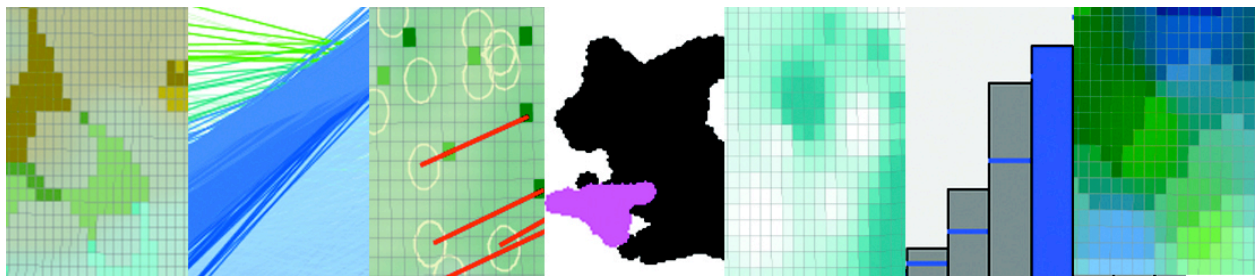


Spring semester 2009
Geomatics Engineering and Planning BSc

Bachelor Thesis

Geovisualization Tools Applied to Landscape Genetics



Author

Elisabeth Leu
Berninastr. 28
8057 Zürich

elisabeth.leu@epfl.ch
eleu@student.ethz.ch

Supervision

Prof. Dr. François Golay

Prof. Dr. Lorenz Hurni

Dr. Stéphane Joost

Dr. Christian Häberling

Laboratoire SIG
EPFL

Institute of Cartography
ETHZ

Motivation

During my exchange semester at EPFL, I seized the chance to get to know not only a new culture, but also a new domain of geographical data treatment. Coming from the strict approach of confirmatory data analysis, it was in the beginning a challenge to get another view and find the differences between confirmatory and exploratory data analysis. Nevertheless, treating an unknown and still developing topic motivated me to go on. On the whole, this thesis was for me an adventure, not only on the side of the topic, but also on the one of the language and the coordination between Lausanne and Zurich. And it was well worth it!

Acknowledgements

In general, I am glad for the support, motivation and feedback of all the people around me, especially the students at EPFL as well as ETHZ.

I am very grateful to Prof. Dr. F. Golay and Prof. Dr. L. Hurni for their consent and the possibility of this thesis, especially because it needed some extra effort to find a topic.

Thanks go also to Dr. Christian Häberling of the IKA in Zurich for his critical view on the topic, the advice regarding the structure of the thesis and his flexibility in the coordination of the work.

A special thank for his support goes to Dr. Stéphane Joost of the LaSIG in Lausanne. He supervised my work and gave me a lot of information, advice and ideas for the topic and answered my numerous questions.

Abstract

This work deals with the approach of geovisualization in the context of the analysis of goat breeds genetic data and environmental information in Switzerland. Geovisualization is an exploratory approach to data analysis, derived from exploratory data analysis and adapted to a spatial context. The main characteristics of this method are interactivity, dynamic visualization and combination of different statistical tools and graphs implemented in several software available on the Internet, like CommonGIS (<http://www.iais.fraunhofer.de/1863.html>), GeoDA (<http://geodacenter.asu.edu>) and GeoViz Toolkit (<http://www.geovista.psu.edu/geoviztoolkit/index.html>). In this research, the latter was used to apply geovisualization to landscape genetics, in order to find relationships between space and / or the different variables. During this process, some tools as RadViz and Parallel Coordinates Plot (PCP) showed their usefulness since they respect the multidimensional aspect of the data set. This approach led to some hypotheses that might be tested in any further confirmatory analysis, e.g. with a standard statistical software and a GIS.

Zusammenfassung

Dieser Bericht behandelt die Geovisualisation angewendet auf genetische Ziegenrassen-Daten in Kombination mit Umweltinformation in der Schweiz. Geovisualisation, abgeleitet von explorativer Datenanalyse, ist ein entdeckender Ansatz Datensätze zu behandeln. Diverse Programme, wie CommonGIS (<http://www.iais.fraunhofer.de/1863.html>), GeoDa (<http://geodacenter.asu.edu>) und GeoViz Toolkit (<http://www.geovista.psu.edu/geoviztoolkit/index.html>), weisen die Hauptcharakteristiken der Geovisualisation auf: Nutzer-Rechner-Interaktivität, Visualisation und Kombination verschiedener Hilfsmittel. GeoViz Toolkit wird in der Folge verwendet, um Geovisualisation auf Landschaftsgenetik anzuwenden, um so räumliche oder inhaltliche Zusammenhänge zu finden. Während dieses Vorgangs zeigten sich vor allem die Hilfsmittel RadViz und Parallell Coordinates Plot (PCP) nützlich, da sie die Mehrdimensionalität des Datensatzes berücksichtigen. Aus der Analyse konnten am Schluss einige Hypothesen über mögliche Zusammenhänge aufgestellt werden, welche in einem klassischen statistischen Test (z.B. GIS-Analyse) geprüft werden könnten.

Résumé

Ce rapport traite de la géovisualisation appliquée à des données génétiques de races de chèvres dont l'analyse a été effectuée en relation avec des variables environnementales en Suisse. La géovisualisation est une approche exploratoire d'analyse des données, dérivée de l'analyse exploratoire mais dans un contexte spatial. Elle est définie par l'interactivité, une visualisation dynamique et une aptitude à combiner des outils statistiques divers, ainsi que plusieurs modes de représentation de l'information. Sur le Web, plusieurs logiciels permettent de faire de la géovisualisation. Parmi eux, CommonGIS (<http://www.iais.fraunhofer.de/1863.html>), GeoDA (<http://geodacenter.asu.edu>) et GeoViz Toolkit (<http://www.geovista.psu.edu/geoviztoolkit/index.html>) ont été examinés plus en détail. C'est finalement GeoViz Toolkit qui a été choisi pour l'application. L'analyse des données de base permet de trouver des relations entre variables dans l'espace géographique. Lors de cette analyse, les outils respectant l'aspect multidimensionnel des données se sont montrés les plus convaincants. Après une démarche exploratoire effectuée avec l'aide du logiciel choisi, des hypothèses de travail peuvent être formulés afin de commencer une analyse de type confirmatoire.

Table of content

Motivation	3
Acknowledgements	3
Abstract	4
Zusammenfassung	4
Résumé	4
List of figures	6
List of tables	6
1 Introduction	7
Goal	7
Overview	7
2 Theory	8
Data analysis	8
Confirmatory Data Analysis	8
Data Mining	8
Exploratory data analysis (EDA)	8
Exploratory spatial data analysis (ESDA)	9
Geovisualization	9
Overview	11
3 Software Overview	12
Available Software	12
CommonGIS	14
GeoDa	15
GeoViz Toolkit	16
Selection	17
4 Application to Landscape Genetics	18
Data Sets	18
Main Questions	20
Possible Analysis Process	20
Derived Hypothesis	28
5 Conclusion	29
6 References	30
7 Appendix	31
Variables of SGGRID-CH.shp	31

List of figures

Fig. 2.1	Different aspects of cartographic communication and geographic visualization. After McEachren (1995, p.357 / Fig PIII.1.)	10
Fig. 2.2	Data Analysis and its related fields	11
Fig. 4.1	Choroplete map of genetic diversity per municipality (left) and per 4 km ² cell (right). High diversity is shown with a dark hue, and low diversity with a light one.	19
Fig. 4.2	Genetic diversity compared to DTRyearlym (top left), FRSyearlym (top right), REHyearlym (bottom left) and TMPyearlym (bottom right). See the appendix to find the description of the variables.	21
Fig. 4.3	An observation attached by different springs to the variables. (From Nováková et. al., 2006, p. 471)	22
Fig. 4.4	GeoMap (left) and RadViz (right) showing selected cells with a low genetic diversity.	22
Fig. 4.5	The region of Poschiavo shows interesting correlations in RadViz.	23
Fig. 4.6	The last frame of the animation shows the highest genetic diversity values.	23
Fig. 4.7	Selecting mean genetic diversity values in the histogram doesn't permit to highlight correlations on the map.	24
Fig. 4.8	A simple parallel coordinate plot of a car. (After Robinson, 2005)	24
Fig. 4.9	A parallel coordinate plot containing all genetic and environmental variables. (The figure is actually a combination of two plots.)	25
Fig. 4.10	Parallel coordinate plot with a selection in upper-middle part of the distribution of SUNyearlym values.	26
Fig. 4.11	Parallel coordinate plot with geographic units showing low genetic diversity.	26
Fig. 4.12	An example of arrangement combining the different tools.	27

List of tables

Table 3.1a	General overview of some geovisualization software (Part 1)	12
Table 3.1b	General overview of some geovisualization software (Part 2)	13
Table 3.2	Evaluation grid of geovisualization software	17

1 Introduction

The geographical analysis of genetic data is a relatively recent field. To study the spatial distribution of the genetic diversity in Swiss goat breeds, data can be analysed directly within GIS software (thematic maps on genetic diversity variables), but with no opportunity to discover yet unknown potential relationships between the genetic information and the environmental variables used to characterize the locations where it was collected (DNA sampling). On this basis, other approaches have to be examined to discover these possible correlations.

