

Development of a Statistical Analysis Tool for GIS with an Application to Social Sciences in Uganda

Karin Gullstrand
Maria Ljungblom

Avdelningen för Fastighetsvetenskap
Lunds Tekniska Högskola
Lunds Universitet

Department of Real Estate Science
Lund Institute of Technology
Lund University, Sweden



ISRN LUTVDG/TVLM 03/5075 SE



Lund Institute of Technology
Lund University

Department of
Real Estate Science
Lund Institute of Technology
Lund University
Box 118
SE-221 00 Lund
SWEDEN

Development of a Statistical Analysis Tool for GIS with an Application to Social Sciences in Uganda

Master of Science Thesis and a Minor Field Study by:
Karin Gullstrand and Maria Ljungblom

Supervisors :

Dr. Lars Harrie at the Department of Technology and Society at Lund Institute of Technology.
Ass. Prof. Petter Pilesjö at the Department of Physical Geography at Lund University.
Dr. Joy C. Kwesiga at the Faculty of Social Sciences at Makerere University.

Examiner:

Dr. Lars Harrie at the Department of Technology and Society at Lund Institute of Technology.

Keywords :

Uganda, Kampala, Makerere University, Faculty of Social Sciences, GIS, integrated geographical statistical analysis tool, Sida

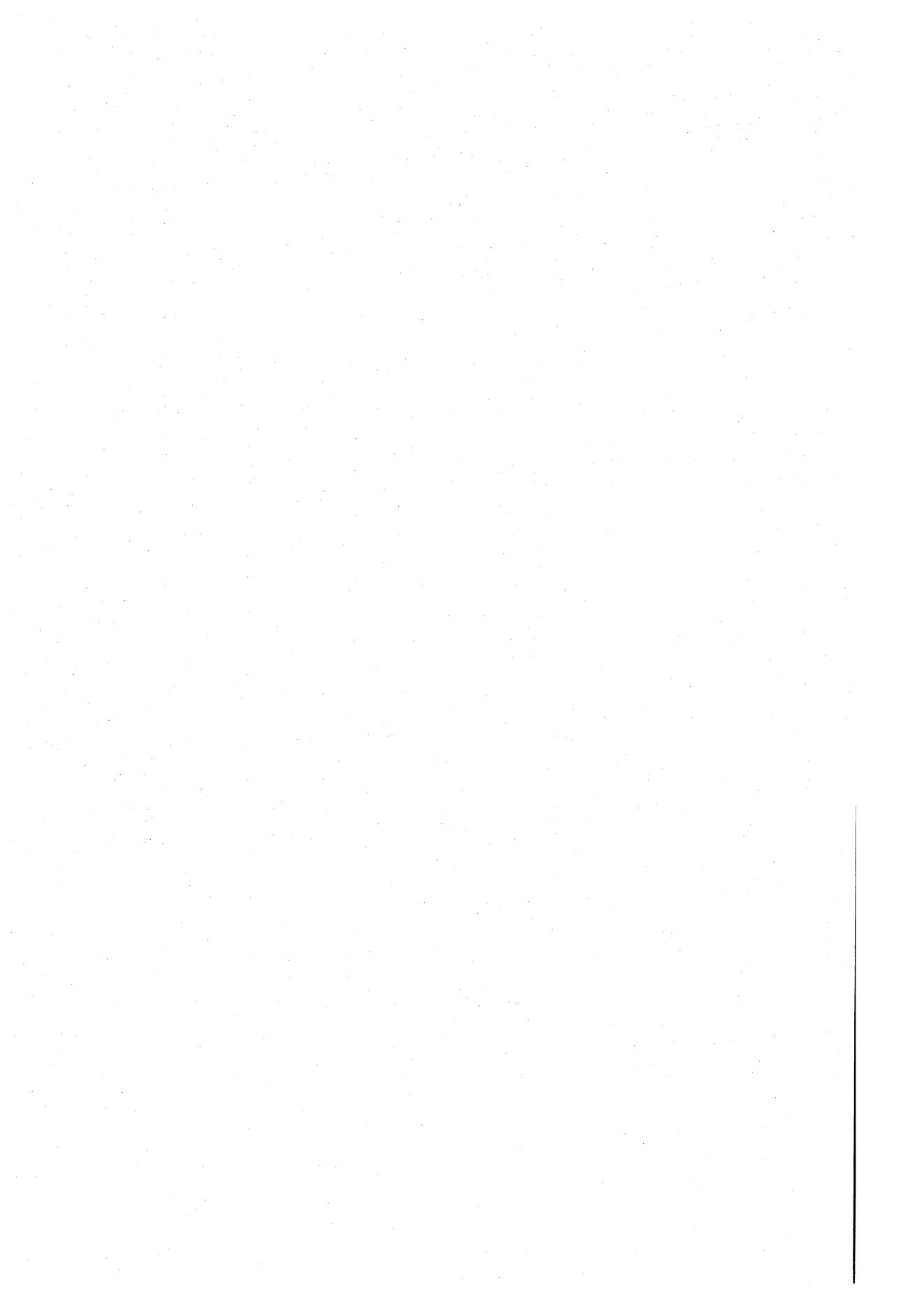


Foreword

This Master of Science thesis, at Lund Institute of Technology in Sweden, corresponds to one semester at the Surveying education. Our project has been performed in cooperation with the Department of Technology and Society at Lund University, the GIS center at Lund University and the Faculty of Social Sciences (FSS) at Makerere University in Kampala, Uganda. Our project has been performed during January 2002 to January 2003.

The supervisors of our project have been Dr. Lars Harrie at the Department of Technology and Society at Lund Institute of Technology, Ass. Prof. Petter Pilesjö at the Department of Physical Geography at Lund University, and Dr. Joy C. Kwesiga at the Faculty of Social Sciences at Makerere University.

We would like to thank our supervisors Lars Harrie, Petter Pilesjö and Joy Kwesiga for their great assistance during our work. We would also like to send a special thanks to Margareta Espling for helping us with the part concerning Iganga and to Bertil Egerö and Edward Kirumira for the HIV/Aids data. Finally, we want to thank Alfred Tingo for all the practical help he gave us during our first confused weeks in Kampala.



Abstract

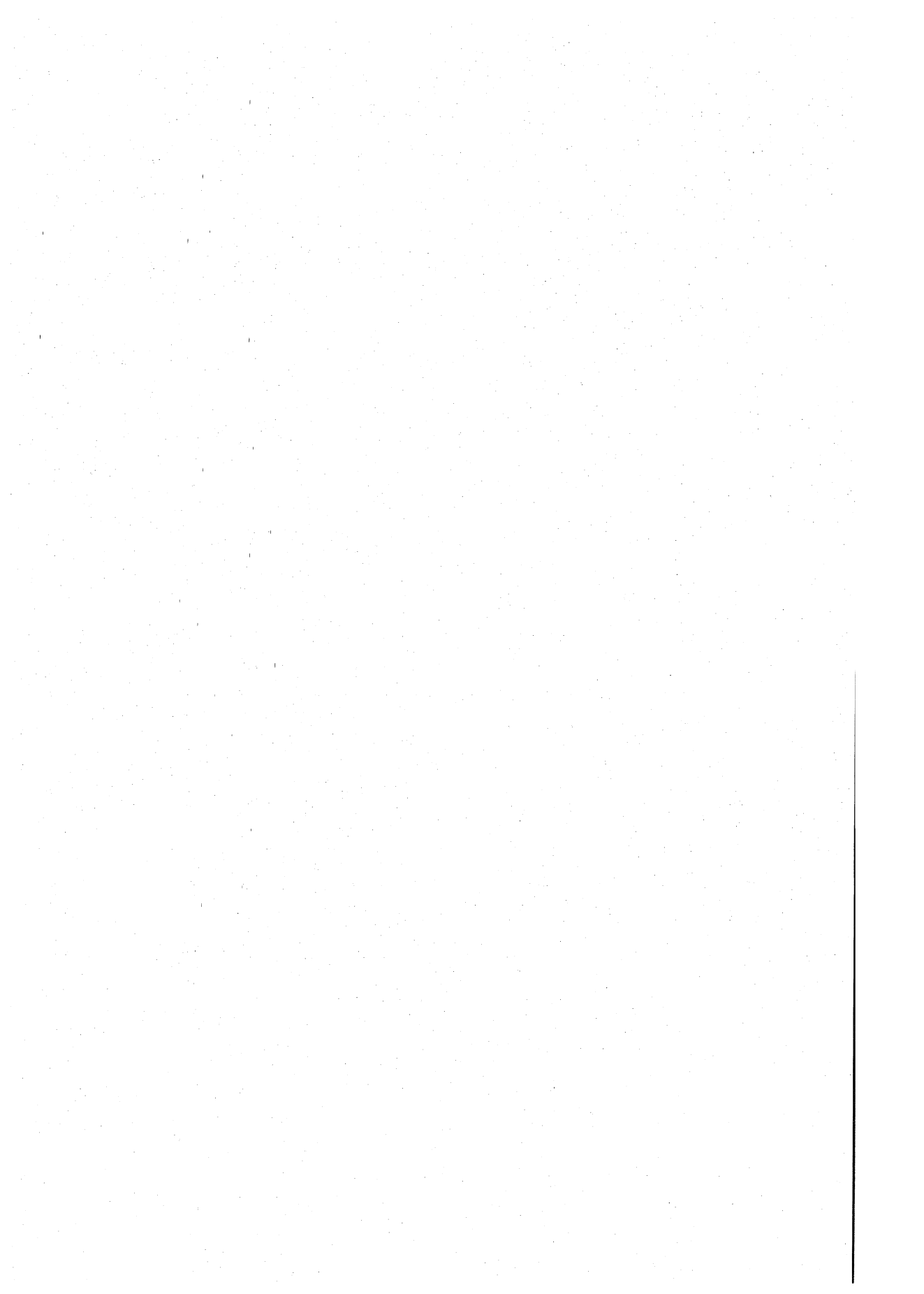
This report describes the development and implementation of a statistical analysis tool integrated in a GIS to be used in conducting research within social sciences. The application is developed at Lund University in Sweden and is followed by a minor study applied to social sciences in Uganda. The purpose of the minor study is to introduce GIS and the tool in research within the Sida project "Consolidation Peace and Development in the Lake Victoria Region and its Environs: The National and Local Responses to Transformation from Turmoil to a more Sustainable Development Process" in Uganda. The Department of Peace and Development Research at Gothenburg University is the Swedish coordinator of the Sida project. The Sida project is performed in cooperation with the Faculty of Social Sciences (FSS) at Makerere University in Kampala, Uganda.

The statistical analysis tool was developed in ArcMap using Visual Basic for Applications (VBA). To manage tables, map, graph etc. in ArcMap the COM-library ArcCatalog was used. The different functions of the tool are: creating new geographical attributes, showing the relationship between two attributes in a graph, and performing a multiple regression analysis.

During a one-week GIS workshop for researchers and teachers at FSS at Makerere University, problems and possibilities concerning the implementation of GIS and the new statistical tool were identified.

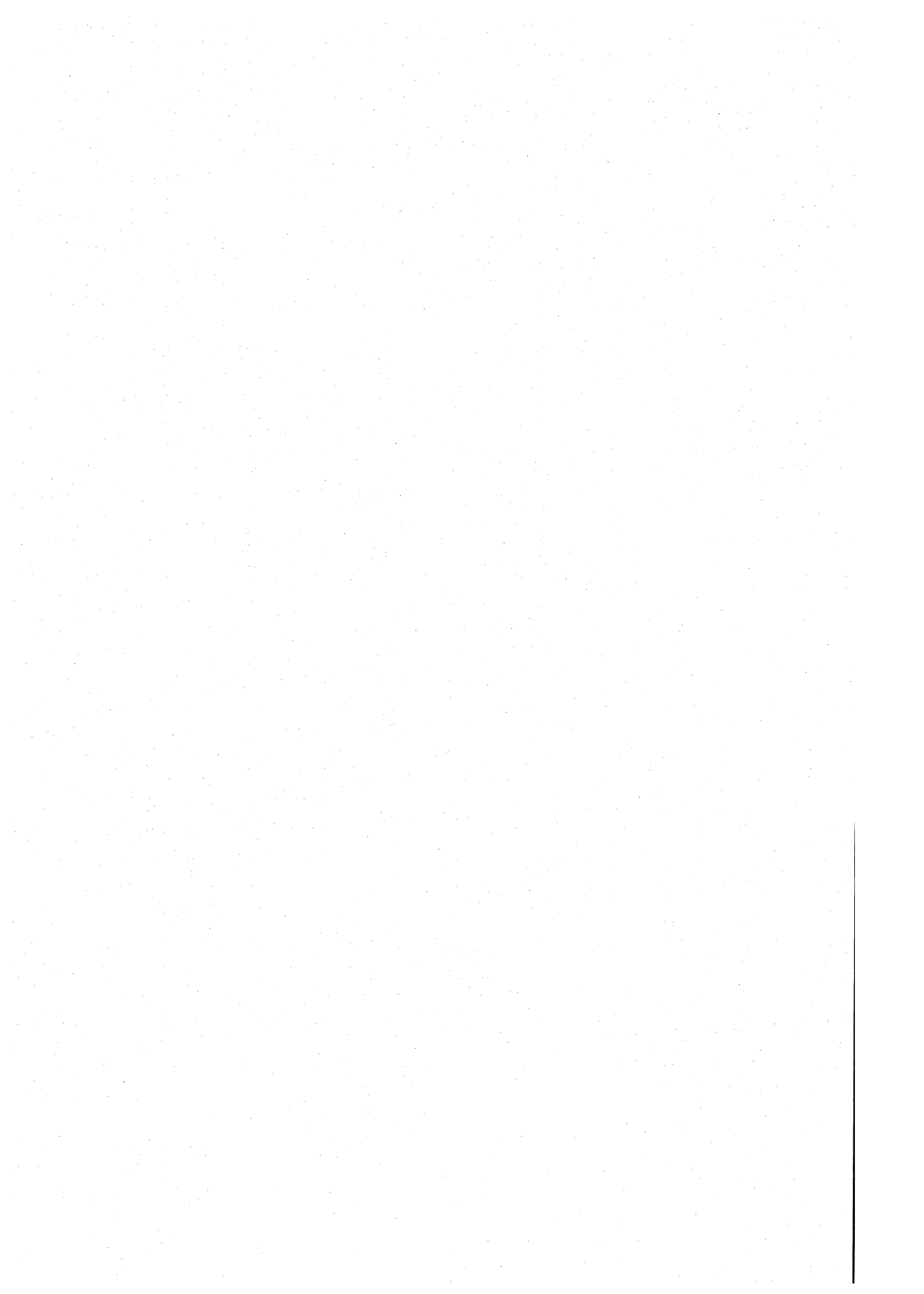
An implementation of the tool, and GIS as such, was made within the parts gender relations and HIV/Aids in the Sida project. Analyses were performed on the currently available data and new collected data. Examples of data used in the analyses are number of women organisations in Kigulu county (Iganga district), population figures, and number of health organisations on district level.

The most important conclusion of our thesis is that GIS can be a useful tool in social sciences but that there is a need of user-friendly tools. The statistical analysis tool is an example of a user-friendly tool that can be used in research within social sciences.



Abbreviations and glossary

ASCII	American Standard Code for Information Interchange, numerical representation of a character
CBO	Community Based Organization
COM	Component Object Model
C++	A computer programming language
ESRI	Company that produces GIS and mapping software, e.g. ArcView
Euclidean distance	The shortest distance between two points
FSS	The Faculty of Social Sciences at Makerere University in Uganda
GIS	Geographical Information System
Grid	A file format that supports raster data
GDP	Gross Domestic Product, an economic index
GPS	Global Positioning System
Macro	A number of recorded commands that can be activated by pressing a key
MFS	Minor Field Study, a student project performed in a developing country, financed by Sida
Mxd-file	A format used for projects in for example ArcView 8
NGO	Non Governmental Organization
Node	The end point of a line
Shapefile	A file format for storing the geometric data of geographic features, used in ESRI products
Sida	Styrelsen för internationellt utvecklingssamarbete, a Swedish organization for international development cooperation
SPSS	A software for statistical analyses
VBA	Visual Basic for Applications, a computer programming language
VB	Visual Basic, a computer programming language
VC++	Visual C++, a computer programming language



Contents

1	INTRODUCTION	1
1.1	BACKGROUND.....	1
1.1.1	<i>The Sida project</i>	1
1.1.2	<i>Geographical statistical analysis tool</i>	1
1.2	PURPOSE.....	2
1.3	METHODOLOGY.....	2
1.3.1	<i>The Swedish phase</i>	2
1.3.2	<i>The Ugandan phase</i>	3
1.4	DELIMITATIONS.....	3
1.5	THESIS ORGANIZATION.....	3
	PART 1 – TECHNICAL FUNDAMENTALS	5
2	GIS AND GIS APPLICATION	5
2.1	GIS.....	5
2.2	ARCGIS.....	5
2.2.1	<i>ArcView</i>	6
2.2.2	<i>Desktop applications</i>	6
2.3	VISUAL BASIC.....	8
2.4	VISUAL BASIC FOR APPLICATIONS.....	10
2.4.1	<i>VBA and ArcObjects</i>	11
3	STATISTICAL ANALYSIS THEORY	13
3.1.1	<i>Standard regression</i>	13
3.1.2	<i>Multiple regression</i>	17
	PART 2 – UGANDA AND GIS IN SOCIAL SCIENCES	20
4	FACTS ABOUT UGANDA	20
4.1	GEOGRAPHY.....	20
4.2	THE UGANDAN PEOPLE.....	21
4.3	POLITICS.....	21
4.4	EDUCATION.....	22
4.5	ECONOMICS.....	22
5	GIS IN SOCIAL SCIENCES	23
	PART 3 – THE STATISTICAL TOOL AND THE MINOR FIELD STUDY	24
6	THE STATISTICAL TOOL	24
6.1	INTRODUCTION.....	24
6.2	DESCRIPTION OF THE STATISTICAL TOOL.....	24
6.2.1	<i>User interface</i>	25
6.3	THE FUNCTIONALITY OF THE TOOL.....	26
6.3.1	<i>New geographical attribute</i>	26
6.3.2	<i>Graph showing relationship between two attributes</i>	27
6.3.3	<i>Multiple regression analysis</i>	28
6.4	TESTING THE STATISTICAL TOOL.....	29
7	THE MINOR FIELD STUDY	30
7.1	EARLIER WORK.....	30
7.2	GIS LABORATORY PREPARATIONS.....	30
7.3	THE GIS WORKSHOP.....	30
7.3.1	<i>Presentation of the statistical tool</i>	31
7.3.2	<i>Problems</i>	31
7.3.3	<i>Discussion with the workshop participants</i>	32
7.4	DATA COLLECTION.....	32

7.5	GIS APPLICATIONS IN SOCIAL SCIENCES.....	33
7.5.1	<i>Application to gender relations.....</i>	33
7.5.2	<i>Application to HIV/Aids.....</i>	37
8	DISCUSSION.....	43
9	CONCLUSIONS	44
10	REFERENCES	45

APPENDIX

- A. Technical specification
- B. Help menus
- C. Exercises
- D. Test results for the geographical statistical analysis tool

1 Introduction

1.1 Background

The idea of our project has its origin in the **Sida** project “Consolidation Peace and Development in the Lake Victoria Region and its Environs: The National and Local Responses to Transformation from Turmoil to a More Sustainable Development Process”, from now on referred to as the Sida project. Ass. Prof. Petter Pilesjö, at the Department of Physical Geography at Lund University, Sweden, is involved in the Sida project. Through further contact with him and Dr. Lars Harrie at the Department of Technology and Society at Lund Institute of Technology, Sweden, the need for a geographical statistical analysis tool in research, especially within developing countries, was realized. This, along with our interests in **GIS** and developing countries, led to the idea of our project.

