance: Combining textures and colors in scientific visualization. IEEE Transactions on Visualization and Computer Graphics 5(2), 145–167 (1999)
14. Ebert, D.S., Rohrer, R.M., Shaw, C.D., Panda, P., Kukla, J.M., Roberts, D.A.: Procedural shape generation for multi-dimensional data visualization. In: Gröller, E., Löffelmann, H., Ribarsky, W. (eds.) Data Visualization 1999, pp. 3–12. Springer, Wien (1999)
15. Brown, M.L., Newsome, S.L., Glinert, E.P.: An experiment into the use of auditory cues to reduce visual workload. In: CHI 1989 Proceedings, New York, USA (1989)
16. Mutanen, J.: Perception of sound source distance, Nokia Research Center (2003)
17. Baluert, J.: Spatial hearing: the psychophysics of human sound localization. The MIT Press, Cambridge (1997)
18. Nagel, H.R., Granum, E.: Vr++ and its application for interactive and dynamic visualization of data in virtual reality. In: Proceedings of the Eleventh Danish Conference on Pattern Recognition and Image Analysis, Copenhagen, Denmark (August 2002)
19. Nagel, H.R., Granum, E.: A software system for temporal data visualization in virtual reality. In: Proceedings of the Workshop on Data Visualization for large data sets and Data Mining, Augsburg, Germany, Department of Computer Oriented Statistics and Data Analysis University of Augsburg (October 2002)
20. Zahorik, P.: Auditory display of sound source distance. In: Proceedings of the 2002 International Conference on Auditory Displays, Kyoto, Japan (July 2002)
21. Blackard, J.A.: Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types. PhD dissertation, Department of Forest Sciences. Colorado State University. Fort Collins, Colorado (1998)
22. Lawrence, J.D.: A Catalog of Special Plane Curves. Dover, New York (1972)
23. Cundy, H., Rollett, A.: Mathematical Models, 3rd edn. Tarquin Pub., Stradbroke (1989)
24. Gray, A.: Modern Differential Geometry of Curves and Surfaces with Mathematica, 2nd edn. CRC Press, Boca Raton (1997)