Goal

Geovisualization is an important tool to explore spatial data sets. This new approach in data analysis is yet rarely applied to specific fields of research, but could have an important potential to analyse huge sets of data. The goal of this bachelor thesis is to show the capacities and limits of exploratory spatial data analysis and geovisualization applied to genetic diversity in Swiss goat breeds and to environmental information characterizing places where these animals are living.

Overview

The thesis is divided in three main parts and a final conclusion.

In the first part, the theory and origins of geovisualization is presented (why was this approach developed, what are the links to GIS / data mining, and what are its main characteristics).

The second part provides an overview of available software related to geovisualization, and a closer look on three of them: CommonGIS, GeoDA and GeoViz Toolkit. It is examined up to what degree these software satisfy the characteristics of geovisualization presented in the previous chapter. Finally, the results are synthesized in a weighted table, leading to the selection of the software used in the next chapter.

The third part contains an application of geovisualization to Swiss goat breeds genetic data and environmental information describing the places where the animals were sampled. The available data sets are used to show how possible working hypotheses can be drawn from a geo-exploratory process. As a result, these hypotheses are formulated and could be used in the context of a subsequent confirmatory analysis process.

In the conclusion, the advantages and limits of geovisualization in this context are outlined, and an outlook for the development, as well as an evaluation of this approach, are given.

2 Theory

In many fields, the amount of data produced increased a lot over the last few years. New approaches were developed to deal with those large data sets, among which geovisualization based on exploratory spatial data analysis (ESDA). The goal of this section is to present this GIScience approach in the context of data analysis.

Data analysis

A set of data is useless in case there is no mean to extract information and knowledge from it. Often information is not directly accessible, and most of it is hidden and requires to be discovered. When the size of a data set becomes very large, the intuitive way of analysis is not possible anymore and specific data analysis approaches are needed. The latter often have to be applied iteratively, to detect patterns or relationships between features.

Data analysis can be divided into three main approaches: confirmatory data analysis, data mining and exploratory data analysis.

Confirmatory Data Analysis

This approach is based on a hypothesis that needs to be tested. The hypothesis involves an assumption about a possible process or way to test the hypothesis. Applied to spatial data, it is often used in connection with standard GIS-analysis and tools (ArcGIS, MapInfo, etc.). In this context, GIS are used to represent the results of the statistical analysis (e.g. choroplethe maps).

Data Mining

Data mining can be presented as a second approach to data analysis. Joost (2007) refers to it as “a nontrivial process allowing to identify valid, novel, potentially useful, and understandable patterns in very large data sets”. It is an automatic way to analyse a structurally known data set. Results are shown as text, tables or charts.

Exploratory data analysis (EDA)

Introduced by the mathematician John Tukey in 1963, exploratory data analysis became increasingly relevant during the last ~10 years. The occurrence of huge and structurally unknown data sets made it necessary to apply other ways to analyse data than usual descriptive statistics and classical hypothesis based methods. The important quantity of information and often its mixing don't allow formulating a hypothesis in advance, since no similar data set were yet explored.

Andrienko and Andrienko (2006) describe EDA as getting acquainted with the data and exploring it in order to see what it might tell us. A hypothesis doesn't exist yet, but could be generated with EDA and subsequently tested with the confirmatory data analysis. This approach is still question-driven, but the questions differ from hypothesis in a way that they are more generally formulated and do not imply an analysis process. For example the initial question to start an EDA could be as follows:

- Why do some goat breeds in Switzerland become endangered?
- What can be done to improve traffic security in Lausanne?
- How could pollution be reduced in the canton of Vaud?

EDA, compared to other data analysis, involves many more questions, whose level of generality varies a lot. In addition, most of the questions evolve during the process, unlike

confirmatory data analysis in the context of which hypotheses are decided from the beginning. After Andrienko and Andrienko, this approach in data analysis is more a philosophy than a specific set of techniques, with a strong accent on exploration and hypotheses generation. Still, a general three-step process could be defined according to the “Information Seeking Mantra” (Shneidermann 1996): First to get an overview over the available data and the structure, secondly filtering and zooming could be applied, and at last to have a special look on details.

In contrast to data mining, which is in general automatic, those processes require high user interactivity in a sense that the actual analysis is done by the human brain. But data mining and EDA contribute to the same purpose and could be well combined (Demšar, 2006). Some authors refer to EDA as visual data mining (Demšar, 2006), and the classical data mining as automatic data mining. Here we call it EDA, knowing that it is often applied in combination with automatic data mining.

EDA later evolved towards ESDA and subsequently to geovisualization.

Exploratory spatial data analysis (ESDA)

If a data set contains spatial information, the general EDA approach can be adapted to the particular characteristics of spatial data. Derived from EDA, it can be divided into three aspects (according to Andrienko and Andrienko, 2006):

- The data containing characteristic components like measurements and referential information, and the context in which those measurements were taken.
- Tasks as parts of the whole process, to analyse behaviour or patterns.
- Tools to achieve the necessary tasks.

The most important tool in ESDA is visualization, because it provides a powerful possibility to explore the data set. Through visualization, the user can detect patterns and features, which would remain undiscovered through automatic and standard processes.

Other tools are display manipulation, data manipulation, querying and computation. Those tools involve a high degree of user interactivity and dynamic display, especially when time is involved as well as spatial references.

Geovisualization

As a further development and extension of ESDA, geovisualization (abbr. for geographic visualization) also involves high interactivity in the exploratory approach, but the emphasis is on the visualization aspects. Visualization in this context means the display of data in a geographic reference, but also mental images produced while exploring the data.

McEachren (1995) differs geovisualization from cartographic communication by their different purpose as shown in figure 2.1 (next page).

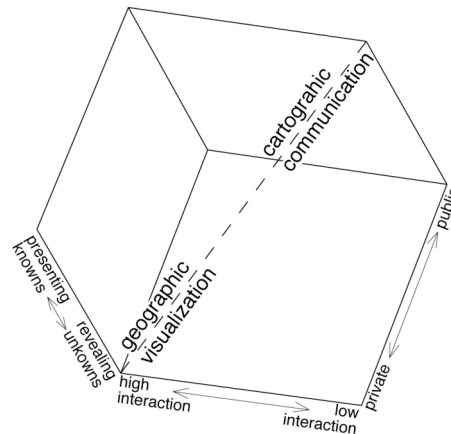


Fig. 2.1 Different aspects of cartographic communication and geographic visualization.
After McEachren (1995, p.357 / Fig PIII.1.)

Cartographic communication focuses on the representation of mostly static, known information with a readable amount of elements to a certain audience. In contrast, geovisualization deals with unknown information to be shown only to the analyst by ways of high interactivity between the user and the computer. The two purposes are not strictly separable; data can be explored in a published map as well, but with a much lower complexity than with geovisualization.

The basic element of geovisualization is not just the produced image, but also how we perceive those images and how our minds deal with this information, in order to find interesting information in the data set. Those processes are described in the *Gestalt* principles, which involve the laws of similarity, closure, symmetry, common fate, etc. Those need to be kept in mind as basis of visual interpretation.

Slocum et al. (2005) collected a set of methods, which are most common in geographic data exploration / visualization:

- **Manipulating data** means that the data itself is changed, e.g. standardizing, transforming or classifying.
- **Varying symbolisation** involves different types of symbols and their specific attributes, e.g. different colour schemes of choroplete maps. The way that symbols are assigned to attributes is another aspect of symbolization.
- **Manipulating the user's viewpoint**, derived from the fact that the large amount of data needs to be displayed on a relatively small screen, while the data and symbolization aren't changed. E.g. zoom and pan.
- **Highlighting portions of a data set** is a fundamental method, which can be seen as classification of data. In this case, the data remain unchanged. E.g. focusing and brushing.
- **Multiple views** allow us to compare data or different symbolization visually, e.g. by small adjoined maps.
- **Animation** gives an additional dimension for data exploration.
- **Linking maps with other forms of display**, such as tables, graphs and numerical summaries, is an important method to combine other visual data analysis to geovisualization (brushing).
- **Access to miscellaneous resources** is needed when additional data should be involved after some time of analysis.
- **Automatic map interpretation** provides a link to data mining, in order to assist the interpretation of the given data.