1.1.1 The Sida project

The Sida project began in autumn 2001 and is a cooperation between Makerere University in Kampala, Uganda, and Gothenburg University and Lund University, Sweden. The overall objective of the Sida project is human resource capacity building, which can contain education for researchers and students and greater possibilities for research. Research, e.g. fieldwork, will be done during the project and will include the collection of both quantitative and qualitative data. The project will be continued until 2004 when a final presentation of the results from the Sida project will take place.

The Sida project focuses on three different problems (FSS, 2001):

1. conflicts and post-conflict conciliation and transformation,
2. public policy, changing gender relations, ideologies and identities, and
3. political economy of diseases in the context of conflict: a study of the HIV/Aids pandemic in the region.

The research within the Sida project also includes analyses with a GIS (Geographical Information System). To enable analyses of the collected data with a GIS, structured digital geographic data are required.

1.1.2 Geographical statistical analysis tool

Statistical analyses are common in social sciences. For example, researchers use statistics to investigate possible relationships between quantities, such as people’s choice of shopping place according to distance, income and prices in the store.

The use of GIS has increased in several fields that partly deal with geographically based data. In many cases researchers are interested in using geographical/statistical data to explain facts in their research. One example might be to study the relationship between income and the number of people with higher education. This can be done visually in GIS by using overlays. Another alternative methodology is to compute

geographical quantities in GIS and then export these quantities to a standard statistical tool (SPSS, Excel, etc.). These statistical tools cannot be integrated into a GIS so that the statistical analyses can be executed from inside e.g. ArcMap. Since such integrated GIS and statistical programs do not really exist and there is a demand for them, it is an interesting task to deal with.

In this project a statistical tool, integrated in a GIS, was used. A first version of the statistical tool was developed by Dr. Lars Harrie, at the Department of Technology and Society at Lund Institute of Technology at the end of 2001. The tool consisted of the functions: "New geographical attribute", "Graph showing relationship between two attributes", and "Multiple regression analysis". This earlier version of the tool has then, in our project, been modified and enlarged.

1.2 Purpose

The overall purpose of our project is to develop and introduce a statistical tool for geographical analyses, as an application to ArcMap, which can be used in research within social sciences. It is important that the tool is user-friendly since many of the researchers within social sciences lack experiences in GIS and related software.

The main purpose of the minor field study is to introduce the statistical tool, and GIS as a whole, in research within social sciences in Uganda, and to evaluate the problems and possibilities concerning the implementation of GIS at the Faculty of Social Sciences (FSS) at Makerere University in Kampala.

1.3 Methodology

The project is divided into two phases, called the Swedish phase and the Ugandan phase. The Swedish phase was performed in Sweden and consisted of development of the statistical tool. The Ugandan phase was performed in Uganda (Kampala and Iganga district) and is the implementation part of the project, i.e. introducing the tool in research at FSS, Makerere University in Kampala.

1.3.1 The Swedish phase

This part of the project was carried out during spring 2002, at the GIS center at Lund University in Sweden. It consisted of further development of a statistical analysis tool as an application to ArcView. The application has been developed in the Visual Basic for Applications environment in ArcMap, and by using the COM-library ArcObjects. Literature treating Visual Basic and further literature in ArcView was studied. To make the tool user-friendly directions from "Software engineering" (Sommerville, 2001) were used. Help menus for each step of the functions were also included in the statistical tool, to make it more user-friendly.

Preparations for the Ugandan phase were also carried out in Lund. For example, we undertook studies of the situation in Uganda to prepare for our stay there. We did also further strengthen our connections with people involved in the project, to make the fieldwork in Uganda easier. These included Dr. Joy C. Kwesiga, our supervisor in Uganda, who helped us with practical details. It also included Dr. Margareta Espling

at Gothenburg University, with whom we performed our fieldwork in Iganga. She helped us with possible parameters for the minor field study that slightly affected our design of the statistical tool. Before going to Uganda a two-day course for the holders of the MFS scholarship was hosted in Sweden.

1.3.2 The Ugandan phase

This part of the project was carried out during two months in autumn 2002. During these two months the analysis tool, developed in the Swedish phase, was installed in the GIS laboratory at the Faculty of Social Sciences at Makerere University.

It was of relevance that the statistical GIS-tool was presented to the researchers in a comprehensible way. The best way to achieve this is to teach the application to the researchers in a face-to-face setting. This was mainly done during a GIS workshop. Exercises and help menus for the tool were also composed to make the statistical tool easier to grasp (see Appendix B and C).

Two minor studies concerning gender relations and HIV/Aids were carried out to show a possible use of the tool. Parameters that were part of the analyses are within the subjects economy, education and health, together with distances. To be able to follow through on these studies several kinds of information was collected. This was done by contacting authorities, participating in interviews, etc. The minor study concerning gender issues was partly performed in Iganga district together with Dr. Margareta Espling.

1.4 Delimitations

The following delimitations have been applied to our work:

The current functionality of the tool is not fully satisfactory. This is mainly due to the time limit of our project.

Our mainly technical background, and the nature of our thesis, made us focus on the more technical matters during our time in Uganda, such as the functionality of the computer lab, helping the researchers using GIS, etc. The minor study was mostly concentrated on the workshop and the problems and possibilities of the implementation of GIS, and not so much on interviewing people or collecting data. Therefore, the social sciences aspects of this report are limited.

1.5 Thesis organization

The thesis is divided into three parts, where the first two parts contain background theory of the third part.

The first part, Technical fundamentals, treats the theory behind the statistical tool, such as regression analysis and ArcGIS. This part is needed to understand the development and the functionality of the tool. Part 1 consists of chapter 2 - GIS and GIS application - and chapter 3 - Statistical analysis theory.

The second part, Uganda and GIS in social sciences, consists of information about Uganda, like economic development and history of the country, and some facts about GIS in social sciences. This gives a background of Uganda and motivates why GIS can be a useful tool in social sciences. Part 2 consists of chapter 4 - Facts about Uganda - and chapter 5 - GIS in social sciences.

The third part, The statistical tool and the minor field study, contains the practical parts of the thesis. Chapter 6 - The statistical tool – gives a description of the functions of the tool, and chapter 7 - The minor field study – treats the implementation of the statistical tool in Uganda.

Words that are of bold type are explained in the abbreviations and glossary list. The words are only marked once throughout the text.

Part 1 – Technical fundamentals

2 GIS and GIS application

This chapter starts with a brief explanation of GIS in general, continues with explaining ArcGIS and Visual Basic and finishes with describing Visual Basic for Applications and ArcObjects.

2.1 GIS

A GIS is a computerized information system for handling and analyzing geographical data (Eklundh, 2000). Geographical data are all types of objects that have a position on the Earth's surface, and include horizontal coordinates (and sometimes height) as well as information about the objects. This information is gathered in tables, which are connected to objects in a geographical database. One of the many advantages with a GIS is that it enables geographical analyses of large databases and maps.

2.2 ArcGIS

ArcGIS is an integrated geographical information system from **ESRI**. The product is a complete system divided into three desktop applications: ArcMap, ArcCatalog, and ArcToolbox. All of the desktop applications can be reached from any of the three software products ArcView, ArcEditor or ArcInfo (see Figure 2.1).

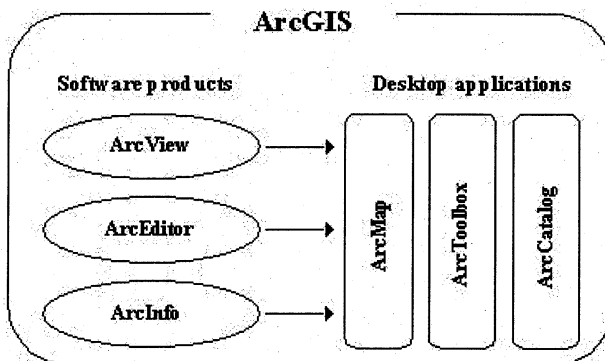


Figure 2.1. Part of the ArcGIS system.

Each software product contains a different level of functionality (more or less advanced) of the desktop applications and is thereby suited for different users. ArcInfo is the most complete product with high access to all desktop applications. ArcView provides high functionality of ArcMap and ArcCatalog and a lighter version of ArcToolbox. ArcInfo includes full functionality of ArcView with the addition of advanced editing capabilities (Booth, 2001).

ArcGIS also contains ArcIMS, and ArcSDE. ArcSDE is a gateway that facilitates managing spatial data in a standard database management system. ArcIMS provides the foundation for distributing GIS and mapping services via the Internet (Esri, 2002a).

In this context ArcView and its components will be presented to give an overview of the part of the ArcGIS system used in our project. The information is based on "Getting started with ArcGIS" (Booth, 2001).

2.2.1 ArcView

ArcView is one of the software products in ArcGIS. It provides comprehensive mapping and visualization tools, analysis tools, and editing and geoprocessing tools (Booth, 2001). The product consists of levels of three desktop applications. These are presented below.

2.2.2 Desktop applications

The three desktop applications ArcMap, ArcToolbox and ArcCatalog in ArcView complement each other. One application can easily be accessed from another.

ArcMap is the central application in ArcGIS. It is an application for viewing, querying, editing, composing and publishing maps (see Figure 2.2) (Booth, 2001). The application contains a Table of contents (to the left in the window) where all the layers and tables of the map are listed. Features with symbols are listed for every layer. The table of contents provides for example the possibility to open the tables to the layers, change symbols and colors in the map, make layers visible or invisible in the map and to join tables. To the right in the window the map is shown. It is possible to click in the map to view the attributes of the selected features. Other functions that ArcMap provides are for example creating buffers and graphs.

ArcMap can be tailored for specific requirements. This is done with the COM-library ArcObject and the programming language Visual Basic for Applications. Details about this are given in subsection 2.4.1.

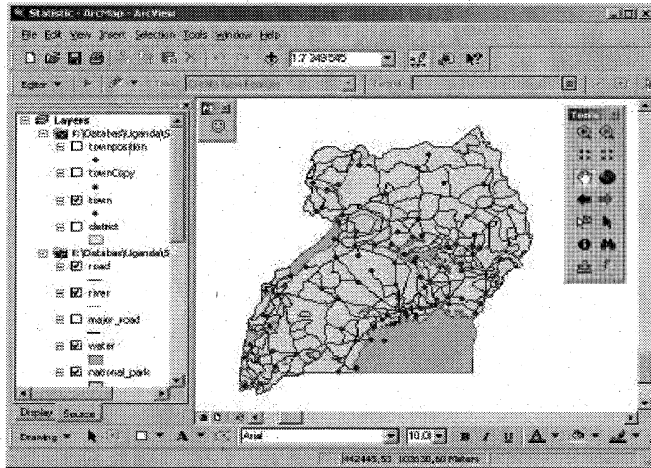


Figure 2.2. A map of Uganda in ArcMap.

ArcCatalog is an application for browsing, organizing, distributing and documenting geographical data sets (see Figure 2.3) (Booth, 2001). The application works as any file catalog with the addition of detailed information of files. ArcCatalog provides the possibility to view metadata such as coordinate system and map projection for a **shapefile**, preview files, create geodatabases, and create new tables.

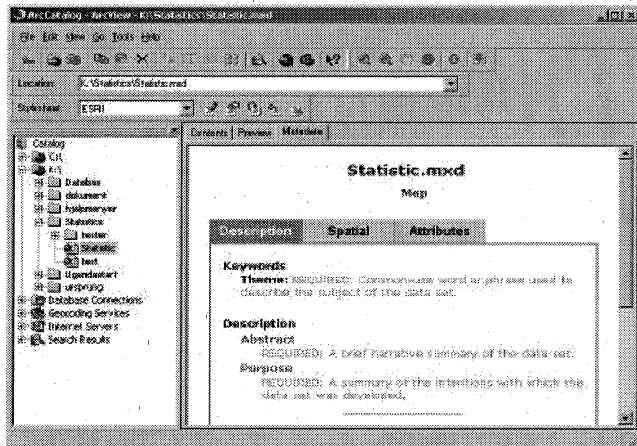


Figure 2.3. ArcCatalog.

ArcToolbox is an application for geoprocessing (Booth, 2001). The application has a catalog tree layout of functions (see Figure 2.4). In ArcToolbox it is possible to convert data between file formats, for example **ASCII** to **Grid** or a geodatabase to shapefile and define projections. Many of the functions that can be performed in ArcToolbox can also be performed in ArcMap. Note that ArcView contains a lighter version of ArcToolbox; the ArcToolbox version accessed from ArcInfo consists of more functions.

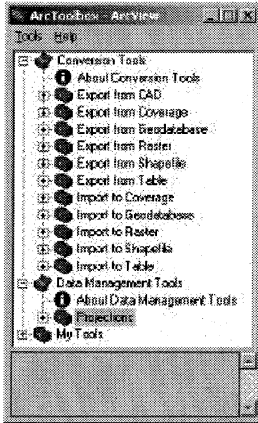


Figure 2.4. ArcToolbox.

2.3 Visual Basic

Microsoft Visual Basic (VB) consists of tools for designing interface and a simplified software programming language (Robin, 1994). Examples of interfaces are menus, texts and graphics. VB is frequently used for interface development in Windows environment, but the programming language is not very suitable for heavy calculations.

VB is built around *objects* (e.g. forms and controls, see Figure 2.5), which have different *properties*, and *methods*. The methods are used to perform actions with the objects. A property is something that characterizes the object, e.g. its name. VB is so-called event-oriented programming, which means that something is executed when the user for example clicks a commandbutton or chooses from a so-called combobox. These buttons and boxes are called controls and are connected to the code (see Figure 2.5).

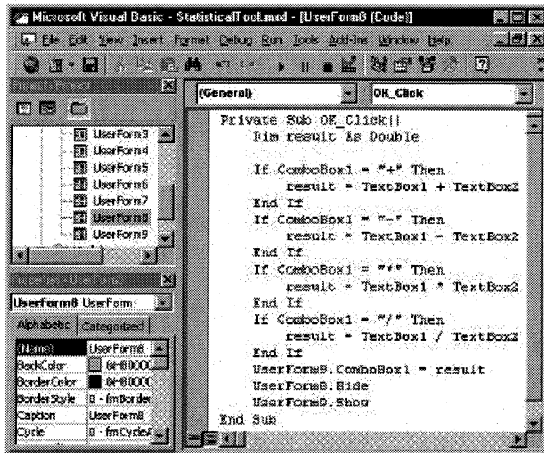
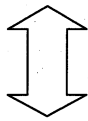
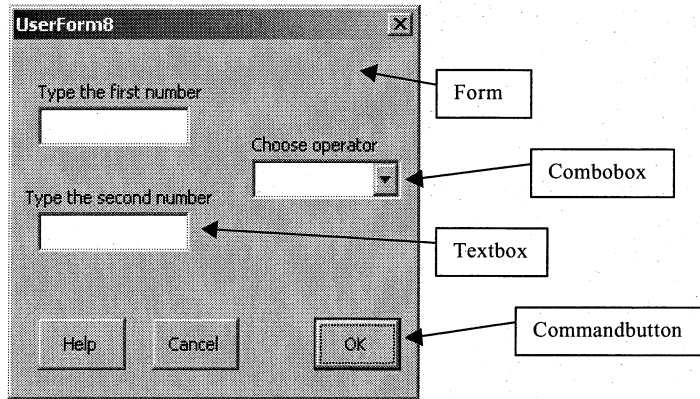


Figure 2.5. The interaction between controls and code.

Below the controls of the form in Figure 2.5 are described:

- Form: A graphic window containing controls and the code that belongs to it (see Figure 2.5).
- Combobox: A box that contains a list of objects to choose from, displayed when pressing the arrow.
- Textbox: A box where text can be filled in using the keyboard.
- Commandbutton: A button connected to code performing an event.

A program in Visual Basic can be built from several forms. The forms are simply connected to each other in the code so that the use of a control in one form displays the next form. When the button OK in the form is clicked, the subroutine OK-Click () is performed and a new form showing the result of the calculation is shown (see Figure 2.5).

Declarations of variables can be done either globally or locally. Variables declared under Declarations are for the whole form. If Global is put in front of the declaration in a module the variables are declared for the whole program. If the variables are declared in a procedure they are local for that very procedure.

Code written in a form cannot be called from other parts of the program. Procedures in a module file are global, i.e. they can be called from any part of the program.

There are three types of procedures:

- 1) Event procedures
- 2) General procedures
- 3) Functions

An event procedure is a procedure linked to an object and an event (see example in Figure 2.5).

A general procedure does not respond to events and must therefore be called from another procedure. A general procedure is advantageous to use if a procedure is long and you want to split the code to make it less hard-to-grasp. Another example of when using a general procedure can be favorable is for parts of code that are used frequently in the program, for example routines for error handling.

A function is a procedure that returns a value and it can therefore be used as a variable in a mathematical operation. If the function $\text{Area}(x,y)$ computes the area of a rectangle with the sides x and y , it can for example be used in the expression: $\text{TotalArea} = \text{Area}(2,2) + \text{Area}(3,4)$, where TotalArea is the name of a variable with the same numerical type as the value of the function Area . A function is advantageous to use e.g. when a complicated mathematical formula is used in several places of the program. This will minimize the errors in the formula since it will exist in only one place.