Overview

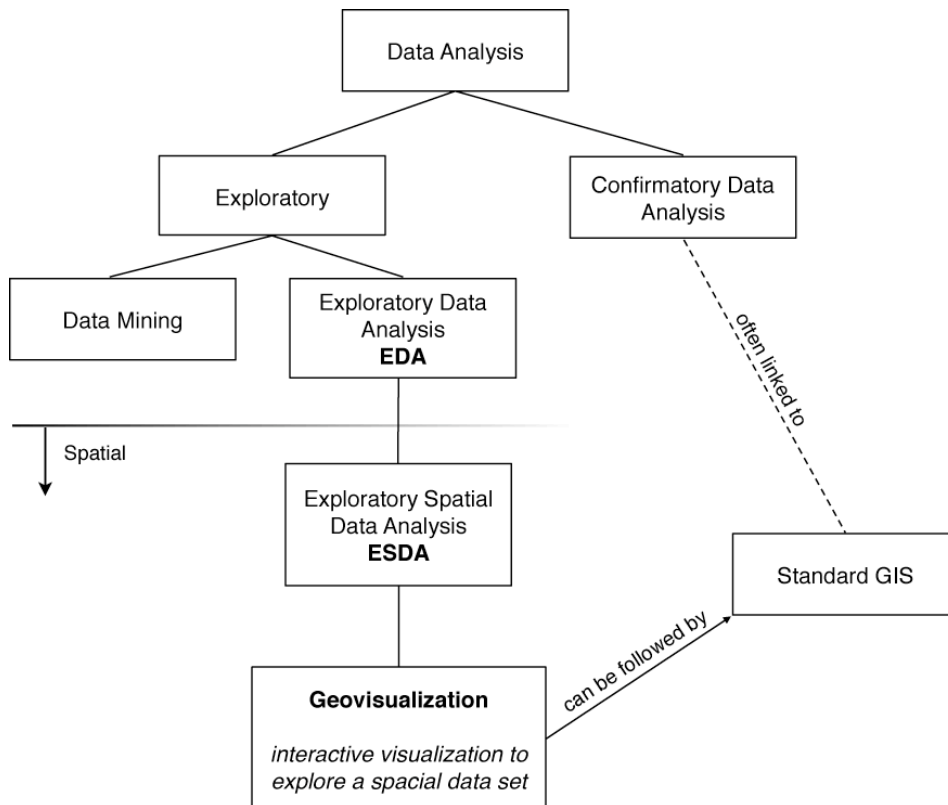


Fig. 2.2 Data Analysis and its related fields

To sum up, data analysis as a part of statistics contains three main approaches: data mining as an automatic process, exploratory data analysis as a way to interactively analyse unknown data sets, and confirmatory data analysis driven by hypothesis. Those approaches all involve and combine descriptive statistics. In addition, they are not exclusive, but combinations are often very useful.

On a spatial level, geovisualization can be followed by a standard GIS-analysis in order to test assumptions derived from geovisualization.

3 Software Overview

This chapter contains a general overview of investigated geovisualization software. Three of them will be given a closer look according to the set of methods proposed by Slocum et al. (2005), and finally evaluated in a quantitative way in order to choose one.

Available Software

Since geovisualization is not a clearly defined domain with strict limits, available software are not always described as typical geovisualization software. Nevertheless, there is a list of main software categorized as geovisualization tools. Among them, Geovista Studio is a development environment, on the basis of which two ready-to-use applications were developed (ESTAT and GeoViz Toolkit). Table 3.1a and 3.1b hereunder show the main characteristics of the software.

	CommonGIS	GeoDa	GeoWizard Lite
Web Adress	http://www.iais.fraunhofer.de/1863.html	http://geodacenter.asu.edu/geodasum	http://vita.itn.liu.se/geowizard
Version	2.2.0 (unsure)	0.9.5-i	2.0
Institution	European Commission / Fraunhofer Institute, Germany	Arizona State University / University of Illinois, USA	Linköping University, Sweden
Involved researcher	G. Andrienko, N. Andrienko	L. Anselin	No information.
History	CommonGIS is an extension of the thematic mapping software “Descartes”	First released Feb. 2003.	No information. (First released 2008?)
Main description	An easy to use tool to facilitate visual thinking, exploring / analysing geo-referenced data and decision making in a spatio-temporal context, involving a high interactivity.	An introduction to spatial data analysis, spatial autocorrelation statistics, as well as basic spatial regression functionality.	An application for visualization of geographical data with multiple attributes.
Applications already carried out in the field of...	Economics, landscape genetics, etc.	Health, economics, etc.	Social science data, demography of Sweden / worldwide.
Language	Java	C++	C# / Net
Readable file format	SHP, GML, CSV, GIF, etc.	SHP	XLS
Platform	Either (Java)	Win	Win

Table 3.1a General overview of some geovisualization software (Part 1)

	Geovista Studio	ESTAT	GeoViz Toolkit
Web Address	http://www.geovistastudio.psu.edu/jsp/index.jsp	http://www.geovista.psu.edu/ESTAT/index.html	http://www.geovista.psu.edu/geoviztoolkit/index.html
Version	1.2	No information.	0.8.5
Institution	Pennsylvania State University	Pennsylvania State University, in cooperation with National Cancer Institute	Pennsylvania State University / University of South Carolina
Involved researcher	M. Gahegan, A. MacEachren	M. Gahegan, A. MacEachren	F. Hardisty, A. Myers, K. Liao
History	Release date of current version: Sept. 2007	Developed initially to support cancer research.	No information.
Main description	An open software development environment designed for geospatial data. It's a programming-free environment that allows users to quickly build applications for geocomputation and geographic visualization.	An user-friendly, open-source software designed to support exploratory geographic visualization. ESTAT is designed to handle any kind of spatial data with attributes. It is based on Geovista Studio.	An application version of Geovista Studio for the visualization of multidimensional, geographic data. All of the components are coordinated with each other (exception: colour scheme).
Applications already carried out in the field of...		Health (National Cancer Institute)	Terrorism, demography
Language	Java	Java	Java
Readable file format	Divers, depending on existing JavaBeans, possibility of programming.	CSV, SHP	SHP
Platform	Either (Java)	Either (Java)	Either (Java)

Table 3.1b General overview of some geovisualization software (Part 2)

On the basis of these short descriptions, the three most interesting and differing software were chosen for further examination.

CommonGIS



CommonGIS is a GIS tool for the interactive analysis of thematic maps, derived from a previous version called Descartes and developed at the Fraunhofer Institute in Germany. It is Java-based and can be used either as a standalone version or as a Java applet. Its installation is very easy (depack and double-click).

Included in the software are some sample data sets, which can be easily loaded by looking for *.app* files in the Menu *Project > Open...* Similar to ArcGIS, there is a layer overview on the right side of each window. A layer contains a vector file and possibly related tables. A large amount of data formats are allowed, among which SHP, GML (vector), CSV (attributes) and GIF (image). The native data format is OVL, which is a vector data format derived from the predecessor Descartes. The relations between the loaded files and symbolisations can be saved as a CommonGIS project in the *.app* format (*Save as*).

During the analysis process, several different maps and charts can be generated, which are located in separate windows allowing comparison. When one part of the data is selected in a window, this part is highlighted also in all others.

In the documentation, the software is well explained by several examples.

Manipulating data: This is a strong side of the software; columns are added very easily to existing tables and there are several possibilities to calculate new data to be stored in existing tables.

Varying symbolisation: For every layer and object, the colour can be changed and different symbolisation possibilities are available: choroplete maps, symbols as circles, pie charts, bars, triangles and even own icons can be loaded.

Manipulating the user's viewpoint: standard navigation functions as scroll and zoom are provided, but are not specially user friendly. For the *3D View*, the viewpoint control lays besides the map, and direct mouse navigation in the map is not possible like it is the case e.g. with Google Earth. For an instant view of the movements, the options dynamic update has to be additionally selected.

Highlighting portions of a data: Data are highlighted in a thicker line style defined in the layer properties. For many types of maps, a dot plot at the right side allows to mask extreme values by dragging the range of the plot.

Multiple views: In the *Data Wizard*, the data can be visualized in one map per variable. Next to the map, a chart can be set to compare. Other windows can be arranged side by side. There is even a small window to show a map overview.

Animation: For time related data, the software offers animated maps with many possibilities to vary the parameters.

Linking maps with other forms of display: Besides maps, there is a large choice of charts as attribute statistics, histograms, graphs, scatter plots, etc. A selection shows up in all open windows.

Access to miscellaneous resources: New layers can be added anytime, as well as new tables and columns. There is a large choice of data formats, even raster data.

Automatic map interpretation: Although data calculations include also multidimensional analysis and decision aids, automatic map interpretation like self-organizing maps is not possible.