2.4 Visual Basic for Applications

Microsoft Visual Basic for Applications (VBA) is a development environment that can be embedded into applications (Microsoft, 2002a). **VBA** contains a set of programming tools based on the Microsoft Visual Basic development system and is designed to enable developers to build custom solutions using the full power of Microsoft Visual Basic. When using applications that host VBA, e.g. ArcMap, automation and extension of the application functionality can be done. An example of this is creating tools with new or simplified functions in ArcMap.

Software that include VBA is called customizable applications, which mean applications that can be suited to fit specific business requirements. With VBA, customers can buy software and tailor it to meet a specific requirement, rather than building solutions from scratch.

There are different ways of programming in VBA. Some examples of this are creating a toolbar, creating a **macro**, or using Visual Basic forms inside the VBA environment. VBA is mainly like VB, but the macros can easily be added to a toolbar within an existing program after they are created.

VBA includes the same elements that are familiar to developers using Microsoft Visual Basic, i.e. a Project window, a Properties window and debugging tools.

VBA is not very suitable for more demanding calculations. For this task it is better to use another programming language, e.g. VC++ in combination with VBA.

2.4.1 VBA and ArcObjects

VBA is used in ArcGIS together with the COM-library ArcObjects. ArcObjects is based on Microsoft's Component Object Model (COM) (Esri, 2002b). In the VBA environment it is possible to reach data, such as table data, using ArcObjects. VBA can be reached from e.g. ArcMap through the Visual Basic Editor. The Visual Basic Editor lets you write a VB macro and then debug it right away in ArcMap.

ArcObjects is a large library of COM (Component Object Model) components containing many hundreds of classes (Zeiler, 2001). COM is a software architecture that allows applications to be built from binary software components (Microsoft, 2002b) and is based on object-oriented technology.

ArcObjects is a collection of software components with GIS functionality and programmable interfaces. ArcObject is not a language itself and does not specify how an application should be structured. It only provides the programmer with COM-objects. This library can be reached from any of the ArcGIS products using the built-in VBA. There are also other programming languages that can handle ArcObjects, e.g. VC++ (Esri, 2002c).

The ArcGIS desktop applications are partly built on components from ArcObjects. When developing special functions in VBA it is possible to use many of the functions from the desktop applications. An example is a special function that creates a new field in a table with some sort of information, for example the sum of the numerical data from two other fields in the table. This can of course also be done directly in ArcMap. In this case "field" is a component or object with belonging properties such as name and type. The code below is an example of how a new field is added to a table and how information is stored in this field. Note that the code is not complete.

```
Set pTable = pMap.Layer(num)
Set pField = New Field
Set pFieldEdit = ppField

'Give the new field its properties
With pFieldEdit
    .Length = 40
    .Name = Difference
    .Type = esriFieldTypeDouble
    .Editable = True
End With
'Add the new field to the table
pTable.AddField pField

'Calculate the distances
Do Until ppRow Is Nothing
    A = ppRow.Value(FieldId1)
```

```

B = ppRow.Value(FieldId2)
distance = (A - B)
Store the information in the new field
ppRow.Value(FieldId3) = distance
ppRow.Store
Set ppRow = ppCursor.NextRow
Loop
    
```

The COM-library ArcObjects is visualized in a diagram that shows all the objects, its properties, interfaces and methods. The components are grouped into subsystems. Each subsystem has its own component object model diagram. Examples of sub systems are ArcMap, ArcCatalog, Geodatabase, IMS, and Spatial Analyst extension. Figure 2.6 shows a component from the sub system Geodatabase.

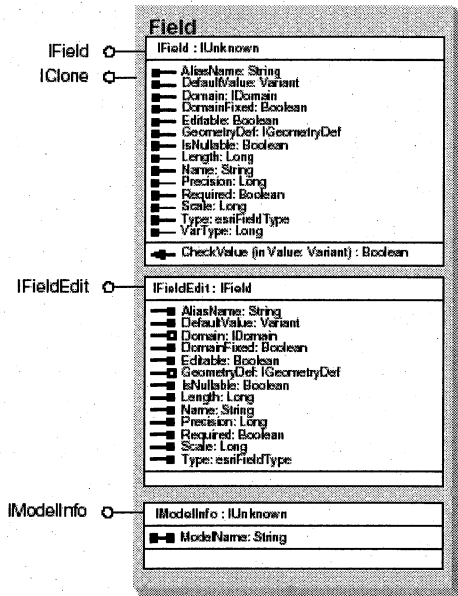


Figure 2.6. The object "Field" in ArcObjects (from ArcObjects developer help).

3 Statistical analysis theory

Regression analysis is a statistical model used for explaining a functional relationship between two or several variables, one dependent variable and one or several independent variable/variables (Haining, 1990). An example of an application of regression analysis in spatial data analysis in social sciences is investigating the possible relationship between illiteracy and distance to a public school. Illiteracy is then the dependent variable and distance is the independent variable. The dependent variable (illiteracy) might as well be explained by several independent variables together, for example the distance to a school and the income of the family. This is performed with a multiple regression.

The standard regression model specifies a functional relationship between the dependent variable (y_i) and the independent variable/variables ($x_{1,i}, x_{2,i}, \dots, x_{k,i}$) (Haining, 1990):

$$y_i = \alpha + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i} + e_i \quad i = 1, \dots, n, \quad (3.1)$$

where

α is the interception of the regression line and the y-axis,

β_1, \dots, β_k are the regression coefficients,

e_i is the error or disturbance term, and

n is the number of values.

The following two sections present the theory of standard regression and multiple regression. The purpose of these sections is to present the theory behind the functions in the statistical tool (see chapter 6). The theory is concentrated on the parameters:

estimation of regression coefficient ($\hat{\beta}$), standard error of regression coefficient

($D(\hat{\beta})$), degree of explanation (R^2), and expectation value of residuals ($\hat{\sigma}$), since these are the parameters calculated in the functions of the statistical tool.

3.1.1 Standard regression

In standard regression, where there are one response variable and one explaining variable, the data set consists of several couples of values $(x_1, y_1), \dots, (x_n, y_n)$ where x_1, \dots, x_n are given quantities and y_1, \dots, y_n are observations depending on the given values x_1, \dots, x_n (Blom, 1989).

One way to examine if there is a relationship between the two variables, i.e. if x explains y , is to plot the values in a graph. If the plotted values seem to create a straight line, then y most likely depends linearly on x . A drawn line fitted to the plotted values is called the theoretical regression line and can be described as (Blom, 1989):

$$y = \alpha' + \beta x, \quad (3.2)$$

where α' is the interception of the regression line and the y-axis.

The regression coefficient β indicates how much y increases as x increases with one unit. It shows the slope of the regression line. If the value of β is zero, y is not dependent on x . The relation between every couple of values can be described as

$$y_i = \alpha' + \beta x_i + e_i, \quad (3.3)$$

where $e \in N(0, \sigma)$.

e_i is the error or disturbance term and represents the vertical divergence between the plotted value and the line (Blom, 1989).

To more precisely decide a regression line, a regression analysis numerical method is needed. If α' is replaced with $\alpha - \beta \bar{x}$ (since this simplifies later expressions), the equation (3.2) is changed into

$$y = \alpha + \beta(x - \bar{x}), \quad (3.4)$$

where

$$\bar{x} = \sum_{i=1}^n x_i / n, \text{ and}$$

α and β are unknown coefficients of the regression line.

α and β can be estimated with help of the least squares method. This method is used because it is simple and gives an unbiased result if the error of disturbance is assumed to be normally distributed. To perform this, the minimum for the following quantity is computed:

$$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - m_i)^2, \quad (3.5)$$

$$\text{where } m_i = \alpha + \beta(x_i - \bar{x}). \quad (3.6)$$

$y_i - m_i$ is the vertical distance e_i from the numerically decided regression line to the plotted points (see Figure 3.1).

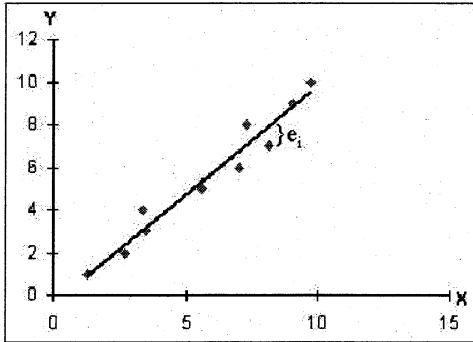


Figure 3.1. Regression line and plotted points. e_i is the error or disturbance term.

To minimize $Q(\alpha, \beta)$ the derivative of the quantity in equation 3.5, with respect to β and α , is set to zero and solved. Simple manipulations then give:

$$\hat{\alpha} = \bar{y}, \text{ and} \quad (3.7)$$

$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}, \quad (3.8)$$

where $\bar{y} = \sum_{i=1}^n y_i / n$.

$\hat{\beta}$ is an estimation of β . $\hat{\beta}$ can be used for explaining the relation between y and x . In the graph below (Figure 3.2) the broken line shows the mean value of y_1, \dots, y_n (\bar{y}) and the other line is the regression line. The plotted values lies closely to the regression line which means that y probably depend on x . The value of $\hat{\beta}$ is positive as the line has a positive slope.

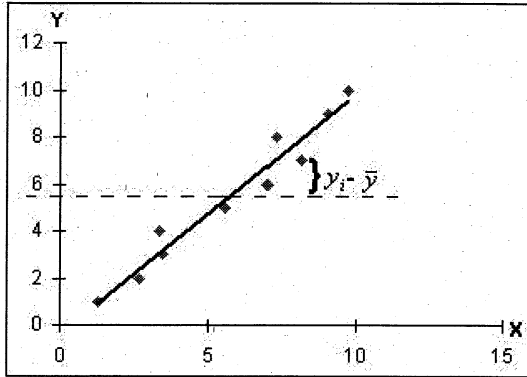


Figure 3.2. Regression line where y depends on x .

The graph below (Figure 3.3) shows a regression where y do not depend on x . y_1, \dots, y_n are almost the same for all the x -values. The mean value coincide with the regression line.

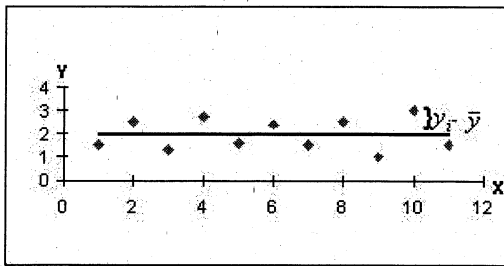


Figure 3.3. Regression line where y does not depend on x .

The standard error of the estimated regression coefficient $D(\hat{\beta})$ is the standard deviation of the estimated regression coefficient. A low value of the standard error means that the regression line is well fitted and that the relation between the dependent variable and the independent variable is strong. If the standard error of the estimated regression coefficient is much less than the estimated regression coefficient there is most likely a relationship. A higher value on the estimated regression coefficient allows a higher value on the standard error.

The standard deviation of the estimated regression coefficient is calculated from the square root of the variance. If the equation 3.8 is written as:

$$\hat{\beta} = \sum c_i y_i, \quad (3.9)$$

$$\text{where } c_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2}, \quad (3.10)$$

the variance is (considering equation 3.9 and 3.10)

$$V(\hat{\beta}) = V\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n c_i^2 \quad (\text{Blom, 1987}). \quad (3.11)$$

Hence, we obtain:

$$V(\hat{\beta}) = \sigma^2 * \left(\frac{1}{\sum (x_i - \bar{x})^2} \right)^2 * \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}. \quad (3.12)$$

The standard deviation is now received from the square root of the variance $V(\hat{\beta})$:

$$D(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}. \quad (3.13)$$

The standard error of $\hat{\beta}$ is small if $(x_i - \bar{x})^2$ is large. This can easily be understood from equation 3.13. If the x-values are widely spread then the fit of the regression line will be more certain. If the values are gathered close together, the fit of the regression line will be uncertain.

3.1.2 Multiple regression

It is often desired to investigate a relationship between more than two variables. This is done with a multiple regression analysis where one attribute can be explained by several others together. Therefore we now extend subsection 3.1.1 to deal with several explanatory variables.

Consider the equation

$$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \dots + \beta_k(x_{ki} - \bar{x}_k) + e_i, \quad (3.14)$$

where $E[e_i e_j] = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$.

This is a multiple regression model where $E[e_i e_j]$ is the expectation operator. β_k is the coefficient for each variable. If β_k equals zero there is no relationship between x_k and y . A high value of β means that y changes a lot, even for small changes in x . The standard error of each coefficient can also be computed. The formula is the same as for standard regression (see equation 3.13)

The constant in the equation 3.14 is the y-value where the estimated regression "line" (strictly speaking not a line since there are more than two dimensions) crosses the y-axis. For example in the equation $y = 0.5 + x$ the constant is 0.5. For equation (3.14)

above, the constant is $\alpha - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \dots - \beta_k \bar{x}_k$. If a constant is not used in the regression this value will be zero and the regression plane will pass through the origin.

The expectation value of the residuals ($\hat{\sigma}$) is a measure of how much the observed values diverge from the values calculated from the regression plane. If this value is zero, there is no divergence at all and the observations coincide with the regression plane. The conclusion is that the smaller this value is the more the observations are gathered round the regression equation. $\hat{\sigma}$ is calculated by:

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{e}_i^2}{n - k - 1}}, \quad (3.15)$$

where

n = the number of observations,

k = the number of independent variables, and

\hat{e} = the estimated error or disturbance term.

The one in the denominator is only used in regression when using a constant (α), because it means the use of an extra parameter.

R^2 is the degree of explanation, i.e. how well the independent variables explain the dependent variable. R^2 is defined as:

$$R^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2}. \quad (3.16)$$

The R^2 value can vary between 0 and 1 for regressions using a constant. If a constant is not used, which is very unusual in social sciences, R^2 can get a value below zero. If R^2 is 1, the independent parameters totally explain the dependent parameter, and if R^2 is 0 they do not explain anything. Also note that when R^2 is 1 $\hat{\sigma}$ is 0, the observations lie on a straight line (Lindgren, 1976). Figure 3.4 illustrates the relationship between R^2 , observations, regression line, and mean value of observations. From looking at the graph it can be realized that the divergence of the observations from the regression line should be smaller than their divergence from the mean value, if the regression model can explain the response variable.

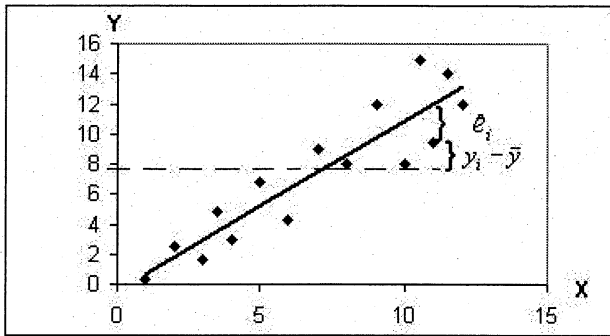


Figure 3.4. R^2 's relationship with observations, regression line and mean value of observations (\bar{y}).

Part 2 – Uganda and GIS in social sciences

4 Facts about Uganda

Uganda is a developing nation situated in eastern Africa with a population of about 25 million people (Uganda Bureau of Statistics, 2002). The country is situated on the equator and has the border countries Sudan, Democratic Republic of Congo, Kenya, Rwanda and Tanzania. The capital of Uganda is Kampala which has approximately 900 000 inhabitants (UI, 2002).

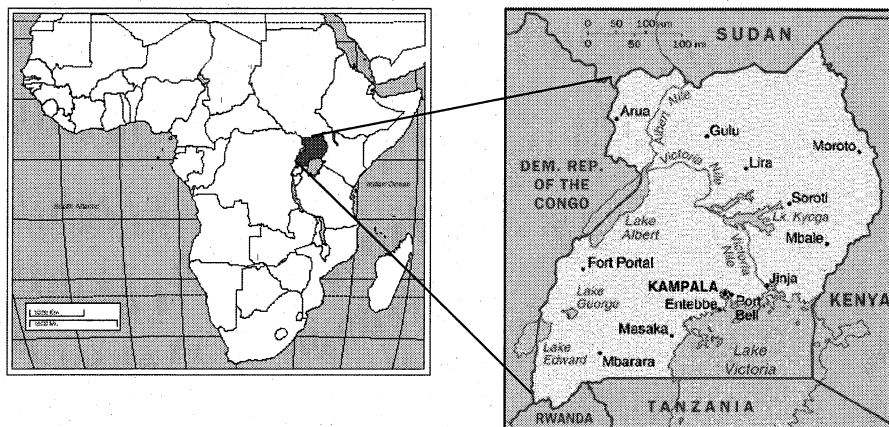


Figure 4.1. Map of Uganda (left: unomaha, 2002; right: CIA, 2002).

4.1 Geography

Uganda has a total area of 241,000 sq km, about half the size of Sweden, and is divided into 56 districts. The country is very fertile and about a fifth of the country consists of lakes. The sources of the Nile can be found in Uganda (Leggett, 2001).

The country is situated on the large East African plateau, approximately 1000 m above sea level, and with mountain masses like Mt. Elgon in the East, and Ruwenzori in the West (Leggett, 2001). The highest point of the country is found on Mount Stanley with 5,110 m above sea level (CIA, 2002a). Forests cover about 8 % of Uganda's land area but have started to shrink due to increased agricultural use of the land and the felling of trees for fuel (SAS, 2002).

Uganda has an equatorial climate. There are two dry seasons, from December to February and from June to August. The other times of the year are wet seasons. In the northeast, the climate is semi-arid (SAS, 2002).

The natural resources in Uganda are copper, cobalt, hydropower, limestone, salt and arable land (CIA, 2002a).

4.2 The Ugandan People

The population of Uganda has increased from 5.9 million 1950 to the estimated 25 million today (SAS, 2002). Over 80 percent of the Ugandan people live in villages and small trading centers (Leggett, 2001). Only 11 % of the population lives in urban areas and 40 % of these live in Kampala (UI, 2002). More than half of the population is between 0 and 14 years old and only 2 % are over 65 (CIA, 2002b). The population density is 85 inhabitants per km² (UI, 2002). Every woman gives birth to 6.8 children on an average (UI, 2002). The average life expectancy is 40.4 years for women and 38.9 years for men (Leggett, 2001).