To sum up, CommonGIS is a geovisualization software focused on data manipulation and reading different file formats. As a drawback, it is rather static, since no brushing possibility is given and deselecting by simply clicking in an empty space is not possible. Symbolisation on the other hand is a good point, since even own symbols can be loaded.

GeoDa



GeoDa is a collection of software tools designed to implement techniques for ESDA. It is developed at the Arizona State University. Different windows, containing maps or graphs, are linked together in the way that a selection of one window shows in all others as well.

To access the software, a login has to be created before downloading and since it is based on C++, installation is necessary. Although only shape files can be opened, there are several possibilities to generate shape files out of ASCII-text. Still, only one shape file can be active, which demands a file containing all data in advance. In general, only one variable can be visualized, with exceptions for charts like parallel coordinate plots (PCP) and 3D-plots. If two variables need to be linked, two maps have to be generated to be visually compared.

The dynamic brushing of the selection tool across maps and charts facilitates interactivity, but it is only possible in a rectangular shape.

Manipulating data: A selection in the table can be directly transformed into a new column, containing a specific value for this selection. Some basic calculations are possible and can be assigned directly to a column.

Varying symbolisation: The colour of the map, the background and a selection can be changed. Colour schemes would help to pick colours faster. Other symbolisation possibilities are very poor: besides choroplete maps and a cartogram in form of different sizes of circles, there are no symbolisation options. Point data are displayed as small squares.

Manipulating the user's viewpoint: Although zoom in, zoom out and full extent options are provided by the menu showing up after a right mouse click, there is no drag option to wander across a section of the map.

Highlighting portions of a data: GeoDA has the possibility not just to select by a rectangular tool, but the selection shape can be chosen out of point, rectangle, circle, line and polygon. Selections have a fine crosshatch, which colour can be changed. Still, a chosen colour may not guarantee a good visibility if different colours show up in a map, e.g. a choroplete map. Nevertheless, a selection shows up in all open maps as well as in tables.

Multiple views: Maps, graphs and charts can be duplicated and set aside in different windows. There is a 3 by 3 map matrix with variable breaking points.

Animation: A cumulative or single map movie allows animating one variable.

Linking maps with other forms of display: Diverse plots, charts and graphs are provided, even more complex forms of display as Moran and LISA visualisation. The connection between maps and charts is strong and interactive.

Access to miscellaneous resources: Layers can be added any time, but only the topmost layer is active and this order cannot be changed. Only shape files can be imported, other formats need previous transformation.

Automatic map interpretation: No data mining features.

Overall, GeoDA is a classic geovisualization tool, offering few map choices and a large set of other forms of display. The fact that only one SHP-file can be analysed forces the user to well think of the structure of his data set. GeoDA main features are dedicated to the analysis of global and local spatial autocorrelation.

GeoViz Toolkit



GeoViz Toolkit is an application of the software development environment GeoVista Studio. Since it is a Java application, it doesn't need installation. When started, the interface offers different windows with an example data set loaded. Other data sets, in the form of ESRI-SHP, DBF or CSV overload the existing data. It is only possible to load one main and one background file into the toolkit. At the top, one can add more windows by choosing one of about 30 different types of tools (*Add Tool*). The tools are all linked with each other for selection.

Interactivity is very high, in the sense that just a mouse movement highlights areas and plot elements. Some operations are nevertheless a bit clumsy to use (e.g. all extra buttons for order, scale, translate, brush). Unfortunately, the available documentation is not complete and not up to date. Therefore, some of the quite nice looking tools remain unexplained. Although most of them are self-explanatory, some would require at least the label written out or shortly explained (e.g. PCAViz, MoranMap).

Manipulating data: The function *Variable Transformer* allows to add content into a new columns. The variables can be transformed by different operation as normalization, addition, etc. The new columns cannot be labelled.

Varying symbolisation: Variables can be plotted as choroplete maps, star plot maps, and classic symbols as triangles, squares, etc. They can only be changed in colour and in size, whereas for the size no graduation is possible. In addition, a sound classification tool is available, which allows the user to have another dimension to perceive the data.

Manipulating the user's viewpoint: In each window with a map, it is possible to drag and zoom, moreover a fisheye lens and magnifying lens offer a quick closer look.

Highlighting portions of a data: A rectangular tool does select the data, but even just wandering over the map highlights the actual area plus its neighbours. Clicking into diagrams highlights the selection in all other maps as well.

Multiple views: Besides the different windows which can be set aside each other, a matrix of maps can be generated, as well as matrices of maps combined with graphs.

Animation: *The Selection Animator* and the *Indication Animator* are available animation tools, but both can be applied only for one variable.

Linking maps with other forms of display: GeoViz offers a large amount of graphs and plots for data visualization: scatter and parallel plot, link graph, space fill (areas are ordered according to an attribute.), histogram, table viewer, etc. Some of the tools would need additional explanation in order to know how to use them correctly.

Access to miscellaneous resources: The toolkit seems to have some satellite image capabilities (*SaTScan*).

Automatic map interpretation: There seems to be no function allowing automatic map interpretation.

On the whole, GeoViz Toolkit offers a lot of clickable areas, allowing a high user interactivity. A large set of tools is available, but there is also the risk of loosing the overview on all open tools. Compared to the amount of graph and plot variants, the choice for different map styles is small. Nevertheless, every window is connected in the sense that a selection has an effect on all windows, highlighting the data. Not highlighted data has a nice blur effect, what gives a good impression for a not so important background. Otherwise, graphics in this toolkit is not very nice, but functional and highly interactive.

Selection

After having examined the three chosen software in more details, I summarized the results in a quantitative way:

Grade 1 means a weak point under the average of all three.

If it's average, or just a feature common to many applications, the grade is 2.

Exceptional ideas and outstanding solutions are set to 3.

In addition to the points derived from Slocum et al. (2005), interactivity, user-friendly design and documentation were added to the evaluation. The weight is lifted to 2 in some cases (characters in bold), because these elements can be seen as the backbone of Geovisualization.

	Weight	CommonGIS	GeoDA	GeoViz Toolkit
Manipulating data	1	3	2	1
Varying symbolisation	2	3	1	2
Manipulating the user's viewpoint	1	2	1	3
Highlighting portions of a data	2	3	2*	3
Multiple views	1	2	2	2
Animation	1	2	2	2
Linking maps with other forms of display	1	2	2	3
Access to miscellaneous resources	1	3	2	1
Automatic map interpretation	1	1	1	1
Interactivity	2	1	2	3
User-friendly design	1	1	2	2
Documentation	1	3	3	1
Total (max. 45)		33	27	32

Table 3.2 Evaluation grid of geovisualization software

CommonGIS and GeoViz Toolkit obtain both a good result, being just one point apart. The advantage of high interactivity on the GeoViz Toolkit side nearly levels the data manipulation capabilities and the detailed software documentation of the CommonGIS side. Since CommonGIS was already applied to landscape genetics (Joost, 2006), GeoViz Toolkit was selected to analyse genetic diversity in Swiss goat breeds, and its relationships with environmental information.

* The choice of different selection shapes in GeoDA is a plus point, but the bad visibility of the crosshatch for some colours equals in an average grade.

4 Application to Landscape Genetics

After having examined the available data and defined some major questions, a model geovisualization process was applied to a 4 km² grid of Switzerland, with interpolated environmental and genetic data. The analysis described in this chapter uses the GeoViz Toolkit, in particular the tools *GeoMap*, *RadViz*, *SingleHistogram* and *ParallelPlot*. A summary of all possible derived hypotheses is proposed at the end of this chapter.

Data Sets¹

The file *CommunesCH.shp* contains socio-economic and environmental variables describing the Swiss municipalities. Moreover, a table of interpolated genetic diversity was added. The genetic diversity was computed from about 100 farms in Switzerland and the neighbouring countries during the Econogene project funded by the European Union².

FarmsBuffer.shp still includes the original genetic data per farm. Nevertheless some changes were applied, since GeoViz Toolkit cannot interpret point data. The raw point data representing farms was therefore transformed into a circular polygon around the farms. In this data set, we can see that about ten different goat breeds were examined: Bionda dell'Adamello (Italy), Camosciata delle Alpi (Italy), German Alpine (Germany), Grisons Striped (Switzerland), Orobica (Italy), Peacock Goat (Switzerland), St. Gallen Booted goat (Switzerland), Swiss Alpine (Switzerland), Valais Black Neck (Switzerland) and Valdostana (Italy).

SGGRID-CH.shp, shows a 4 km² cell grid (vector) on the whole area of Switzerland with genetic diversity, original farm data and some environmental variables. The genetic diversity variable was interpolated from farm data, and environmental data were produced according New et al. (2002). It has to be kept in mind, that interpolation regionalizes data and therefore is likely to lead to approximations.