There are several ethnic groups living in Uganda. The largest is Baganda with 17 % of the population. The religions in Uganda are Roman Catholic (33 %), Protestant (33 %), Muslim (16 %) and local religions (18 %) (UI, 2002).

The languages spoken are English, Luganda (one type of Bantu languages), other Niger-Congo languages, Nilo-Saharan languages, Swahili and Arabic. English is the official national language and is taught in grade schools, used in courts of law and by most newspapers and some radio broadcasts. The Bantu languages are the most widely used of the Niger-Congo languages. They are preferred for native language publications in the capital and may be taught in schools. The Bantu-speaking majority lives in the central, southern and western parts of the country. (CIA, 2002a)

4.3 Politics

Uganda obtained independence from the British colonization in 1962 (BBL, 2000). After the time of colonization the country has been colored by the dictatorial regime of Idi Amin (1971-79), who was responsible for the deaths of about 300,000 opponents. During the presidency of Milton Obote (1980-85) another 100,000 lives were claimed from guerrilla war and human rights abuses. When Obote was elected as President the politician Yoweri Museveni rejected the results and started a guerrilla war (Leggett, 2001). His army was known as the National Resistance Army (NRA) and in 1986 the army captured Kampala and Museveni became the President of Uganda (Leggett, 2001).

The government of the 1990's announced non-party presidential and legislative elections. The President is both chief of state and head of government. Uganda has presidential elections every fifth year and the last election was held 12 March 2001 (CIA, 2002a).

The political development has involved an increased spread of democracy. This has led to a stronger position for the Parliament, decentralization, school reforms, and an increased insight of the press and the general public into democratic rights (Sida,

2001). Although Uganda is a progressive developing country it still suffers from many problems and far from everything in Uganda has developed in a positive direction. Military commitment in Democratic Republic of Congo, extended armed conflict in the northern and western parts of the country, and ravages of the army in the rebel areas are examples of the government's difficulty to bring about a durable peace in these areas (Sida, 2001).

4.4 Education

The adult literacy rate in Uganda is 64 % (Leggett, 2001). Formal education in Uganda starts with seven years of primary school. There is a possibility that primary school will be eight or nine years in the future because resources from secondary and post-secondary education have been moved to primary education. In 1998 only 29 % of primary school children were girls, but there has been a wish for gender equality in primary education (SAS, 2002).

Makerere University, founded in 1922, stands for 95 % of all post-secondary education in Uganda (SAS, 2002). It has over 20 000 students of which 75 % are fee paying (Leggett, 2001). Except for the Government-owned Makerere University and three training and vocational colleges, there are three private universities in the country (SAS, 2002.)

4.5 Economics

Uganda is one of the poorest countries in the world. There have been many ups and downs in the economy since the independence in 1962. The times of civil war and dictatorship totally destroyed the economy. The country has also suffered from large inflations when trying to build up the country (SAS, 2002). Still, the economy in Uganda has improved within the last 15 years. Uganda has pursued economic policy reforms that have imposed fiscal discipline, restructuring public expenditure and liberalization of the economy. Because of the cautious macro economic policies, Uganda has recorded an impressive economic performance over the last decade with average real rate of annual growth in **GDP** recorded at 6.9 percent (MyUganda, 2002). In spite of the progress, Uganda is still highly dependent on foreign aid. The country receives \$1,4 billion (2002) in foreign aid every year (CIA, 2002a).

Agriculture has been Uganda's predominant economic activity since pre-colonial times. Over half of Uganda's economic earnings are derived from coffee exports (MyUganda, 2002). Other important exports are fish, tea, tobacco, cotton and cut flowers (CIA, 2002a).

The currency in Uganda is Ugandan shillings, with a rate of about 1800 shillings per US dollar (in 2002).

5 GIS in social sciences

The economical and political development in Uganda over the last few years has been progressive (Sida, 2001). It is important to analyze changes caused by this progress to learn from the positive development (FSS, 2001). It is also of great interest to discover and analyze spatial patterns to see what needs the country has. Examples of subjects to be studied are the spread of HIV/Aids, conflict changes, and changes in gender related issues.

In these types of analyses GIS can be a very useful tool for many reasons. The main benefits connected to using GIS in such analyses, particularly in the Sida project, are (FSS, 2001):

- improved possibilities to share large amounts of data of common interest between different analyses (e.g. digital maps, infrastructure, income and population),
- improved access, organization and storage of data in a common database,
- possibilities to discover and visualize spatial patterns (e.g. how income varies geographically),
- possibilities to analyze if these patterns are related to other factors (such as conflict zones), and
- new ways to visualize and analyze qualitative data, by converting this data into quantitative data, by the use of GIS.

In spite of all the advantages connected to the use of GIS in social sciences, there are a number of problems concerning the implementation of GIS. It has been stated that the main obstacles to the implementation are (Pilesjö, 2001):

- lack of knowledge/training,
- non-available or expensive data, and
- hardware and software costs.

Even if most of these problems are related to money, some actions can be suggested to facilitate the implementation (Pilesjö, 2001):

- At least one person responsible for the GIS environment (hardware, software, updating, meta data, backup, further education etc.) is needed.
- There is a need of education within GIS and technical skills, such as computer knowledge, and cartography, etc.
- The fastest computers and the latest software are not always needed.

Generally, it can be said that training is of highest importance for the implementation of GIS in social sciences.

Part 3 – The statistical tool and the minor field study

6 The statistical tool

This chapter describes the design, functionality and test of the statistical analysis tool.

6.1 Introduction

The statistical tool is a geographical analysis tool produced for analyses of geometrical data and tabular data together. It is built as an application to the ESRI GIS component ArcMap using VBA (see section 2.2 and 2.4). The tool has been developed at Lund University within the Sida project (see subsection 1.1.1) and is produced for social sciences research, in e.g. developing countries. The purpose is to enable analyses between geometrical data and other data and to create new geometrical data in a simple way. The tool is placed in a GIS to enable easy access to databases containing information in for example district level as well as geographical attributes. In the following text the statistical tool will be referred to as "the tool".

A very important part of developing the tool is to make it user-friendly for people that have not much experience in statistics or in working with a GIS. For this task, VBA is suitable since it is easy to learn and enables development of user-friendly applications.

The tool is partly tailored for data within the Sida project, but only some adjustments are required to make it general.

6.2 Description of the statistical tool

The tool can be reached by opening the project (**mx-d-file**) containing the tool in ArcMap. To access the statistical tool a certain button, called My own tool, is used (see Figure 6.1).

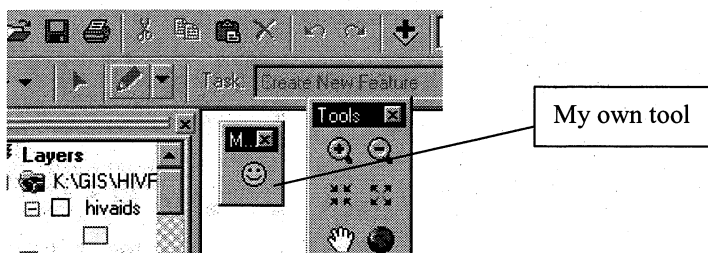


Figure 6.1. The "My own tool" button in ArcMap that accesses the statistical tool.

The opened project must consist of paths to the tables in the database. The tables used in the statistical tool must be stored as shapefiles.

The tool consists of three main functions:

- create new geographical attribute,
- graph showing relationship between two attributes, and
- multiple regression (including standard regression).

These are presented in subsections 6.3.1- 6.3.3. The methods used for analyzing data are different regression analyses. The function that creates new geographical data supports the user with the new information for the analyses. This function can be enlarged with more types of geographical attributes.

An example of the procedure of an analysis could be:

1. Calculate the distance between several villages and a nearby hospital with the function called "New geographical attribute".
2. Examine if there is any relationship between the new attribute (from step one) and the number of people dying from a certain disease in every village. This can be done with the function called "Graph showing relationship between two attributes".

Every step of the functions is connected to help menus. These describe what the different functions can do, how to perform them and how to understand the results. For more detailed information about the tool see the technical specification (Appendix A) and the help menus (Appendix B).

Although the statistical tool to a great extent is general, some functions have been tailored for a specific purpose. The following description of the tool is not general, but specific for the current design of the tool.

6.2.1 User interface

The user interface is the part of a software program that is visible for the user, i.e. not the code. The user interface of the statistical tool is ordinary forms with command buttons, comboboxes and textboxes that lead the user through the program.

The tool has a graphical interface, which is easy for the user to understand. Recommendations of how to make the tool user friendly have been considered during the development. The recommendations are taken from the book "Software Engineering" (Sommerville, 2001). Examples of how the recommendations have affected the design are:

- The forms have similar design.
- The user can always cancel an operation or go to a help menu.
- The forms have little and simple text to describe to the user what to fill in or what to select.
- If the user needs more detailed information he/she is requested to use the help menus.

- After having completed an operation, a message is shown with information of where to find the result.
- Terms have been used in a consequent way.

In spite of this, there are some details in the tool that could be disturbing for the user. These are:

- The tool blocks the rest of the functions in ArcMap. The user can for example not open a table when the tool is in use. The help menus and the graph also block ArcMap functions.
- The help menu and the graph window can only be closed by pressing the cross in the upper right corner.
- When the graph window is maximized and then restored the whole graph is not shown in the restored window.

These problems are not solved due to the time limit of the project and restrictions in VBA and ArcObjects.

6.3 The functionality of the tool

Below the functionality of the statistical tool is described. Three different functions can be chosen from the main menu (see Figure 6.2).

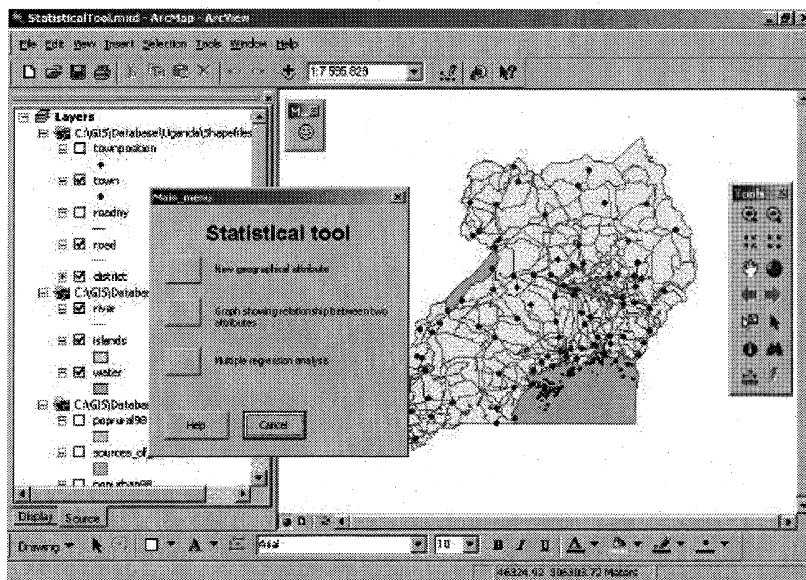


Figure 6.2. The main menu of the statistical tool.

6.3.1 New geographical attribute

This function consists of two sub functions:

1. Compute the shortest distance between geometric objects and
2. Compute traveling distance between cities.

The results of the functions are new geographical attributes added to one of the tables used in the function.

Compute the shortest distance between geometric objects

The first function computes the **Euclidean distance** between one point object and several other point objects. The user first selects one table with point attributes and then one object to where the distances should be computed (possibly from another table). The results are stored in a new field in the selected table as shown in figure 6.3 below. The distances are computed using the theorem of Pythagoras.

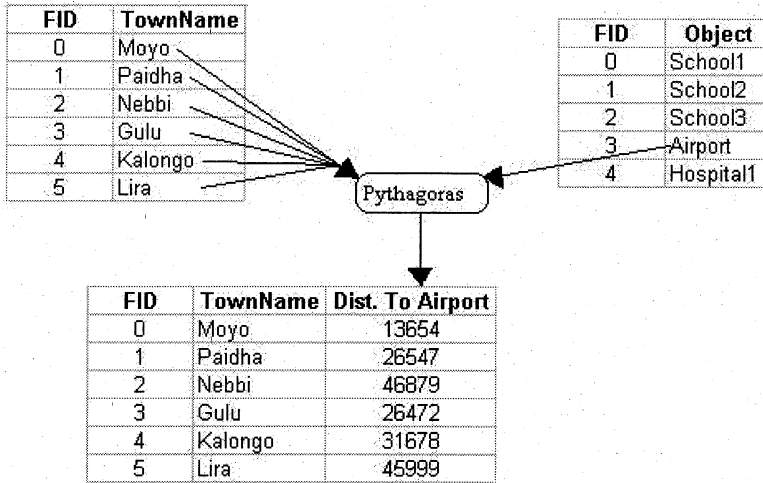


Figure 6.3. The principle of the shortest distance functionality.

Compute traveling distance between cities

The second function computes the traveling distances from one arbitrary city, in this case in Uganda, to other cities (in Uganda). The traveling distance is the nearest distance if a network in a map is followed. The distance is computed with Dijkstras algorithm (Dijkstra 1959, as described in Worboys, 1993). Today, this function is locked to specific tables with specific information on Uganda. The only selection the user has to do is to choose to which city the distances should be computed. The selected city can be any of the cities in Uganda (in the table "town"). The results are shown in a new field in the table containing the district cities. The function also uses the table "townposition". This table is not a shape file and only consists of city names and their node numbers. The table is created by hand inside ArcMap specifically for the traveling distance function. So is the table "road" which is also used when computing the distances. This table contains the length of the road segments, which are split at each city. The splitting of the roads had to be done to create nodes and a useful network.

6.3.2 Graph showing relationship between two attributes

This function shows if there is any linear relationship between two attributes. A standard regression (see subsection 3.1.1) between two attributes is performed to investigate the relationship. The result is presented as a graph where two quantities of

attributes are plotted (see Figure 6.4). The two parameters regression coefficient and standard error of regression coefficient are also calculated and presented together with the graph.

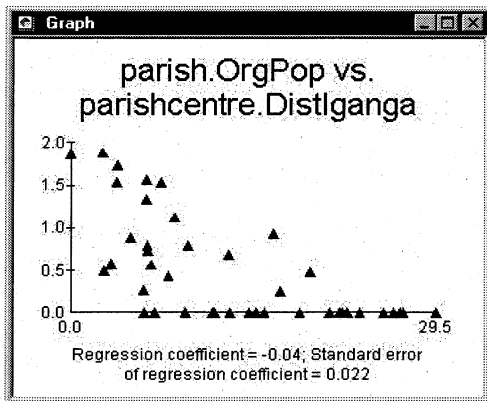


Figure 6.4. The result of the graph functionality. The graph shows the relationship between the number of women organizations (y-axis) and the distance to Iganga town for each parish in Kigulu county (x-axis) (see chapter 7).

6.3.3 Multiple regression analysis

This function shows if there is any relationship between one dependent variable and several independent variables by performing a multiple regression.

The user selects the table containing the attributes, the dependent variable and the independent variables. At least one and no more than six explanatory variables have to be selected. It is also possible to choose if the regression is to be performed with or without a constant.

The result is shown as the parameters degree of explanation (R^2), expectation value of the residuals ($\hat{\sigma}$), regression coefficients and standard error of the regression coefficients (see Figure 6.5). R^2 shows how well the independent variables explain the dependent variable and is defined in equation 2.16. R^2 can vary between 0 and 1, where 1 means that the independent parameters totally explain the dependent parameter, and 0 means that they do not explain anything. $\hat{\sigma}$ is a measure of how much the observed values diverge from the values calculated from the regression plane and is defined in equation 2.15. The smaller this value is the more the observations are gathered round the regression equation (see chapter 3).

Result Multiple Regression

Statistics of the multiple regression analysis

	Name	Coefficient	Standard error
Dependent parameter	GDP_INDEX		
Independent parameter 1	LITERACY 1990	0.01	0.00
Independent parameter 2	Distance to Kampala	0.50	0.00
Independent parameter 3			
Independent parameter 4			
Independent parameter 5			
Independent parameter 6			
Constant		0.00	0.00

R-square	0.14	OK
F2	0.12	
adjusted value of residuals	0.07	OK

Figure 6.5. The result table of multiple regression

6.4 Testing the statistical tool

Before using the statistical tool in research it is important that it is reliable. The tool has to give, to the fullest possible extent, correct results and should not contain any misleading or ambiguous information. It is also important that the tool consists of functions that are relevant for research within social sciences, and that instructions and help menus are correct and on a suitable level for the users. To assure that all these requirements are fulfilled different tests and inspections are executed. Some of the tests are gathered in appendix D.

7 The minor field study

A minor project was performed in Kampala district and in Iganga district in Uganda. The purpose of this project was to evaluate, demonstrate and teach how to use GIS and the statistical tool in research within social sciences. A GIS workshop was held at FSS at Makerere University during one week, where the participants learned the basics of the GIS software ArcView 3.2 and the statistical tool in ArcView 8.1, but also how to add and use their own data in a GIS. Data concerning women organizations were collected on parish level in Iganga district and data concerning HIV/Aids on district level for Uganda as a whole. Statistical analyses as well as several map visualizations were performed on the data in a GIS. During the workshop several problems and obstacles were identified and they are presented later in this chapter. The needs and problems related to data collection are also presented in this chapter.

7.1 Earlier work

During autumn 2001, a database containing new and old data was built by two MFS students from the Department of Physical Geography at Lund University. This database consists of both national data as well as district data. During the same period a GIS laboratory was built at Makerere University. This laboratory was built for research within the Sida project. To facilitate the analyses, researchers at the university are given courses in GIS. A first GIS workshop was held in January 2002.

7.2 GIS laboratory preparations

The first two weeks at Makerere University were spent performing preparatory work at the computer laboratory including updating and organizing of the geographical database. Installations were made of the GIS software ArcView 3.2 and ArcView 8.1, including the statistical tool.

The main obstacles during the preparations were connected to computer access, such as complicated logins and a crowded computer laboratory.

7.3 The GIS workshop

A GIS workshop was held during one week in October at FSS. The workshop was part of the GIS capacity building within the Sida project. Earlier this year a three-day GIS workshop was held at FSS but the October workshop was not a continuation of this workshop as the participants were not all the same. The October workshop was an introduction course in GIS and consisted of both theoretical and practical parts. The theoretical parts were held by Petter Pilesjö, except for the theory concerning the statistical tool, which was held by us. The practical parts consisted of exercises and were also prepared and led by us. The persons who attended the course six times out of eight were rewarded with a GIS certificate from Lund University.

There were about eight participants in the workshop and most of them were teachers and research assistants. Some of them had participated in the earlier course and the computer skills were varying. The participants showed a great interest in getting GIS skills, which made the workshop successful. Some of the participants even continued with the exercises weeks after the workshop to get more GIS experience.

After the closing of the GIS workshop, preparations in form of installments and upgrading were done in a smaller computer laboratory that is planned to be the GIS laboratory for researchers and teachers. Time was also spent on helping participants from the workshop that wanted to finish their GIS exercises.

7.3.1 Presentation of the statistical tool

The last day of the workshop we presented the statistical tool. During the presentation we tried to explain the statistical theory in a comprehensive way and also how to use the statistical tool. The presentation was followed by an exercise on the tool that was held during the afternoon (see Appendix C).

The presentation went well; people seemed interested and posed many questions. As we had expected, there were problems with fully understanding the statistical theory as some of the participants had little knowledge in mathematical statistics, but they seemed very eager to understand. The participants thought that the tool could be useful if there were any relevant data to analyze.

It was easy to follow the exercise since it is designed in a simple way and the participants seemed to manage without greater problems. In the exercise the participants were encouraged to use the help menus when needed. Most of them used the help menus and seemed to understand their contents.

7.3.2 Problems

Preconceived notions

The fairly low number of participants at the workshop was partly because some people from the staff at the faculty were reluctant to join the GIS workshop. The week before the GIS workshop we tried to find out what the staff thought about working with GIS and what their experiences from the earlier workshop were. Many of them were afraid that working with GIS would be difficult for them and some did not understand how GIS could be useful in research within social sciences. A few people expressed that the recent database was not relevant for their research. We had not succeeded in informing the users that they were supposed to add their own research data to the database in the future. They also thought that it was a problem that the last workshop had not been followed up afterwards and that they had forgotten what they had learned in that course.

Planning

Unfortunately the workshop was held during the first week of the semester, which meant that the staff at the faculty had other duties that prevented them from participating in the course. There were 25 persons signed up for the course but only about eight of them attended. This was a result of bad planning since the workshop

was planned in May the same year. The first day of the workshop was cancelled because only a few people appeared.

Lack of a relevant database

The GIS database that was used in the workshop is at the moment quite limited. The data is based on the districts in Uganda and therefore it was only possible to make analyses and visualize data for the entire Uganda on district level. The participants want to be able to make analyses and visualize data on lower levels as they often do research within a specific district. The workshop could not present as interesting examples of analyses in GIS on the current data as wanted. Better examples are very important to create comprehension of how GIS can be used in research.

Computer skills

The computer skills of the participants were sometimes low which prevented them from learning GIS in an effective way, since much time was spent learning basic computer usage.

Lack of a responsible GIS person

The faculty needs a person to be in charge of the GIS. This person should be someone with good computer and GIS skills and with a genuine interest in GIS. The person in charge should be available for other people who need help when working with GIS. The faculty seems to have a problem with finding this kind of person. One person was selected after the first GIS workshop to be in charge of the GIS. He also had an opportunity to go to Sweden to be trained in GIS, but he never came because of a misunderstanding. This person never attended the GIS workshop in October either.

7.3.3 Discussion with the workshop participants

After the course a small discussion was held with the participants to get suggestions for improvements for the next workshop. One suggestion was that next workshop should contain all the parts of a GIS analysis. The participants wanted to be able to follow the analysis from the collection of data in field with a **GPS** to the result. Some also wanted to learn how to add sound and pictures to their maps. The participant wanted to be able to be prepared for the exercises by getting handouts one day ahead. Generally they thought that the certificate was a good idea to motivate people to attend the GIS workshop. The most important thing was to be able to make analyses for a specific district, which is more applicable in their own research.

7.4 Data collection for research

There is a problem with collecting data, as statistics can be hard to get access to. The authorities keeping the data may be reluctant to expose them, and anyway, the process of getting information from an authority is often long. Another problem concerning data collection is that statistical data sometimes do not exist at all.

Maps and data on lower levels than district level (e.g. county or parish) need to be collected, as research often is performed within one district. Hence, there is a problem with digital maps as paper maps need to be digitized. For this task, equipment and skilled persons are required. The database also needs to be updated with the new

districts in Uganda. Another reason for expanding the database is to let the users realize the advantages of GIS by demonstrating relevant examples.

As the researchers collect their own data they need to know what kind of data can be used in a GIS and in what shape. Data in social sciences are often qualitative. If analyses in GIS are to be done on these data they need to be converted into quantitative data or visualized in a correct way.

To be able to collect and expand the database with data that are possible to use in GIS analyses, some basic database knowledge is required. An understanding of the structure of data storage is needed; that there is only room for one piece of information in each attribute connected to a certain object. For example, a table containing all cities in a country and information about how many restaurants there are in each city cannot also contain detailed information about the specific restaurants. This must be separated into different tables since the information is on different levels. The researchers might see it as limiting not being able to analyze very detailed data on a high level (e.g. a whole nation) at the same time.

7.5 GIS applications in social sciences

To demonstrate the use of GIS and the statistical tool within social sciences, a number of analyses have been performed. The analyses are within two research areas in the Sida project: gender relations and HIV/Aids. The GIS application to HIV/Aids demonstrates how GIS analyses can be carried out with data on district level. In this case data from all districts in Uganda were collected. The gender application shows how analyses can be carried out on a lower level, in this case within a county. For this application Iganga district was chosen and data on parish level were collected.

7.5.1 Application to gender relations

The aim of this field study is to demonstrate analyses in GIS and the statistical tool on a lower level than district level. In this case the application was carried out within a county (Kigulu) and information about women organizations on parish level in Iganga district (see Figure 7.1) was collected.

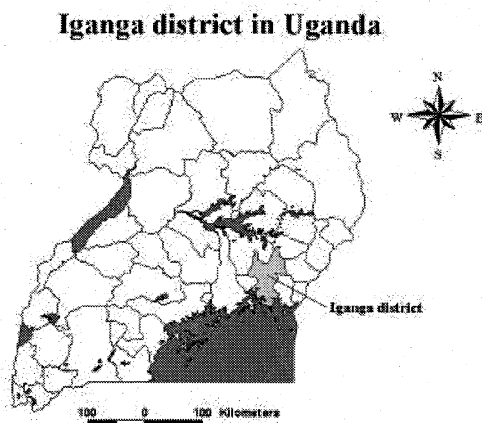


Figure 7.1. The placement of Iganga district in Uganda

Iganga and women organizations

There are many women organizations in Uganda and in Iganga district. They mainly have the aim to improve women's situation. The organizations generally work with issues like economic empowerment, health, education, etc.

Women organizations in Uganda consist of NGO:s (Non Governmental Organizations) and CBO:s (Community Based Organizations). In some cases they are well established with an office and several employees. These organizations work with issues like education. Other organizations consist of for example the women in a village. In these kinds of organizations the members often perform common activities such as brick making or selling handicraft to improve their economical situation.

Iganga is one of 56 districts in Uganda. The district is situated in the east of Uganda close to Lake Victoria. It has slightly more than one million inhabitants. The districts of Uganda are divided into different levels: county, sub county, parish and community. We have chosen Iganga district for our study because research on gender issues within the Sida project is performed in that district. Kigulu county contained the majority of the organizations and was therefore chosen for our analysis.

Collection of data

A fieldwork was carried out during one week in Iganga district together with Dr. Margareta Espling from Gothenburg University. The purpose of our fieldwork was to collect information about women organizations in the district. Data were collected by participating in interviews and copying data at different authorities.

Through participation in interviews of organization staff, together with Dr. Espling, we created a rough image of what kinds of organizations there are in Iganga and how they work. Through the interviews we also understood how organizations register and where we could get more information.

NGO:s and CBO:s have the possibility to register at the Community Development Office at the District Administration. If they register it is possible for them to get contributions from the government. Because of this, most of the established organizations can probably be found in the registers. The registers consist of all kinds of NGO:s and CBO:s and therefore the women organizations needed to be sorted out. The registers contain information about the organizations' names, when they registered, their location, how many members they have, and what their activities are.

At the Iganga District Planning Unit we found information about female and male population in all parishes in Iganga. The population data is from year 2002 and is a projection from 1991 population census by Iganga District Planning Unit. Data for one of the parishes are missing.

Dr.Espling also helped us find detailed maps from 1998 over the counties in Iganga, with sub county and parish borders, at The Department of Surveys and Mapping in Entebbe.

Performance of statistical analyses and visualizations

We have chosen to perform the GIS application within the county Kigulu in Iganga district. This county is situated in the middle of Iganga and the district town (Iganga town) can be found in Kigulu.

The map over Kigulu county was digitized in ArcMap and the digital map consists of the parish borders. Information about women organizations and population for Kigulu was stored in tables. These tables were then connected to the digital map. As the distance function in the statistical tool compute distances between point objects we also created a point layer containing points near the center of each parish.

With the function "Graph showing relationship between two attributes" in the statistical tool (see subsection 6.3.2) we investigated statistical relationships in the collected data. The parameters we used were:

- the number of women organizations standardized with population in each parish,
- the number of members in women organizations standardized with population in each parish, and
- the distance to Iganga town from each parish.

The figures used in the analyses are scaled to be in the same size. Also note that the distances to Iganga town from each parish are computed from an estimated point near the centre of each parish. The results from the analyses are shown in Figure 7.2 and Figure 7.3.

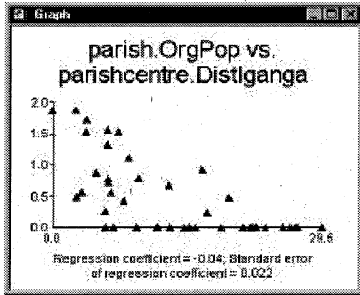


Figure 7.2. A Graph showing the relationship between the number of organizations (y-axis) and distance to Iganga town (x-axis).

parish.OrgPop = number of organizations divided by population in each parish and multiplied by 1000.

parishcentre.DistIganga = distance from all parishes to Iganga town divided by 1000.

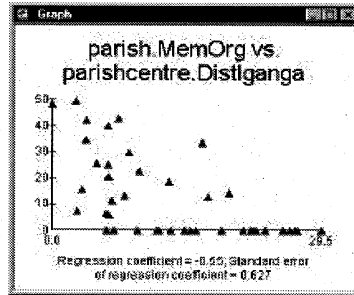


Figure 7.3. A Graph showing the relationship between the number of members in women organizations (y-axis) and distance to Iganga town (x-axis).

parish.MemOrg = number of members in women organizations divided by population in each parish and multiplied by 1000.

parishcentre.DistIganga = distance from all parishes to Iganga town divided by 1000.

In Figure 7.2 a relationship between the two parameters can be suspected. This would mean that the number of organizations in each parish in Kigulu decreases with the distance to Iganga Town. The graph in Figure 7.3 does not show any strong relationship, as the error of the regression coefficient is even larger than the regression coefficient. Some of the values in the graphs are placed on the x-axis. These values correspond to the parishes with no organizations. The value placed on the y-axis is the distance from Iganga town to Iganga town, which is zero.

GIS makes it possible to visualize data with graduated colors, charts etc. Figure 7.4 shows an example of how data can be visualized. This visualization on parish level shows that parishes with many organizations are concentrated to the area around Iganga town. This agrees with the result in the graph in Figure 7.2.

Women organizations in Kigulu county, Iganga 2002

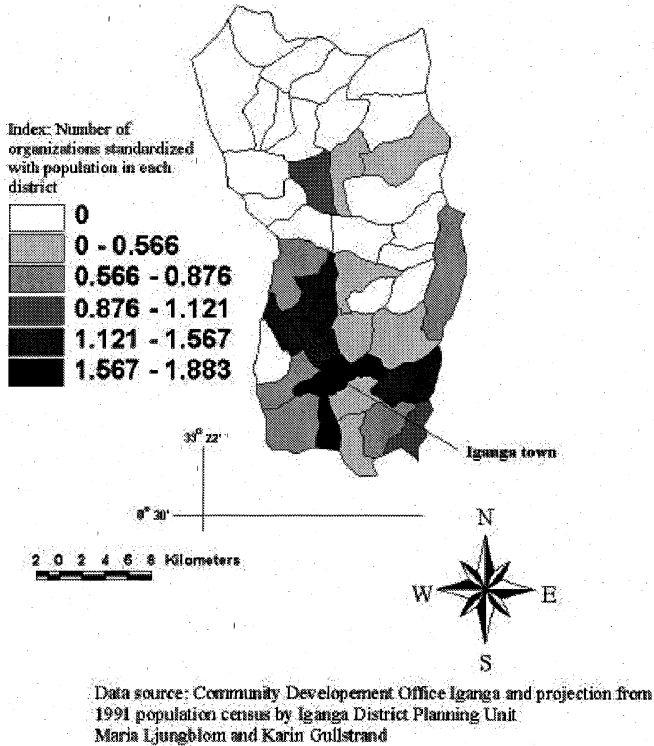


Figure 7.4. Layout from ArcView 3.2 showing number of women organizations standardized with population on parish level in Kigulu county.

Problems

Throughout the study a few problems occurred. Most of them were related to the digitizing of the map and consisted of:

- There is no coordinate system or map projection stated on our paper map. Therefore we did not get georeferenced coordinates (although this does not affect the analyses since we used the correct scale).
- The names of the communities, parishes, etc. and the placement of the borders are not always the same on the map as in other documents, which makes the data connection to the map sometimes difficult.

7.5.2 Application to HIV/Aids

The aim of this GIS application is to introduce GIS in the HIV/Aids research field and also to demonstrate an analysis on district level. With help from Ass. Prof. Bertil

Egerö, one of the Swedish researchers in the Sida project, and Dr. Edward Kirumira, the Dean of the Faculty of Social Sciences at Makerere University, we found statistics for organizations working with HIV/Aids in Uganda. The statistics were taken from the report “Inventory of agencies with HIV/Aids activities and HIV/Aids interventions in Uganda 2001” (AMREF, 2001). The report contains topical information about how many and what kind of organizations there are in each district in Uganda. We also found current information about the HIV infection prevalence rates in eight districts in Uganda.

With the functions “Graph showing relationship between two attributes” and “multiple regression” in the statistical tool, we investigated statistical relationships between the HIV/Aids data and other parameters. The parameters we used are:

- number of organizations in each district 2001 (AMREF, 2001),
- GDP index for each district (Uganda Bureau of Statistics, 2000),
- reported Aids cases 1998 (Ministry of Finance, 2001),
- distance to Kampala (from each district capital town),
- population in each district (Uganda Bureau of Statistics, 2002), and
- literacy rate in each district 1996.

The figures used in the analyses are scaled to be in the same size. Also note that the data are from different years. This slightly affects the reliability of the analyses. The results from the analyses are shown in the figures below.

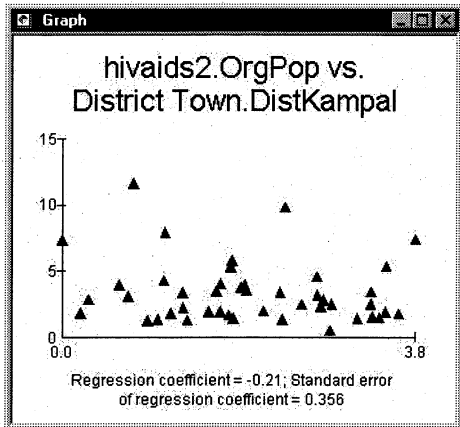


Figure 7.5. A Graph showing the relationship between the number of HIV/aids organizations (y-axis) and the distance to Kampala (x-axis).

hivaids2.OrgPop = number of organizations divided by the population in each district and multiplied by 100000.

District Town.DistKampal = distance from all districts capital cities to Kampala divided by 100000.

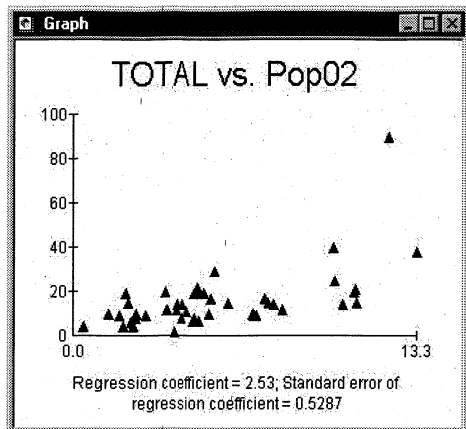


Figure 7.6. A Graph showing the relationship between the number of HIV/AIDS organizations (y-axis) and population in each district (x-axis).

TOTAL = number of organizations in each district.

Pop02 = population in each district divided by 100000.

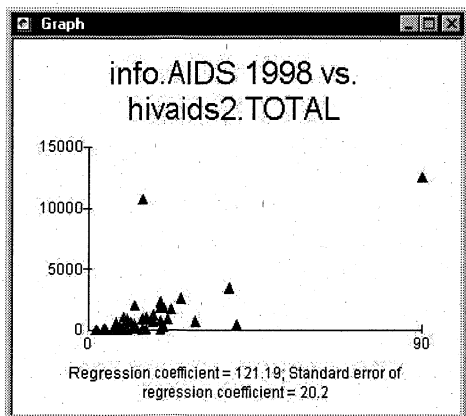


Figure 7.7. A Graph showing the relationship between reported Aids cases 1998 (y-axis) and the number of HIV/AIDS organizations in each district (x-axis).

info.AIDS 1998 = reported Aids cases 1998.

hivaid2.TOTAL = number of organizations in each district

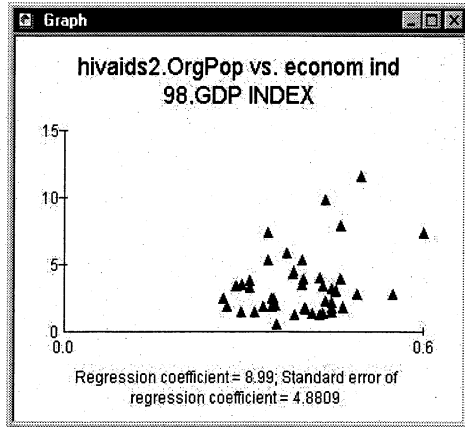


Figure 7.8. A Graph showing the relationship between the number of HIV/Aids organizations in each district (y-axis) and GDP index 1998 (x-axis).

hiv aids2.OrgPop= number of organizations divided by the population in each district and multiplied by 100000.

econom ind 98. GDP INDEX = the Gross Domestic Product 1998.