When the data was first imported into GeoViz Toolkit, the projection system was a problem to solve. In general, the software requires the global longitude-latitude system. Initially, the file *CommunesCH.shp* was not projected (non earth-system). This was modified in the Manifold software and the file was first transformed into the New Swiss Grid projection system, and then projected into the longitude-latitude system, in decimal degrees.

Some attention had to be paid to text data type while importing data into GeoViz Toolkit, because this software does not support text variables.

Concerns about the interpolation of genetic diversity in *CommunesCH.shp* were already mentioned. Comparing the genetic diversity within municipalities and within the cells of the grid clearly highlights differences (see Fig. 4.1).

¹ All data sets were provided by Stéphane Joost.

² www.econogene.eu, Retrieved 7.5.09

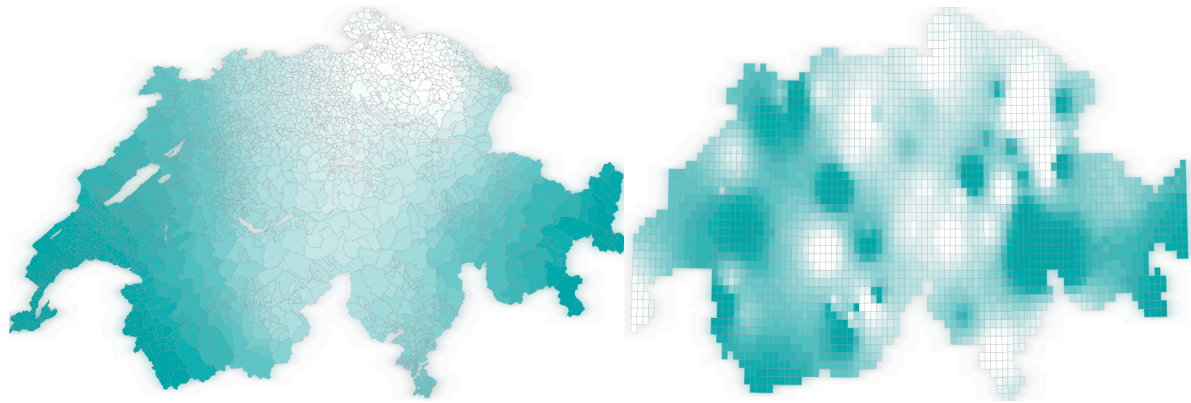


Fig. 4.1 Choroplete map of genetic diversity per municipality (left) and per 4 km² cell (right). High diversity is shown with a dark hue, and low diversity with a light one.

Since the information is regionalized, the grid of 4 km² doesn't show the real, primary data. Nevertheless, the grid data is enough accurate to get an impression of the data and to understand possible correlations. It seemed to be the best suitable for geovisualization analysis in GeoViz Toolkit and was therefore used in the rest of the analysis.

Main Questions

Before starting data analysis in GeoViz Toolkit, I elaborated some questions as a guideline for data exploration. Since the process is exploratory, questions are of a more general type and do not imply hypothesis like in the context of confirmatory analysis.

The main goal of the analysis is to find possible influences of any environmental parameter affecting genetic diversity in Switzerland. Do relationships between environmental and genetic data exist, and if yes, what particular association could be detected?

Possible Analysis Process

Geovisualization, as mentioned before, is a highly interactive approach to data analysis. Therefore, an overall workflow cannot be decided in advance, nor can it be adopted from prior projects. But some important stages in the exploratory process have to be applied: viewing the data and its structure, filtering or zooming, and as a consequence having a particular look on details (Shneidermann 1996). In this case, the data and its structure are visualized in the beginning, and an iterative process of filtering, zooming and looking at the details is then applied.

After having loaded *SGGRID-CH.shp* into GeoViz Toolkit, I start with a look at the table (named *SpreadSheetBean*). The variables are listed into columns and each observation has a line. Looking on the data in the *SpreadSheetBean*, the amount of the data doesn't allow a direct interpretation of the relations. With this fact in mind, a geovisualization approach can be justified.

In order to get acquainted with the data, each variable is plotted in a simple choroplete map (*GeoMapUni*). A quantile classification was applied to the data to display the maps.

The genetic diversity (variable GeneticDiv) can be considered as the main variable to be tested in combination with all others. It is a measure of heterozygosity and was calculated according to Joost (2006, p. 52-53). Expressed in percent, it ranges from 0.456472 (45.6%) to 0.692443 (69.2%). Since the mean value exceeds 50%, a normalized column newCol was created from GeneticDiv. Whenever a comparison with over or under average genetic diversity is required, the values of newCol could be used as well.

After a first look at each variable, a combination of two of them can be applied in *GeoMap*. For GeneticDiv, a nearly continuous colour scheme was chosen with blue indicating a low genetic diversity and green a high one. The second variable changes the colour scheme with different degrees of brightness. As an example, a dark green square signifies a high genetic diversity and a high diurnal temperature range (correlation). Three or even four different colours, black and white excluded, result in a highly complex colour scheme and is not recommendable.

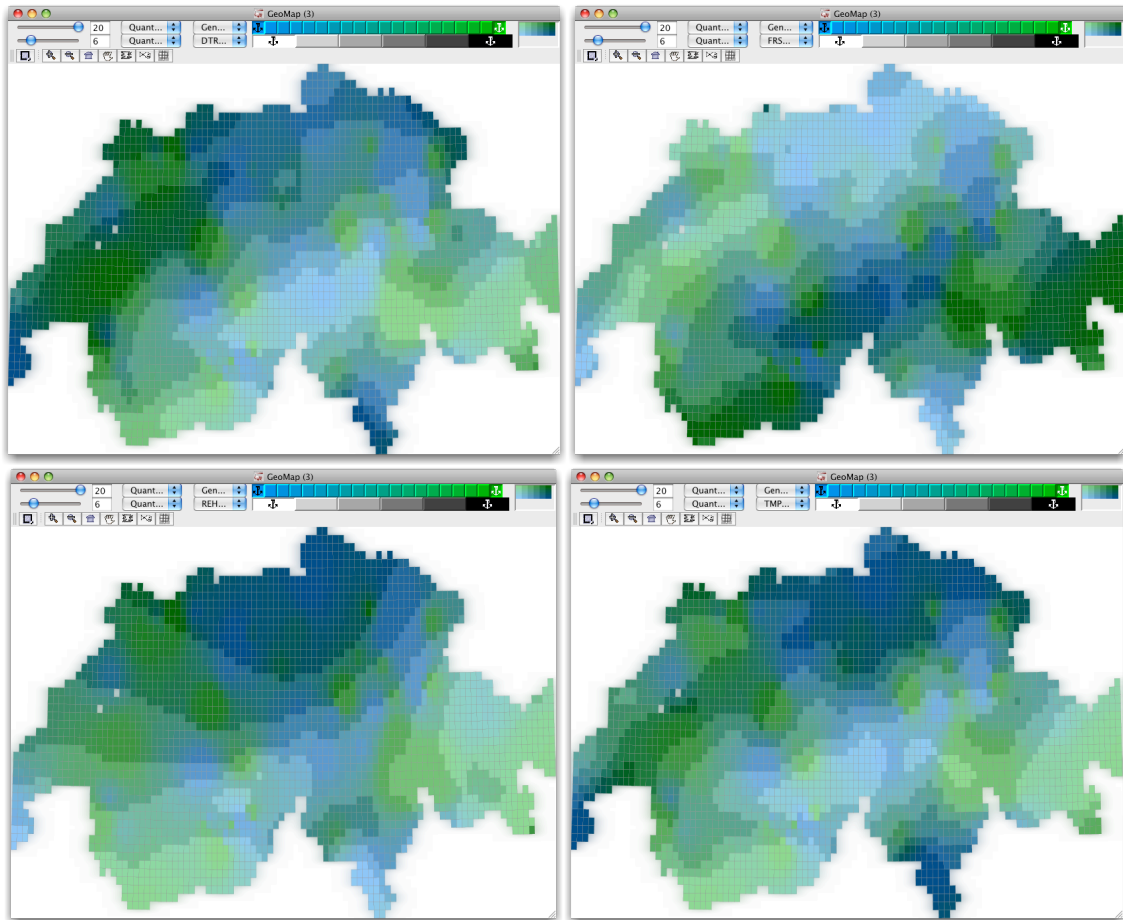


Fig. 4.2 Genetic diversity compared to DTRyearlym (top left), FRSyearlym (top right), REHyearlym (bottom left) and TMPyearlym (bottom right). See the appendix to find the description of the variables.

Correlations between genetic diversity and environmental parameters can hardly be detected in figure 4.2. This way of comparing two variables apparently doesn't permit to reveal any possible hypothesis in this case.