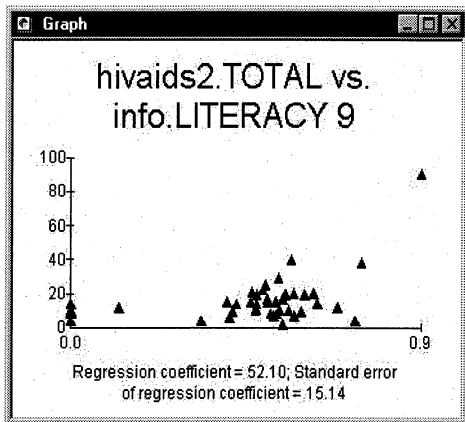


Figure 7.9. A Graph showing the relationship between the number of HIV/Aids organizations in each district (y-axis) and literacy rate 1996 (x-axis).

hiv aids2.TOTAL = number of organizations in each district.

info.LITERACY 9 = literacy rate 1996.

In some of the graphs in Figure 7.5 to Figure 7.9 a relationship between parameters can be suspected. Figure 7.5 does not seem to show any strong relationship between the parameters as the plotted values in the graph are spread out and the standard error of the regression coefficient is even larger than the regression coefficient. Figure 7.6 shows how the number of organizations depends on the population in each district. It seems likely that the number of organizations would increase as the population increases. Figure 7.7 also seems to show a relationship. It is possible that the graph and the statistical parameters show that there are many reported Aids cases if the district has many organizations. From the graph and the statistical parameters in Figure 7.9 it seems likely that there are more organizations if there is a higher literacy rate.

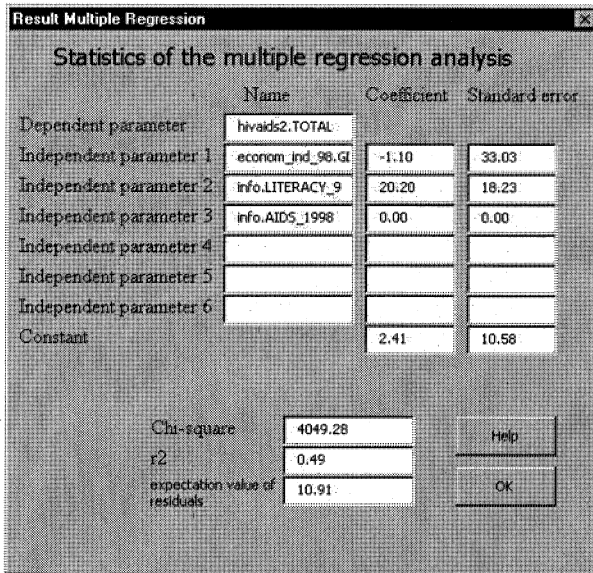


Figure 7.10. Result from a multiple regression analysis with number of HIV/Aids organizations in each district, GDP index 1998, literacy rate 1996 and reported AIDS cases 1998 as parameters.

hivaids2.TOTAL = number of organizations in each district.

econom_ind_98.GDP = the Gross Domestic Product 1998.

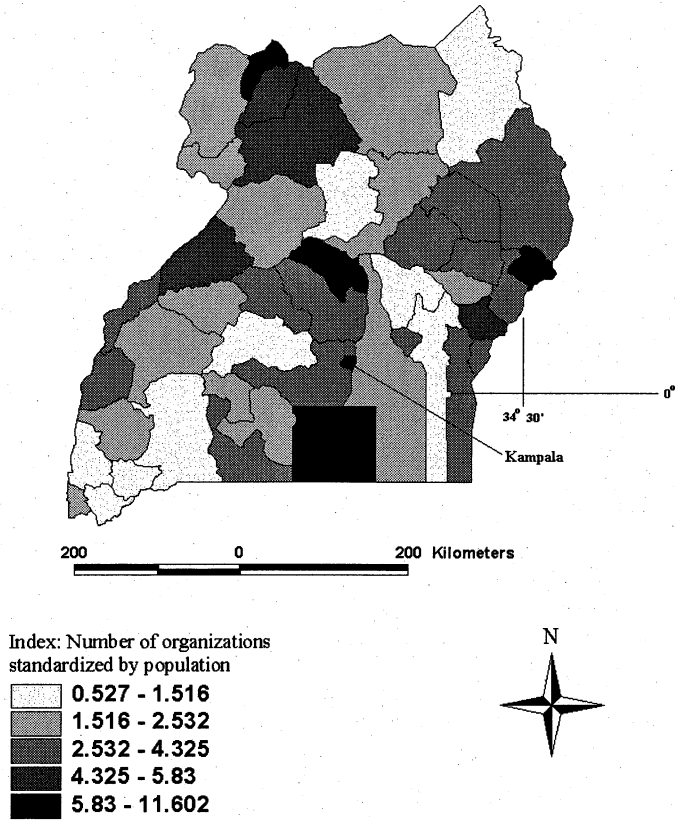
info.LITERACY_9 = literacy rate 1996.

info.AIDS_1998 = reported Aids cases 1998.

Figure 7.10 shows the result of a multiple regression. In this case no strong relationship can be seen. If the relationship is strong the parameter r^2 is supposed to have a value close to one. A value of 0.5 for r^2 approximately means that half of the values of the dependent parameter are explained by the independent parameters. The coefficients and the standard errors do not seem to show any strong relationship either. The standard error is supposed to be much less than the coefficient and that is not the case for any of these parameters. The only case where a relationship can be suspected is for the parameter literacy rate 1996, where the standard error is less than the value of the coefficient.

The figure below shows a visualization on district level of the HIV/Aids organizations in Uganda. No special pattern can be seen in the map. This figure can be compared with Figure 7.5 above, where no relationship between distances to Kampala and number of organizations can be seen.

HIV/AIDS organizations in Uganda 2002



Data source AIDS commission report 2002 and Bureau of Statistics, Uganda
Maria Ljungblom and Karin Gullstrand

Figure 7.11 Layout from ArcView 3.2 showing HIV/Aids organizations standardized with population per district in Uganda.

8 Discussion

This discussion will first treat the technical part of our project, and then continue with the implementation part.

The statistical tool can be further developed. Examples of how it could be expanded are:

- create more geographical attributes, e.g. distance between polygons,
- improved design of the graph, and
- more statistical parameters, e.g. Morans I (Eklundh, 2000) for correlation of the residuals etc.

An expansion of the tool would make it possible for the researchers within social sciences to perform more advanced statistical analyses. This would further motivate the researchers to perform their analyses using an integrated statistical tool in a GIS.

A lot of work is required for introducing GIS in research areas such as social sciences, in e.g. developing countries. The researchers at the Faculty of Social Sciences (FSS) at Makerere University are generally not very familiar with computers, partly because there has been a lack of technical equipment at the university, and partly because they work within a non-technical field. Therefore, much effort needs to be concentrated on basic computer knowledge if GIS is going to be used. With this in mind it is well-founded to pose the questions:

1. Is GIS a useful tool for people that hardly even have basic computer skills?
2. Will the implementation of GIS continue without the support from e.g. a Sida project?
3. Do the researchers themselves find the implementation of GIS important or is it just forced on by the GIS spokesmen?

These questions are hard to answer, but still important to have in mind when performing these kinds of projects. On the basis of our experience we will try to answer the questions.

1. Certain basic computer knowledge, such as file management, is required for GIS to be a useful tool. This could be solved by a course in basic computer knowledge, for those who need it.
2. To make it possible to proceed with the use of GIS without support from outside, it is important that at least one person is educated in computer use and GIS. This education could be held at the most suitable institution or in a country with more GIS knowledge.
3. Some people, especially those with less computer knowledge, do not think that GIS is important, but as long as there is an adequate interest among the researchers it is motivated with further implementation of GIS.

Another topic to discuss is whether the digitizing of new maps should be done at FSS or in Sweden. For this task, equipment and skilled persons are required, which are currently not found at FSS. Since it would require a lot of money and work to establish digitizing possibilities at FSS, for the moment it might be best to perform this work in Sweden. In the future, cooperation with other skilled faculties (e.g. the Department of Geography at Makerere University) might be a better alternative than digitizing maps in Sweden.

9 Conclusions

The use of GIS as a research tool in social sciences has many advantages. For example, it is possible to discover and analyze spatial patterns, such as the spread of HIV/Aids or the economical distribution throughout a country or a district. This can be useful to realize where and what efforts are needed for a positive development of the country.

Since the computer skills among researchers in non-technical fields can be rather low, there is a need for user-friendly tools to enable the implementation of GIS in social sciences. When integrating a statistical tool in ArcMap, VBA was a suitable programming language to use. It is integrated in the program and easy to learn. VBA makes it possible to create user-friendly applications in a simple way, since it is a graphic programming language. For more demanding calculations other programming languages, such as VC++, are more suitable.

During the implementation of the statistical tool, the user-friendly help menus and the exercise on the statistical tool seemed to be on a suitable level for the workshop participants. The help menus were useful when performing analyses with the statistical tool and the participants found the tool user-friendly and therefore helpful in their research.

The most important issues that currently affect the implementation of GIS in research at the Faculty of Social Sciences are:

- a person, responsible for the GIS activity at the faculty, needs to be selected,
- the existing database at FSS is rather poor and therefore needs to be expanded, and
- training in GIS, basic computer knowledge, etc is needed.

There are several problems connected to these three points. For example, the responsible GIS person should be someone with good knowledge in GIS and computers. A problem connected to data collection is that statistical data might not always exist. Another obvious problem is the economic aspect of training, equipment, etc.

Although there is a need of user-friendly tools, one must not forget the importance of continuous training for the users.

10 References

African Medical and Research Foundation (AMREF-Uganda) and The Uganda Aids Commission Secretariat, 2001. *Inventory of Agencies with HIV/Aids Activities and HIV/Aids Interventions in Uganda*. Pfizer International Corporation, Uganda.

Blom, G., 1989. *Sannolikhetsteori och statistikteori med tillämpningar*. Studentlitteratur, Lund, Sweden.

Blomh, L., 1975. *Multipel regressionsanalys och fastighetsvärdering*. Svensk Lantmäteritidskrift (SLT) 1975:1.

Booth, B., and Mitchell, A., 2001. *Getting started with ArcGIS*. ESRI, Redlands, California.

Bra böckers lexikon 2000, Bra böckers förlag AB, Höganäs, Sweden, 1999

Dijkstra, E. W., 1959. *A Note on Two Problems in Connexion with Graphs*, *Numerische Mathematik*, Vol. 1, pp. 269-271.

Eklundh, L., Arnberg, W., Arnborg, S., Harrie, L., Hauska, H., Olsson, L., Pilesjö, P., Rystedt, B., and Sandgren, U., 2000. *Geografisk informationsbehandling – metoder och tillämpningar*. Bygghörsningsrådet, Stockholm.

Enger, J., 1990. *Linjära modeller inklusive regressions- och variansanalysmodeller*. Matematisk statistik, KTH, Stockholm.

Faculty of Social Sciences (FSS), 2001. *Consolidating peace and development in the Lake Victoria region and its environs: The national and local responses to transformation from turmoil to a more sustainable development process*. SIDA application.

Haining, R., 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge.

Leggett, I., 2001. *The background, the issues, the people - Uganda*. Oxfam GB, Fountain Publishers LTD, 2001.

Lindgren, B. W., 1976. *Statistical theory*. Macmillan Publishing Co, New York.

Ministry of Finance, 2001. *Poverty Eradication Action Plan (2001-2003) (Volume1)*. Planning and Economic Development Kampala February 2001.

Pilesjö, P., 2001. *Geographical Information Systems in Development Studies*. In Närman, A and Karunanayake, K., (eds), 2002, *Towards a new regional and local development research agenda*. Göteborg University and University of Kelaniya, Göteborg, Sweden.

Robin, A., Peterson, A., and Tjernlund, M., 1994. *Visual Basic – programutveckling från början*. Studentlitteratur, Lund, Sweden.

Sida, 2001. *Landstrategi Sverige Uganda 2001-2005*.

Sommerville, I., 2001. *Software Engineering, 6th Edition*. Pearson Education Limited, Essex, England.

Uganda Bureau of Statistics, 2000. *Human Development Indicators District Profile 1998*. Human Development Report, UNDP.

Uganda Bureau of Statistics, 2002. *Population census 2002*. In *The New Vision*, Kampala, Uganda, October 2002.

Zeiler, M., 2001. *Exploring ArcObjects, Vol.1*. ESRI, Redlands, California.

Ministry of Finance, 2001. *Poverty Eradication Action Plan (2001-2003) (Volume1)*. Planning and Economic Development, Kampala, Uganda, February 2001.

Internet addresses:

Africaaction, 2002 <http://www.africaaction.org/action/women.htm> 2002-12-18.

CIA, 2002 <http://www.cia.gov/cia/publications/factbook/geos/ug.html> 2002-05-02.

University of Nebraska, Omaha (unomaha), 2002.
http://maps.unomaha.edu/Peterson/funda/Quiz***/SubAfricaNAmerica/maps/Uganda.GIF
2002-05-02.

Esri, 2002a. <http://www.esri.com/software/index.html> 2002-05-15.

Esri, 2002b. <http://www.esri.com/library/brochures/pdfs/arcob81bro.pdf>, 2002-05-15.

Esri, 2002c. <http://www.esri.com/software/devsolutions/arcobjects/index.html> 2002-05-15.

Microsoft, 2002a <http://msdn.microsoft.com>, 2002-05-15.

Microsoft, 2002b <http://www.microsoft.com/com/tech/com.asp>, 2002-05-09.

Microsoft, 2002c <http://msdn.microsoft.com/vba/prodinfo/backgroundunder.asp>, 2002-05-13.

MyUganda, 2002 <http://www.myuganda.co.ug/categories/about/economy/index.htm>
2002-12-09.

SAS, 2002 http://www.sas.upenn.edu/African_Studies/NEH/ug.html 2002-05-03.

Sida, 2002 <http://www.Sida.se>, 2002-04-10.

UI, 2002 <http://www.ui.se/fakta/afrika/uganda.htm> (Utrikespolitiska Institutet), 2002-04-10.

Appendix A

Lars Harrie, 2002-01-16

Updated by: Karin Gullstrand and Maria Ljungblom, 2002-12-16

Statistical module in ArcView

Technical specification

This document contains a specification of a statistical module in ArcMap (from ESRI). The module can compute new parameters from geometric quantities, show graphs with simple statistical relationships, and perform a multiple regression. Notice that this is only a technical specification; user instructions (help menus) are found in another document.

Contents

1.1.	PROGRAM ENVIRONMENT	1
1.2.	DEMANDS ON COMPUTER ENVIRONMENT	1
1.3.	TO START A PROJECT WITH THE STATISTICAL MODULE.....	1
1.4.	DEMANDS ON GEOGRAPHICAL DATA	1
1.5.	KNOWN LIMITATIONS.....	2
2.	PROGRAM OVERVIEW	3

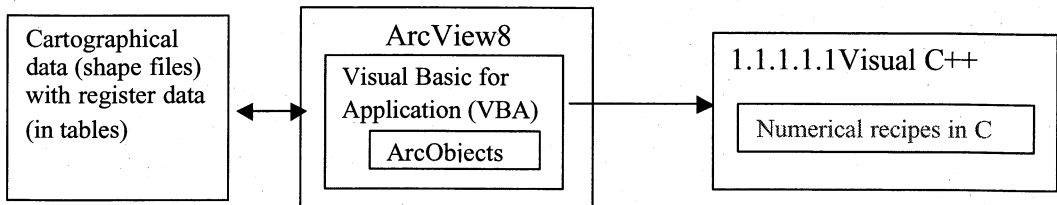
Appendix I: Visual C++ program for computing multiple regression statistics

Appendix II: Installation instructions

Appendix A

1.1. Program environment

The module is written in *Visual Basic for Application (VBA)* inside *ArcView8 (ArcMap)* using the COM-library *ArcObjects*. The more demanding calculations (multiple regression) are written in *Visual C++* (using the mathematical library *Numerical Recipes*).



1.2. Demands on computer environment

The statistical module requires the following equipment:

- PC with Windows NT or 2000, or newer version
- ArcView8 (ESRI) or ArcInfo8
- Explorer (Microsoft) – used for the help menus

1.3. To start a project with the statistical module

- 1) After starting ArcMap the project *Statistic.mxd* is chosen
- 2) Start the analyses by clicking the "smiley" (*My Own Tools*).

1.4. Demands on geographical data

- The data must be stored in *shape format*.
- The cells that are missing data must contain the value 0 in the database.
- When adding new data showing the shortest distance (Euclidian distance) between objects the attributes must be found in point tables where the geometry is stored as any *point* format
- When adding new data that show the distances along roads (traveling distance) between objects, the following requirements must be fulfilled (it is very important that no names or suchlike are changed in these tables):
 - 1) A district table (polygon) that contains an attribute with the name of the district must exist. This table should be called "district".
 - 2) A town table containing province cities for all districts and the node of each city. This table should be called "townposition".
 - 3) A road table containing the length of each segment between two nodes. The table should have the name "roads".
- When using the graph function the included attributes must be found in the same table. If they are not, then a join of the two tables must be performed.
- When using the multiple regression function the same thing concerning join as above is applicable.

Appendix A

1.5. Known limitations

The statistical module has the following known limitations:

- The objects are not listed in alphabetic order in the menus (this can be done in new versions of VB but not in the version of VBA that is installed in the current version of *ArcView8*).
- Statistical parameters (Moran1) for correlation between residuals etc. are not fully implemented in the current version of the tool (2002-12-12). This can be implemented when needed.
- It is not possible to use more than 6 explanatory variables in the multiple regression analysis.
- The independent variables are treated as independent in the multiple regression. A principal component analysis can be implemented to treat this problem, although this is relatively difficult.
- The calculations of the parameters in the multiple regression strictly assumes that data are normal distributed, which is an uncertain assumption. It is possible to test the assumption with a X^2 test (see *Numerical recipes* section 15.1) or by using build-in functions in ArcView – Geostatistical analysis, such as Histogram or QQ Plot (see ESRI manual *ArcGIS Geostatistical Analyst*, ch. 4). Possibly the Robust Estimator must be used for this (see *Numerical recipes* manual 15.7). Since we do not know σ_i we have to approximate it with 1 for all the observations.
- No weighting of the observations is implemented.
- The multiple regression function cannot handle very high values for the independent variables (in the size of 10^5 and bigger). This is probably due to problems in the communication between VBA and VC++.
- The distance along roads can only be computed from all the district cities to one arbitrary city and using tables with certain contents and names (see above).
- Quality of roads, speed limits, etc are not taken into consideration when the traveling distance is computed, because data concerning this was not found in the database.