When more than two-dimensional data have to be visualized, *RadViz* serves as an add-on tool for the map. *RadViz* allows visualizing multidimensional data using the Hooke's law taken from mechanics. This law of elasticity is an approximation stating that the extension of a spring is in direct proportion with the load added to it as long as this load does not exceed the elastic limit. In the geovisualization context, each variable has an anchor point, often aligned on a circle. From each anchor, a spring connects the data point under investigation. Depending on the value of the observation, the stiffness of the spring varies. Levelling the forces of all the attached springs therefore derives the position of a data point (Nováková et al., 2006; Brunson et al., 2009).

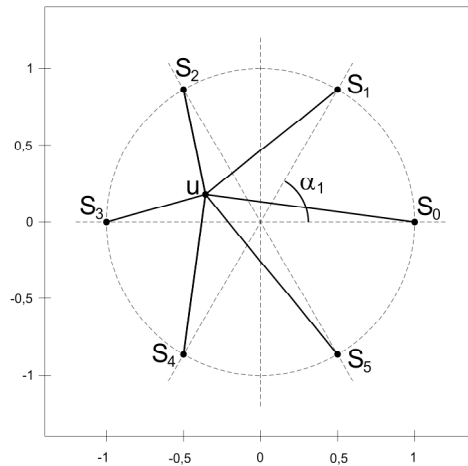


Fig. 4.3 An observation attached by different springs to the variables. (From Nováková et. al., 2006, p. 471)

I applied *RadViz* to four variables related to precipitation, humidity, and genetic diversity. The arrangement of the variables around the circle can be changed, since a different order strongly affects the distribution of the points. Guidelines to determine best arrangements don't exist, so, after some tests, I chose the one resulting in the most interesting shape. As usual with GeoViz Toolkit, the tools are linked together, showing respectively the same selection of geographical units, and brushing is always active. But as a drawback, it is not possible to select individuals in the *RadViz* tool.

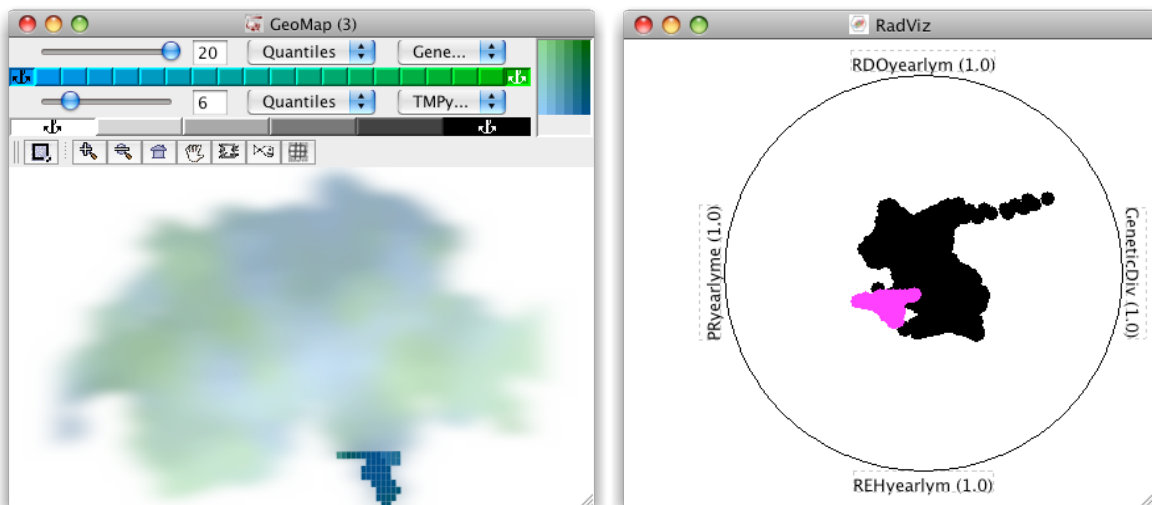


Fig. 4.4 GeoMap (left) and RadViz (right) showing selected cells with a low genetic diversity.

The analysis with *RadViz* of variables related to precipitation and humidity in combination with genetic diversity shows a significant regional relationship on a choroplete map. As GeneticDiv and RDO show relatively low values, PRyearlyme and REHyearlym are quite high. It means that in this region (southern Tessin) there are more wet days, and rain is more intense than in other parts of Switzerland, while GeneticDiv is quite weak.

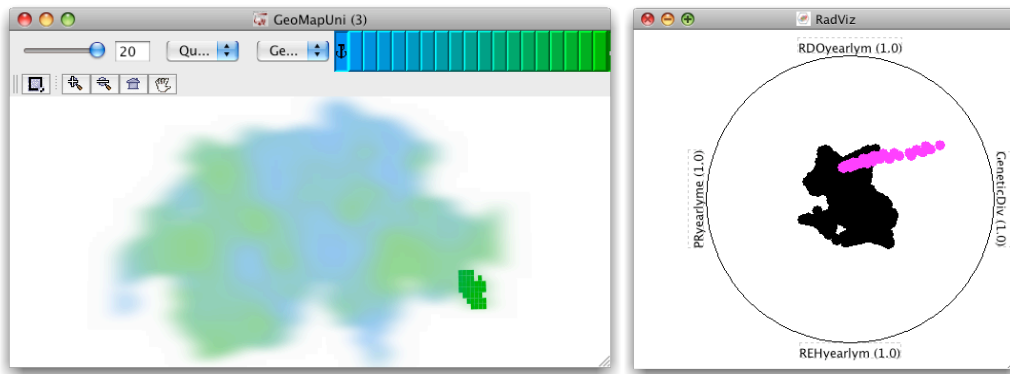


Fig. 4.5 The region of Poschiavo shows interesting correlations in RadViz.

On the other hand, I was interested in the reasons that could be found to explain the high genetic diversity (and high RDOyearlym) in the upper right part of *RadViz*. Selecting the corresponding region (valley of Poschiavo) in the map reveals a strait line in the *RadViz*. While the genetic diversity is constantly rising, RDOyearlym is slightly increasing and the other two variables are weakening. We can see here that a low humidity level and little precipitation possibly favour a higher diversity. This hypothesis has to be tested, and perhaps also other influences supposed.

Examining this hypothesis with the help of an animation tool (*SelectionAnimator*) didn't underline this correlation. It rather showed the valley of Poschiavo as a special case, with low precipitation values, unlike other region with high genetic diversity. As a consequence, there are probably other reasons than just precipitation data. An advanced analysis could imply the search for other explanatory parameters by exploratory data analysis, as well as a further reasoning in order to reject or not the hypothesis in a confirmatory context.

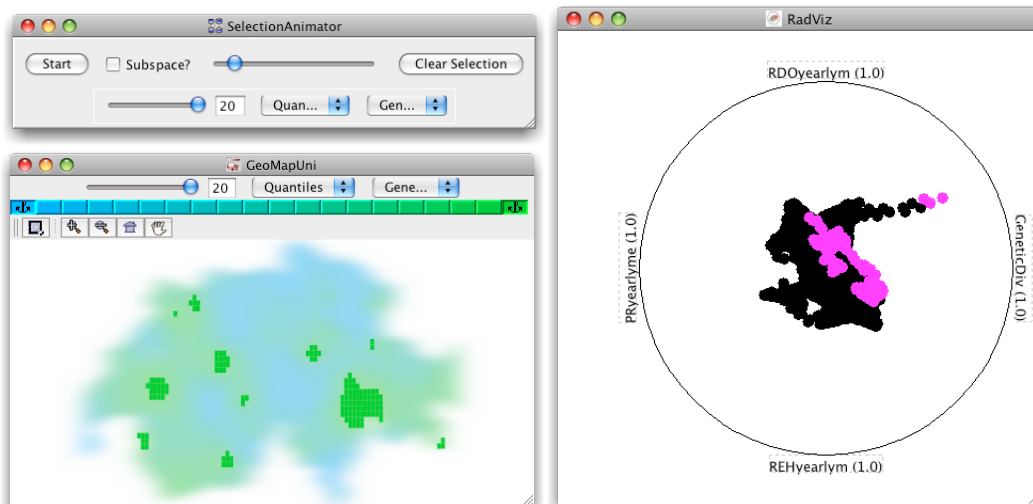


Fig. 4.6 The last frame of the animation shows the highest genetic diversity values.

As an additional step, I chose to include temperature data. A two-dimensional choroplete map was produced (figure 4.7). Following an intuitive perception, blue as a cold and yellow as a warm colour were used to represent temperature. At the same time, any saturated, dark colour indicates a high diurnal temperature range, and a light colour a low one.

The third dimension, the genetic diversity, was inserted with the help of a histogram plot (*SingleHistogram*). It shows the distribution of the variable over the domain. Selecting the mean class permit to visualize the corresponding geographical units on the map. As shown in figure 4.7 hereunder, no apparent correlation can be found this way. All temperature classes seem to be represented.