2. Program overview

In this chapter a summary of the program structure is given.

The program is roughly structured as follows. A toolbox with a control button is created in Visual Basic. When pushing this button a form is called (UserForm3). After this, the forms are called depending on the choices of the user (see Figure 3.1). Visual C++ has been used to create a function called *multiple regression* to compute the parameters of the multiple regression. This function is mainly based on Single Value Decomposition (SVD) functions in *Numerical Recipes* (see the book *Numerical Recipes in C, 2nd edition*, section 2.6 and 15.4). The C++ function is stored in a DLL file (Dynamic Link Library), which always has to be copied to a system library (e.g. c:\winnt\system32).

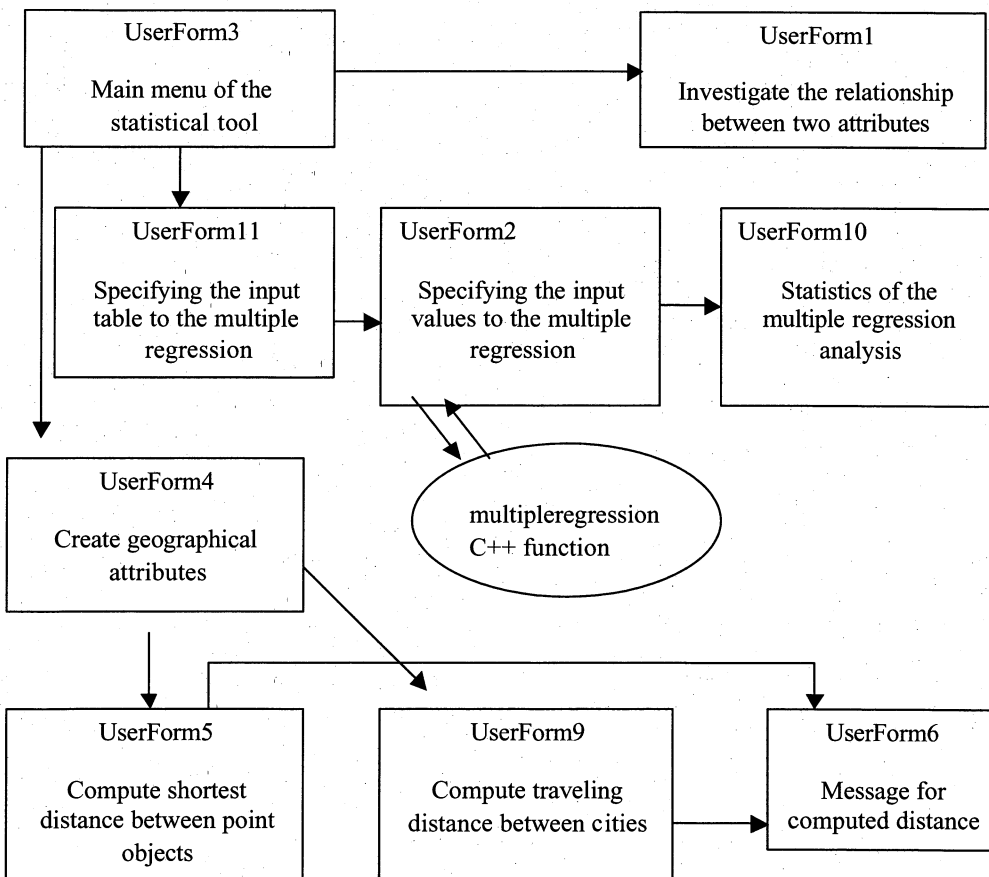


Figure 2.1. Overview of how the different user forms are connected in the VBA program.

Appendix I: Visual C++ program for computing multiple regression statistics

Author: Lars Harrie, 2002-01-16

This routine computes the multiple regression statistics from the values of the dependent and independent data. The function *xsvd.cpp* is written for this application and *svdfit.cpp* is partly changed. The remaining function is original *Numerical Recipes* functions for single value decomposition (SVD) (see *Numerical Recipes in C*, 2nd edition, sections 2.6 and 15.4).

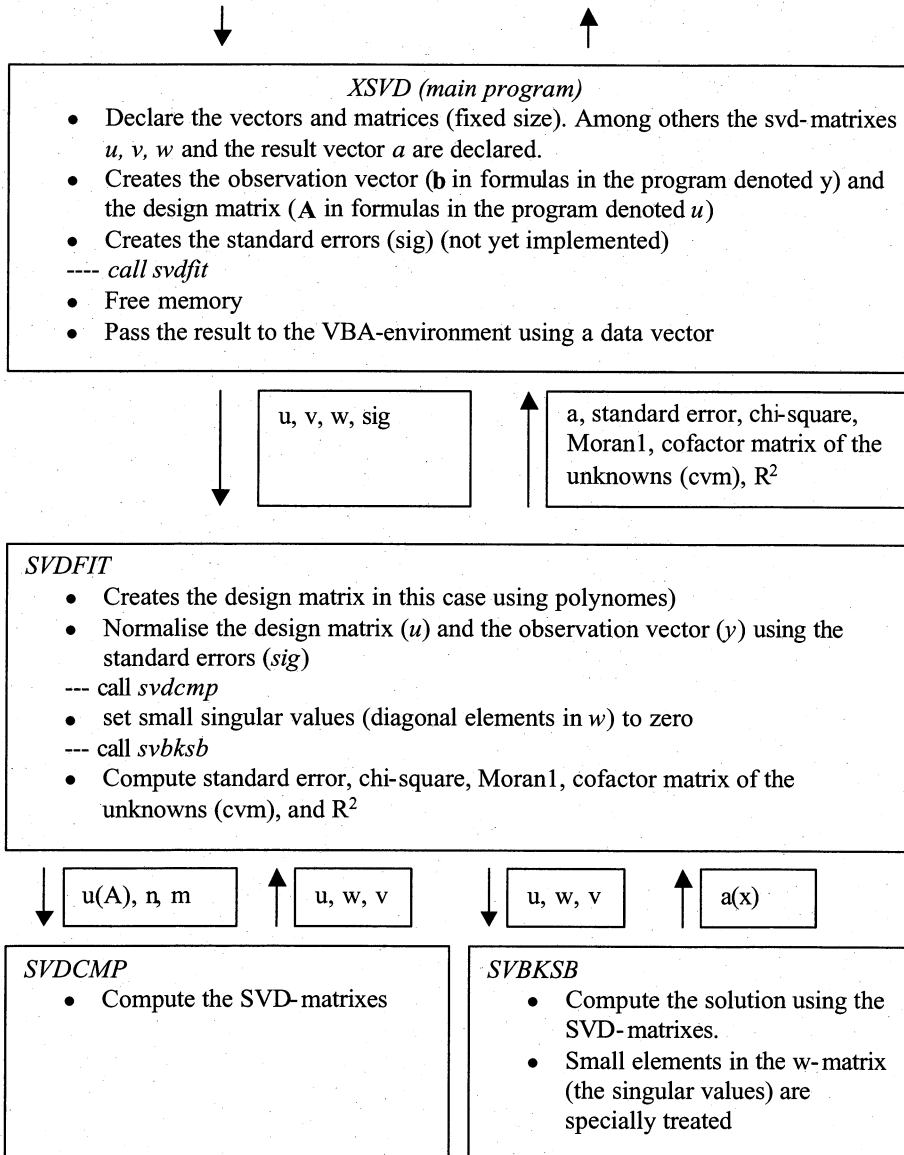
Workflow of the SVD-programs (see also next page)

The statistical module in Arcview implemented in VBA and ArcObjects

* The value of the design matrix (A) and the observation vector (b) stored in a data vector.
* Two integers describing the size of the design matrix and the observation vector.
* The standard errors used for weighting
* (Binary) adjacency matrix for the Moran statistics (The two last items are not yet implemented)

* The solution vector (a)
* The standard errors of the solution vector
* The R²-value
* Moran-1 statistics
* Chi-square value

Appendix A



Appendix A

Appendix II: Installation instructions

The following must be done before the statistical module can be used:

- Install ArcView8.
- Copy the database to the folder you want to use when adding files to ArcMap.
- Copy the file *Statistic.mxd* to a suitable folder, e.g. the same as above.
- Copy the dll-file *nrdll.dll* to the system library (e.g. C:\WINNT\system32).
- Choose point as decimal symbol in *Windows* (set under Start -> Control Panel -> Regional Options).

Open *StatisticalTool.mxd* in ArcMap and add the relevant layers from the database.



Appendix B: Help menus

Help Menu – Statistical tool

New geographical attributes

Here a new geographical attribute is created and will be stored in a new column in an existing table. The following geographical attributes can be computed:

- The shortest distances (Euclidean distance) between a chosen geometric object and all the geometric objects in a chosen table. Click [here](#) to see an example.
- The nearest travelling distances between an arbitrary city and the main city in each district. Click [here](#) to see an example.

Graph showing relationship between two attributes

This function creates a graph showing the relationship between two chosen attributes. It also gives parameters that show how much the attributes are related to each other. Click [here](#) to see an example.

Note! The attributes must be in the same table. Before you use this function you need to join the tables containing the attributes you wish to see in the graph. How to join tables is described [here](#).

Multiple regression analysis

This function investigates the relationship between two or several attributes. The result shows how one attribute depends on one or several other attributes. The result will be shown as the degree of explanation (R^2), chi-square, the expectation value of the residuals (σ), and the standard error and coefficients for all of the independent attributes. Click [here](#) to see an example.

Note! The attributes must be in the same table. Before you use this function you need to join the tables containing the attributes you wish to see the relationship between. How to join tables is described [here](#).

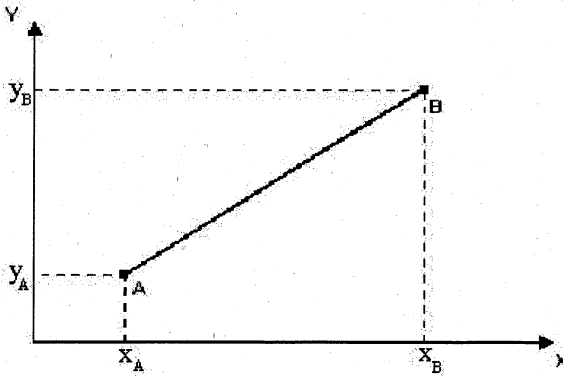
Help Menu – Create geographical attributes

Compute shortest distance between geometric objects

This function computes the shortest distances (i.e. not along roads) between one chosen geometric object and all the geometric objects in a chosen table. The shortest distance between two positions is along a straight line between these two positions. Click [here](#) to see an example.

To compute the shortest distance between A and B (see picture below) we use the theorem of Pythagoras:

$$\text{Distance} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$



The result will be added in the table containing the geometric objects from which the distances are computed.

Compute travelling distance

This function computes the shortest travelling distances between one chosen city and the district capital cities. Travelling distance means the distance from one point to another along roads and is therefore different from the shortest distance. The distances are computed by using Dijkstras algorithm. Computing the travelling distances with this function takes approximately 5-10 minutes. The result will be added in the table containing the district cities. Click [here](#) to see an example.

Help Menu – Compute shortest distance between geometric objects

Click [here](#) to see an example.

In the first box you specify the name of the new field. This field will be added to the table containing the geometric objects from which the distances are computed. The name of the new field should preferably be called something that reveals what it will contain, for example dKampala (d as in distance and the name of the city to which the distance is computed). Note that it is only possible to use ten characters!

In the next box to the left you select the table containing the objects you wish to compute the distances from.

In the boxes to the right you first select the table containing the object to which the distances should be computed. The fields in your selected table are then listed in the next box. Choose the field containing the wanted object. In the last box you choose the object.

After pressing OK you can find the result in the table to which the new field was added. The geometric object you have chosen to compute the distance to will get the value zero. This is because there is no distance between the chosen object and itself.

Help Menu – Example of computing shortest distance between geometric attributes

This example will show how to compute the distances between the capital of Uganda (found in the table "town") and all the other cities in Uganda that in this example also are collected in the table called "town".

- After clicking the "smiley" you reach the main menu.
- Click the button saying "New geographical attributes".
- Click the button saying "Compute shortest distance between geometric attributes.
- In the next menu (see Figure 1 below):
 - 1) Fill in the name of the new field that later will contain the computed distances.
 - 2) Select the table "town" (which contains all the cities) by pressing the arrow and scroll down to "town" and click on it.
 - 3) Select the table "town" (which contains the capital Kampala) again.
 - 4) Select the field "name", which contains the names of all the cities.
 - 5) Select Kampala (since the distances should be computed to the capital in this example).
 - 6) Click OK.

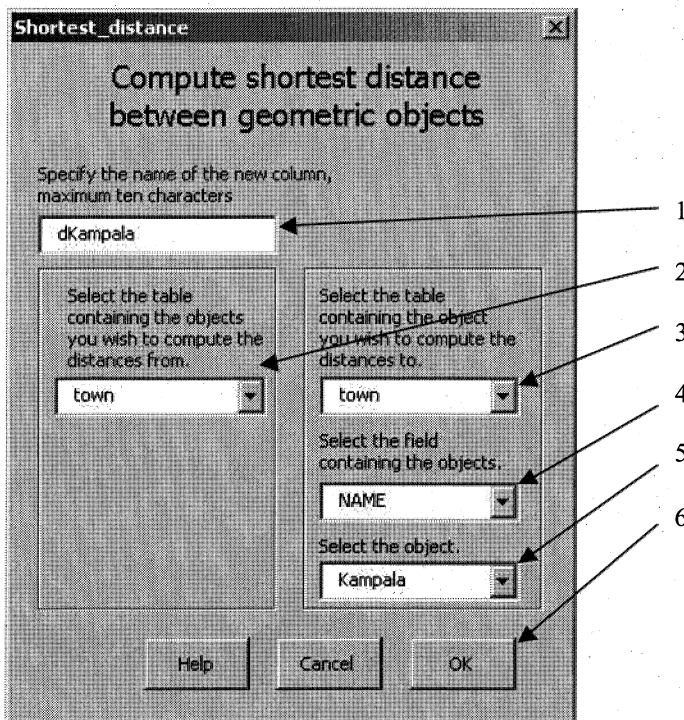
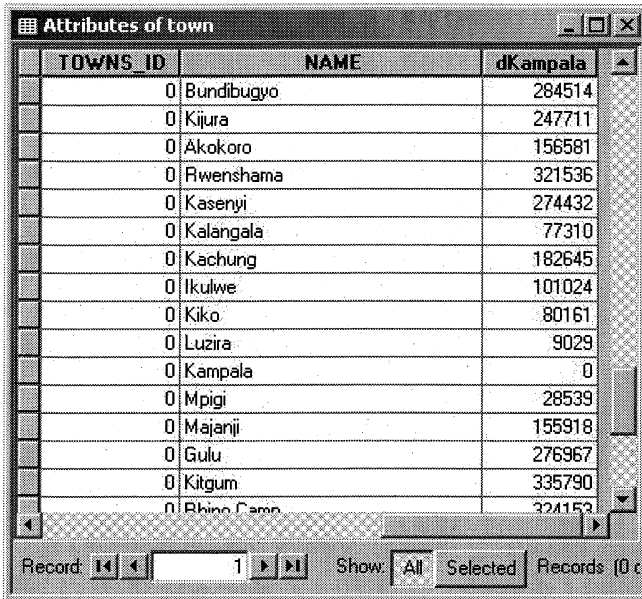


Figure 1. An example of how to fill in the last menu for the distance functionality.

- Click OK when you get the message that the distances are computed.
- Open the table "town" containing all the cities to see the result. To open the table, right-click on "town" in the table of contents and click "Open Attribute Table".

Appendix B

- Scroll to the last field, which you named earlier, and see the result (see Figure 2 below). You can for example see that the distance from Akokoro to Kampala is 156581 meter. Also note that the distance from Kampala to Kampala is 0.



TOWNS_ID	NAME	dKampala
0	Bundibugyo	284514
0	Kijura	247711
0	Akokoro	156581
0	Rwenshama	321536
0	Kasenyi	274432
0	Kalangala	77310
0	Kachung	182645
0	Ikulwe	101024
0	Kiko	80161
0	Luzira	9029
0	Kampala	0
0	Mpigi	28539
0	Majanji	155918
0	Gulu	276967
0	Kitgum	335790
0	Rhino Camp	324153

Figure 2. Result from computing the Euclidean distance between Kampala and the other cities in the table "town"

Help Menu – Compute travelling distance

Click [here](#) to see an example.

In the first box you specify the name of the new field. This field will be added to the table containing the district cities. The name of the new field should preferably be called something that reveals what it will contain, for example tdKampala (td for travelling distance and the name of the city to which the distance is computed). In the next box you select the city to which you wish to compute the travelling distance from the district cities.

After pressing OK you can find the result in the district table, to which the new field was added. For the city you have chosen and the city Kalangala there will be a zero. This is because there is no distance between the chosen city and itself and because Kalangala is situated on an island with no connecting roads.

Help Menu – Example of computing travelling distance between cities

This example will show how to compute the travelling (along roads) distances between the capital of Uganda and the district capital cities in Uganda (found in the table "District_Town").

- After clicking the "smiley" you reach the main menu.
- Click the button saying "New geographical attributes".
- Click the button saying "Compute travelling distance between cities."
- In the next menu (see Figure 1 below):
 - 1) Fill in the name of the new field that later will contain the computed travelling distances.
 - 2) Select the capital by pressing the arrow and scroll down to "Kampala" and click on it.
 - 3) Click OK.

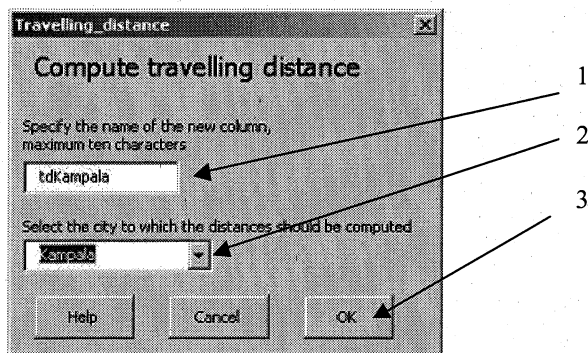


Figure 1. An example of how to fill in the last menu for the travelling distance functionality.