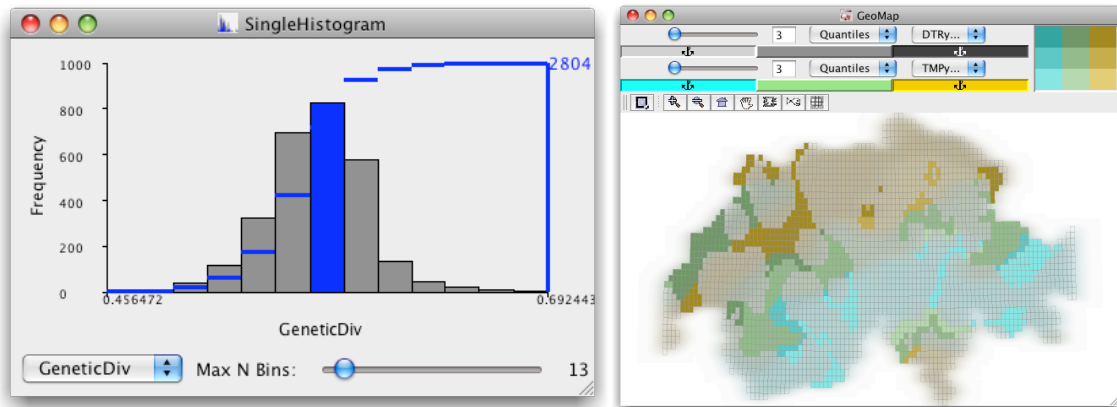


Fig. 4.7 Selecting mean genetic diversity values in the histogram doesn't permit to highlight correlations on the map.

As we are not that much interested in mean genetic diversity, and as high values in the histogram are difficult to select, the *ConditionManager* tool was used. This tool allows selecting a certain range of one variable according to a quantitative criteria. The selection is visualized in all other active tools. Thus, using the *ConditionManager*, a genetic diversity higher than 60% was chosen. As a result, again, no apparent correlation in the map could be found.

The parallel coordinates plot tool (*ParallelPlot*) was applied for a new step. In the tutorial of ESTAT, another ready-to-use software from Geovista Studio, the parallel coordinates plot is described as “a tool designed to display multi-dimensional data in a visually accessible format. It works by converting data categories into axes and then drawing lines from one category to another based on the values at each intercept.” (Robinson, 2005)

For instance, several variables characterising cars would look as follows in a parallel coordinates plot:

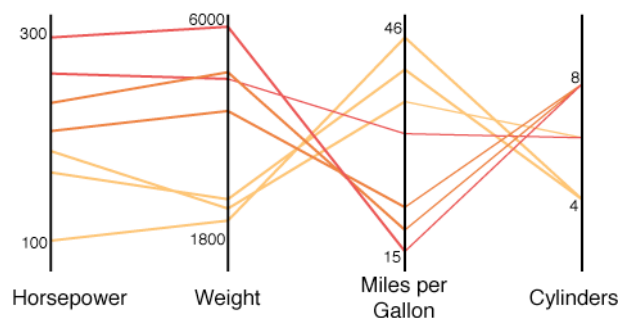


Fig. 4.8 A simple parallel coordinate plot of a car. (After Robinson, 2005)

Observations of a variable spread along the vertical axes. Lines between the axes represent the different observations. With the help of a parallel plot, positive and negative correlations can be detected easily. In the example above, there is a positive correlation between horsepower and weight, and a negative one between the weight and miles per gallon. A parallel coordinates plot is also very useful to detect outliers.

In GeoViz Toolkit, the parallel coordinates plot can be drawn for up to six variables. As in *RadViz*, different arrangements of the axes are crucial to have a view of all relations in the data. They can be moved interactively, the scaling can be changed and the axes inverted. In the cases described hereunder, I used the *min-max scale* as default. It scales each axis in order to fit the display window. If negative correlations occur, axes are inverted to get a smoother view of the data. The colour is set according to the genetic diversity with a similar colour scheme like with the *RadViz* analysis.

In the figure below, two parallel coordinate plots were produced and are here set side by side to have an overall view. In general, a rather high genetic diversity is highlighted.



Fig. 4.9 A parallel coordinate plot containing all genetic and environmental variables.
(The figure is actually a combination of two plots.)

This shows an interesting structure. The highlighted data concentrates in two strings, either with relatively high values and well bundled, or with relatively low values, containing fewer observations and more irregularities. Especially the low value string is clearly parted at SUNyearlym and precipitat. For the upper string, a very narrow distribution is found in the cases of precipitat, REHyearlym, and SUNyearlym. Since the upper one contains more and “greener” observations, we could derive a hypothesis from it: A relative high genetic diversity needs relative high values for PRyearlyme, precipitat, FRSyearlym, RDOyearlym and SUNyearlym and relative low values for DTRyearlym, REHyearlym and TMPyearlm.

Inversing this hypothesis leads to a selection in the same domain, but on another axis. But this doesn't produce a corresponding result; there are much more values with an upper-average sunshine than just those with high genetic diversity. So, the hypothesis can probably not be knocked over. Nevertheless, the highlighted sunshine range doesn't involve any of the low genetic diversity values.

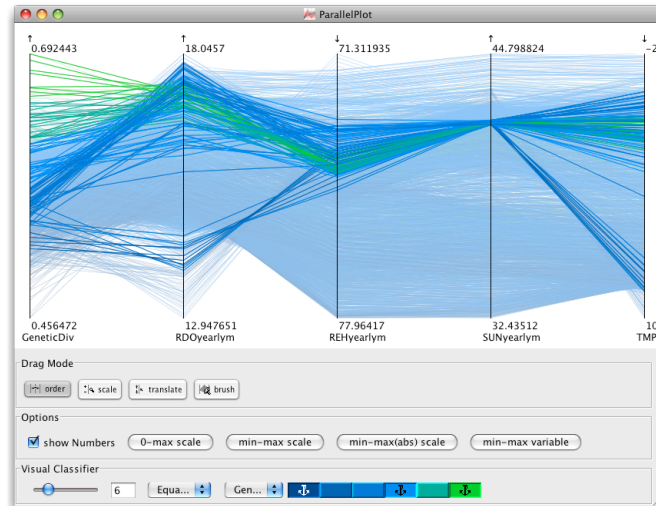


Fig. 4.10 Parallel coordinate plot with a selection in upper-middle part of the distribution of SUNyearlym values.

Until now, we looked at the parallel coordinate plot with a selection of high genetic diversity values. Correlations for low genetic diversity are roughly the same, but the range is larger and therefore interpretation is not that clear as before.

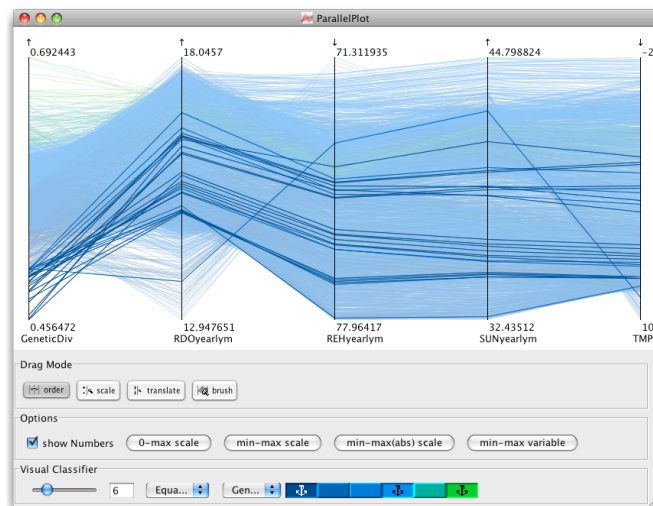


Fig. 4.11 Parallel coordinate plot with geographic units showing low genetic diversity.

Further comparisons with a choroplete map and other tools can be carried out. This leads in the end to an analysis interface adjusted for this type of data set (see figure 4.12).

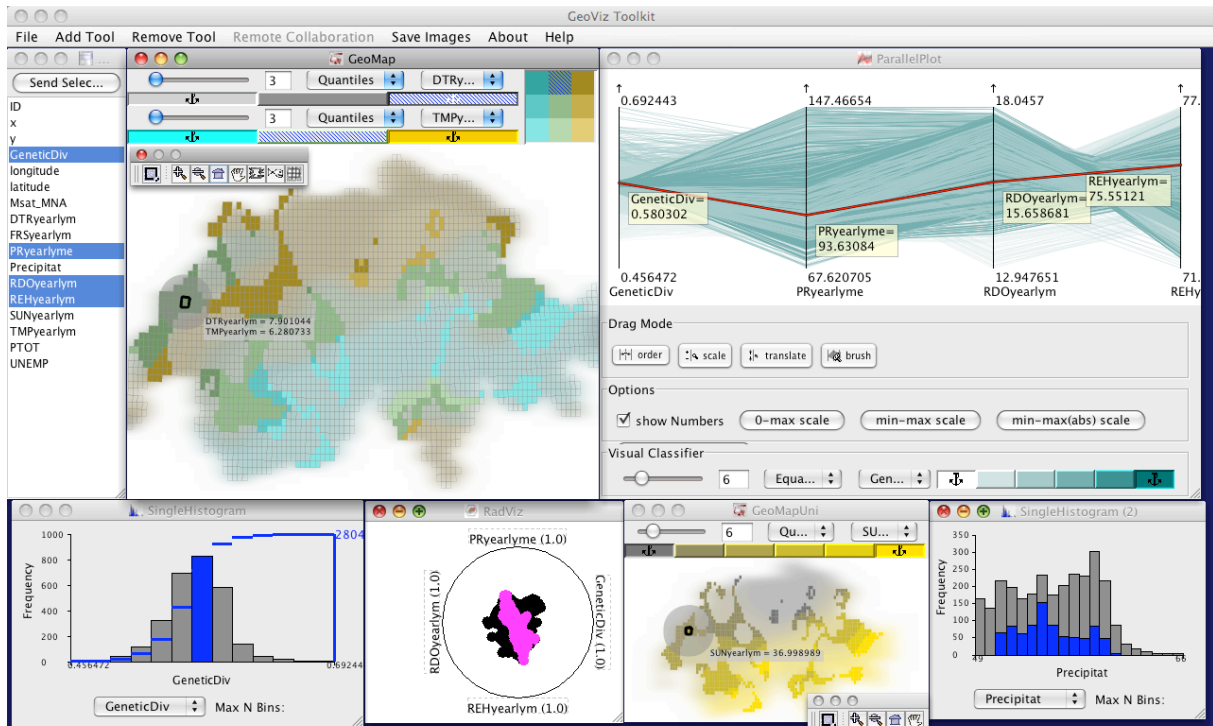


Fig. 4.12 An example of arrangement combining the different tools.

Other tools as *SonicClassifier*, *GeoMapCartogram*, *SingleScatterPlot*, *LinkGraph*, *StarPlot*, etc. were not described here, since some of them don't work correctly and others are not sufficiently efficient to treat this kind of data (*StarPlot*). Nevertheless, the set of tools I exploited in the context of this exploratory process permitted to obtain a good insight into the data.

Derived Hypothesis

Geovisualization doesn't lead to ready-to-use solutions. It has to be considered as a starting point to investigate any unknown data set. Often, it is followed by GIS, which can show results provided by statistical tools in the context of confirmatory analysis. Therefore, I collected the diverse hypotheses I was able to produce during geovisualization, to provide a basis for a possible further confirmatory analysis

I found five hypotheses:

1. In southern Tessin, there are more wet days, and rain is more heavy or intense than in other parts of Switzerland, while GeneticDiv is quite weak.
2. According to results found in the Poschiavo valley, we suppose that low humidity and few precipitation favour a higher genetic diversity.
3. The valley of Poschiavo is a special case, with low values in precipitation or humidity variables, unlike other region with high genetic diversity. Other explanatory factors probably influence the level of genetic diversity.
4. A relatively high genetic diversity requires relatively high values of PRyearlyme, precipitat, FRSyearlym, RDOyearlym and SUNyearlym, and relatively low values of DTRyearlym, REHyearlym and TMPyearlm.
5. The hypothesis n° 4 cannot be inverted.

5 Conclusion

Even though geovisualization is not yet widely applied, it can be a useful way to explore unknown and complex data structures. Nevertheless, it requires an excellent knowledge of its limits. It is not an “all-in-one” and “easy-to-use” solution, but rather an approach helping to get a first impression of a data structure and, possibly, to extract some knowledge from a data set, and encouraging further investigation. Beyond these observations, it is important to master each visualisation tool very well, e.g. the RadViz tool with its limit in ordering the variables around the circle.

Regarding GeoViz Toolkit, the software has a huge potential for development, as it is not very stable and some features simply don't work. Anyway, it was really adequate to show the possibilities in geovisualization: the combination of tools and the way they can be arranged allow the user to get a good overview on the data. The selection of subgroups and the intuitive way of manipulating the interface permitted very quickly to detect interesting facts and anomalies, which deserve further analyses.

In the future, standard GIS-software, along with the development of the digital world, will probably improve in terms of interactivity, and therefore a fusion with geovisualization tools should be considered. Even though the cost of such integration would probably be quite important, this will increase the quality and the usability of GIS tools.

6 References

- "GeoVISTA Studio Project." Retrieved 25.3.2009, from <http://www.geovistastudio.psu.edu/jsp/index.jsp>.
- "GeoViz Toolkit | GeoVISTA Center." Retrieved 25.3.2009, from <http://www.geovista.psu.edu/geoviztoolkit/index.html>.
- "CommonGIS." Retrieved 25.3.2009, from <http://www.iais.fraunhofer.de/1863.html>.
- "ESTAT - Exploratory Spatio-Temporal Analysis Toolkit." Retrieved 25.4.2009, from <http://www.geovista.psu.edu/ESTAT/>.
- "GeoDa Center for Geospatial Analysis and Computation." Retrieved 25.3.2009, from <http://geodacenter.asu.edu/>.
- "GeoWizard Lite: geowizard: VITA: Linköping University." Retrieved 2.4.2009, from <http://vita.itn.liu.se/geowizard>.
- Andrienko, N., Andrienko, G. (2006). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin, Springer.
- Brunsdon, C., Fotheringham, A. S., Charlton, M.E. "An Investigation of Methods for Visualising Highly Multivariate Datasets". Retrieved 24.4.2009, from <http://www.agocg.ac.uk/reports/visual/casestud/brunsdon/brunsdon.pdf>.
- Demšar, U. (2006). *Data mining of geospatial data: combining visual and automatic methods, Urban Planning*: 91.
- Joost, S. (2006). *The geographical dimension of genetic diversity a GIScience contribution for the conservation of animal genetic resources*. Lausanne.
- Joost, S., Pointet, A. (2007). *Animal Farming and GIS service. WAAP Book of the year 2007*: 175 - 185.
- MacEachren, A. M. (1995). *How Maps Work: Representation, Visualization, and Design*. New York [etc.], Guilford.
- New, M., Lister, D., Hulme, M., Makin, I. (2002). *A high-resolution data set of surface climate over global land areas. Climate Research: Vol. 21*: 1-25.
- Nováková, L., Stepánková, O. (2006). *Multidimensional clusters in RadViz. Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*. Lisbon, Portugal, World Scientific and Engineering Academy and Society (WSEAS).
- Robinson, A. (2005). *Getting to know ESTAT: The Exploratory Spatio-Temporal Analysis Toolkit*.
- Shneiderman, B. (1996). *The eyes have it: a task by data type taxonomy for information visualizations. Visual Languages*.
- Slocum, T. A. (2005). *Thematic Cartography and Geographic Visualization*. Upper Saddle River, N.J., Pearson-Prentice Hall.

7 Appendix

Variables of SGGRID-CH.shp

Variable	Description	Format	Value range	
ID	Identification number		9469.0	12272.0
x	Longitude in decimal degrees		6.183605	10.351171
y	Latitude in decimal degrees		45.805652	47.824988
GeneticDiv	Genetic diversity (a rate)		0.456472	0.692443
farmid	ID of the farm where the animal was sampled	ANSI text		
longitude			0.0	9.88302
latitude			0.0	47.5596
regionFarm	Place where the farm is located	ANSI text		
breed	The name of the breed	ANSI text		
ANIMAL	ID of the animal	ANSI text		
Msat_MNA	Another variable for genetic diversity (a number of alleles)		0.0	3.266667
DTRyearlym	Diurnal temperature range, yearly mean		5.165095	8.945284
FRSyearlym	Number of days with ground frost, yearly mean		7.581879	23.308244
PRyearlyme	Precipitation in mm/month, yearly mean		67.620705	147.466537
Precipitat	Coefficient of variation of precipitation, yearly mean		49.768912	66.866592
RDOyearlym	Number of days with > 0.1mm rain per month, yearly mean		12.947651	18.045701
REHyearlym	Relative humidity in percent, yearly mean of monthly values		71.311937	77.964176
SUNyearlym	Percent of maximum possible sunshine (percent of daylength), yearly mean		32.43512	44.798823
TMPyearlym	Temperature, yearly mean		-2.570599	10.761986
PTOT	Population indication		0.0	363273.0
UNEMP	Unemployment indication		0.0	5.030181