- Click OK when you get the message that the distances are computed.
- Open the table "District_Town" containing all the cities to see the result. To open the table, right-click on " District_Town " in the table of contents and click "Open Attribute Table".
- Scroll to the last field, which you named earlier, and see the result (see Figure 2 below). You can for example see that the distance from Tororo to Kampala is 209500 meters. Also note that the distance from Kampala to Kampala is 0. The distance from Kalangala to Kampala is also indicated by 0, since there is no road connected to the city.

Appendix B

DISTRICT_T	POPGROWTH	SAFE_WATER	tdKampala
Tororo	2.82	21.27	209500
Fort Portal	3.29	5.98	283823
Mubende	2.72	11.94	145460
Bugiri	0	0	146371
Jinja	2.15	66.7	79141
Busia	0	0	191031
Mpigi	2.94	25.66	30150
Kasese	1.94	37.37	352115
Kampala	4.76	87.17	0
Sembabule	0	0	148495
Mbarara	2.75	18.93	261304
Masaka	2.71	10.4	124517
Rakai	3.04	5.58	185749
Bushenyi	3.08	14.77	311750
Kalangala	5.88	6.27	0
Rukungiri	2.51	26.63	364884
Ntungamo	0	0	320662
Kisoro	3.53	22.33	438584
Kabale	2.17	58	384973

Record: 0 Show: All Selected Records (0 out of 45 Selected.)

Figure 2. Result from computing the travelling distance between Kampala and the district cities in the table "district".

Help Menu – Graph showing relationship between two attributes

Observe! Before you use this function you need to join the tables containing the attributes you wish to see in the graph, if they are not already fields in the same table. How to join tables is described [here](#).

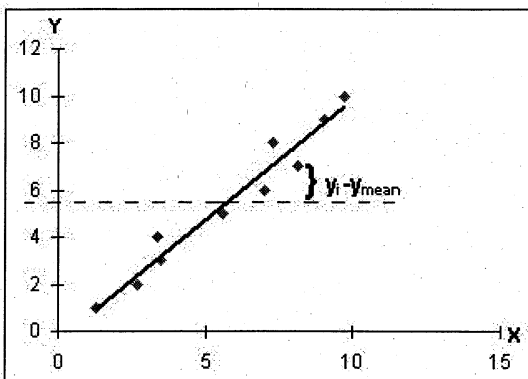
Click [here](#) to see an example of the graph functionality.

In the first box to the left you select the table containing the variables between which you like to investigate the relationship. In the other boxes you select the dependent and the independent variables. The dependent variable is what you wish to explain, for example illiteracy, and the independent variable is what you wish to explain with, for example distance to school.

After clicking OK the result will be shown in a graph where the values of the attributes are plotted. If the dots in the graph seem to create a line it is likely that the attributes are related to each other. The regression coefficient (β^*) shows the slope of the regression line. If the coefficient is zero (notice that it is only shown with two decimals) there is no relationship. The standard error of the regression coefficient is the standard deviation of the estimated regression coefficient. If the standard error of the regression coefficient is much less than the regression coefficient there is most likely a relationship. A higher value on the regression coefficient allows a higher value on the standard error.

The graph below shows a regression where the y-values depend on the x-values (the red dots). The broken line shows the mean value of the y-values and the other line is the calculated regression line. In the graph you can see that β^* must be larger than zero because the slope is positive. This is given by

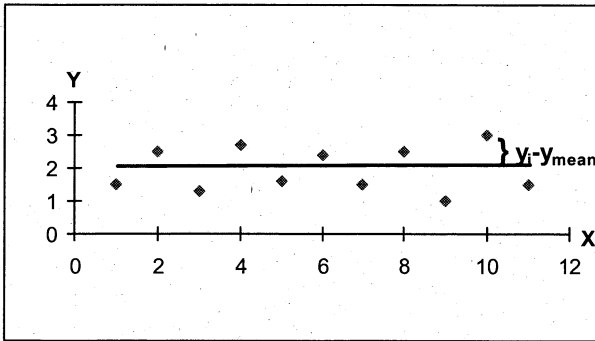
$$\beta^* = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \text{ where } \bar{y} (=y_{\text{mean}}) \text{ is the mean value of the y-values.}$$



In this case, the regression coefficient is 1,04 and the standard error is 0,09

Appendix B

The graph below shows the regression where the y-values do not depend on the x-values. The y-values are almost the same for all the x-values. The mean value coincide with the regression line. In this case β^* is almost zero because the slope of the regression line is zero.



In this case, the regression coefficient is 0,03 and the standard error is 0,08

Observe! After using this function you should remove the join to avoid unnecessary large tables. How to do this is described [here](#).

Help Menu – Example of creating a graph showing the relationship between two attributes

This example will show how to create a graph showing the relationship between the attributes area (in km²) and number of cattle in the districts of Uganda.

- After clicking the "smiley" you reach the main menu.
- Click the button saying "Graph showing relationship between two attributes".
- In the next menu (see Figure 1 below):
 - 1) Select the table "sources of income" (which contains the two attributes) by pressing the first left arrow and scroll down to "sources of income" and click on it.
 - 2) Select the dependent variable area by pressing the second left arrow and scroll down to "real area" and click on it.
 - 3) Select the independent variable cattle by pressing the right arrow and scroll down to "cattle_199" and click on it
 - 4) Click OK.

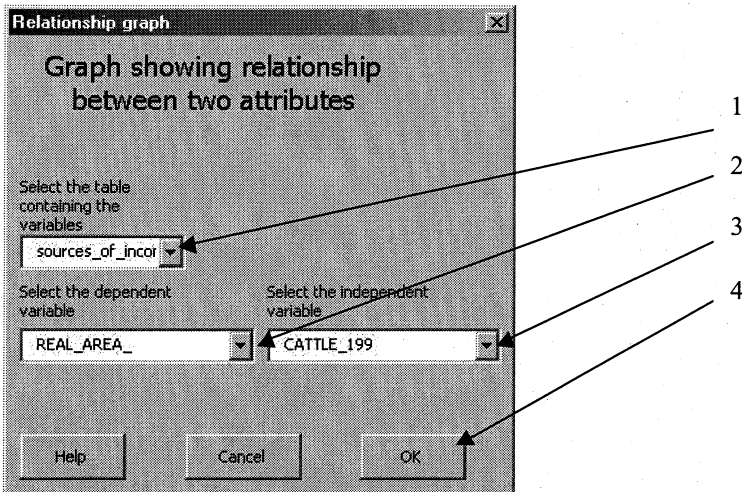


Figure 1. An example of how to fill in the last menu for the graph functionality.

- The result is shown in a graph where the two variables are plotted (see Figure 2 below). At the bottom of the graph two regression parameters are also shown, which show if there is a relationship or not. The relationship can also be seen in the graph by the creation of a line from the plotted values.

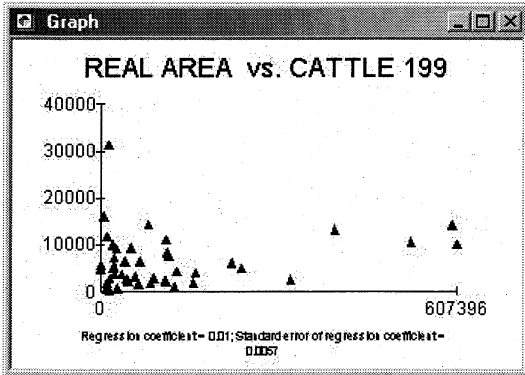


Figure 2. Graph showing relationship between area and number of cattle in the districts of Uganda.

Help menu - Multiple regression

Note! Before you use this function you need to join the tables containing the attributes you wish to see the relationship between if they are not already fields in the same table. How to join tables is described [here](#).

Click [here](#) to see an example.

Select the table containing the variables you wish to use in the regression. This table should be the one you have joined before. Then click OK to move to the next step.

Help menu - Multiple regression

Click [here](#) to see an example.

The fields from your earlier selected table are now listed in the boxes. In the left box you choose the variable you wish to explain (the dependent). In the right boxes you choose the explaining variables (the independent variables). You can choose minimum one and maximum six variables to the right. Observe that the variables have to be chosen from the top and down, i.e. you cannot choose one variable in the first box and one in the third, without choosing one in the second box.

Help menu - Statistics of the multiple regression analysis

In the result table you can see how well the chosen independent parameters explain the dependent parameter. The names of the parameters you have selected are listed.

For each independent parameter you can see the coefficient and the standard error of the coefficient. The coefficient is the β -value in the equation

$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \dots + \beta_m(x_{mi} - \bar{x}_m) + e_i$. It is a measure of the slope of the regression line. If β equals zero there is no relationship between x and y ; y is constant for any x . A high value of β means that y changes a lot even for small changes in x .

The constant is the y -value where the estimated regression line crosses the y -axis. For example in the equation $y = 2 + x$, the constant is 2. If you have chosen not to use a constant in the regression this value will be zero and the regression line will pass through the origin.

The expectation value of the residuals ($\hat{\sigma}$) is a measure of how much the observed values diverge from the values calculated from the equation for the regression line. If this value is zero, there is no divergence at all and the observations coincide with the regression line. The conclusion is that the smaller this value is the more the observations are gathered round the regression equation. $\hat{\sigma}$ is calculated by extracting the root of the sum of squared residuals (chi-square) divided with the number of observations minus the number of independent parameters minus 1 (if using a constant):

$$\hat{\sigma} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - m - 1}}$$
, where n is the number of observations and m is the number of independent variables.

R^2 is the degree of explanation, i.e. how well the independent parameters explain the dependent parameter, where

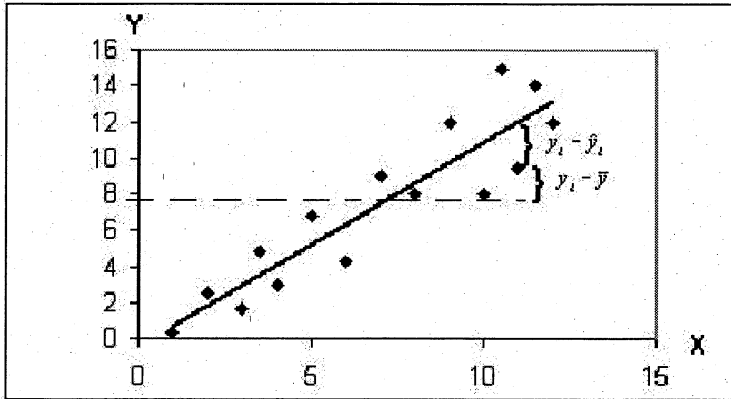
$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

where

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1(x_{1i} - \bar{x}_1) + \hat{\beta}_2(x_{2i} - \bar{x}_2) + \dots + \hat{\beta}_m(x_{mi} - \bar{x}_m) + e_i.$$

This value can vary between 0 and 1 for regression using a constant. If a constant is not used, which is very unusual in social sciences, R^2 can get a value below zero. If R^2 is 1, the independent parameters totally explain the dependent parameter, and if R^2 is 0 they do not explain anything. Also note that when R^2 is 1 $\hat{\sigma}$ is 0, the observations lie on a straight line. See figure x below to understand the relationship between R^2 and observations, regression line and mean value of observations. From looking at the graph below it can be realised that the divergence of the observations from the regression line should be smaller than their divergence from the mean value, if the regression model can explain the dependent variable.

Appendix B



Note! After using this function you should remove the joins to avoid unnecessary large tables. How to do this is described [here](#).

Help Menu – Example of multiple regression analysis

This example will show how to perform a multiple regression analysis on the variables cattle, area (real_area_), the amount of goats (goats_1998) and the amount of sheep (sheep_1998), all collected in the table "sources_of_income", in the districts of Uganda.

- After clicking the "smiley" you reach the main menu.
- Click the button saying "Multiple regression analysis".
- In the next menu (see Figure 1 below):
 - 1) Select the table "sources_of_income" (which contains the attributes) by pressing the left arrow and scroll down to "sources_of_income" and click on it.
 - 2) Click OK.

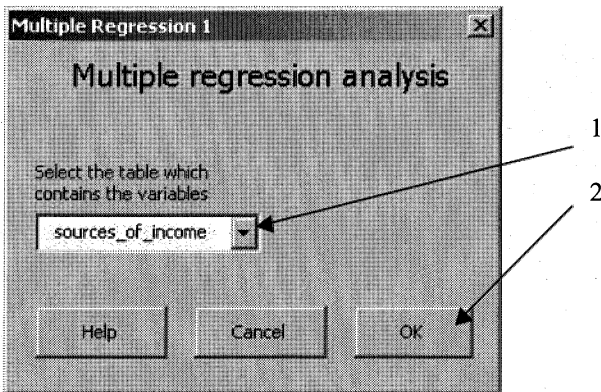


Figure 1. Example of how to fill in the first menu in the multiple regression functionality.

- In the next menu (see Figure 2 below):
 - 1) Select the dependent variable (in this case "goats_1998") by pressing the left arrow and scroll down to "goats_1998" and click on it.
 - 2) Select the first independent variable by pressing the first right arrow and scroll down to "sheep_1998" and click on it.
 - 3) Select the second independent variable by pressing the second right arrow and scroll down to "real_area".
 - 4) Click OK.

Appendix B

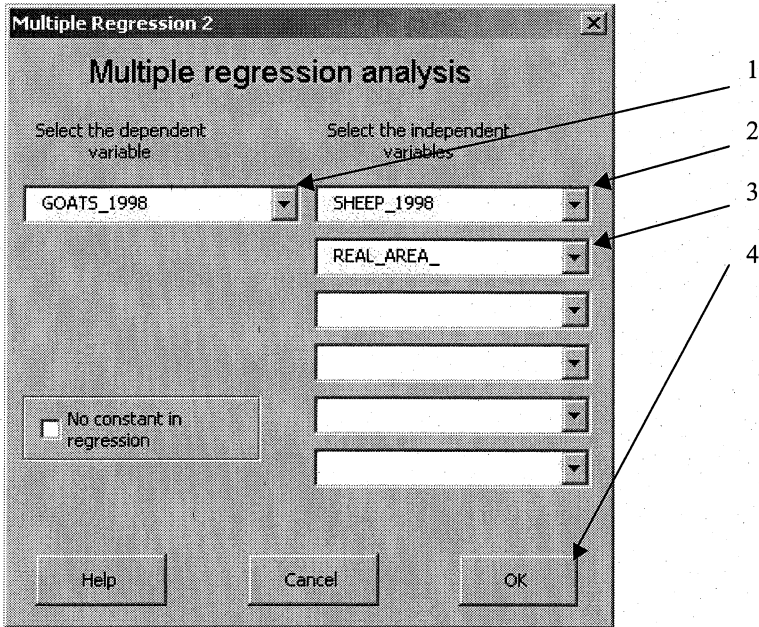


Figure 2. An example of how to fill in the second menu for the multiple regression functionality.

- The result is shown in a menu where the regression parameters are shown (see Figure 3 below).
- After observing the relationship between the variables, press OK.

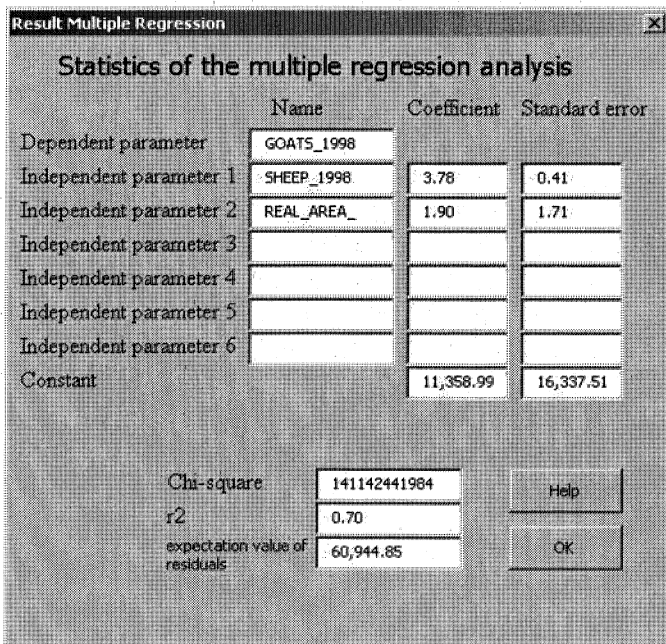


Figure 3. The result from an example of a multiple regression analysis.

Help menu - Joining tables

To join tables means that information from one table is put in another table. To be able to perform a join between two tables there must be fields in the two tables that contain the same type of information, e.g district names. The information from a chosen table will be added to the table on which you perform the join (the layer you right-click). To perform a join follow the instructions below.

- 1) In the table of contents, right-click the layer you want to join, point to Joins and Relates, and click Join.
- 2) Click the first dropdown arrow and click Join attributes from a table.
- 3) Click the next dropdown arrow and click the field name in the layer that the join will be based on (the field with the same type of data as a field in the other table).
- 4) Click the next dropdown to choose the table to join to the layer.
- 5) Click the next dropdown arrow and click the field in the table to base the join on (the field with the same type of data as in step 3).
- 6) Click OK.

The attributes of the table are appended to the layer's attribute table. The field names in the table will have the name of their source table in front of them. If you for example joined information from a table called **population** with a field named **popurban** to the table **district**, this field will be called **population.popurban** in the **district** table.

If you want to join another table to the joined table then repeat all the steps above.

Help menu - Remove joins

After using a joined table in a function from the statistical tool (e.g. graph showing a relationship or multiple regression) you should remove the joins from the table. After removing a join from a table it will look as it did before the join, which is desired. Follow the instructions below to remove a join after using it in a function.

- 1) In the table of contents, right-click the layer containing a join you want to remove (the same that you joined earlier) and point to Joins and Relates.
- 2) Point to Remove Join(s) and click remove all joins.

The table will now look as it did before the join.

Appendix C

GIS

Exercise on the Statistical Tool

Statistical Analysis on Geographical data

ArcView 8.1

By

Karin Gullstrand and Maria Ljungblom

Appendix C

Index

INTRODUCTION TO THE STATISTICAL TOOL.....	1
STARTING ARCMAP	1
USING THE STATISTICAL TOOL.....	2
NEW GEOGRAPHICAL ATTRIBUTES.....	2
GRAPH SHOWING RELATIONSHIP BETWEEN TWO ATTRIBUTES.....	4
JOINING TABLES - GRAPH SHOWING RELATIONSHIP BETWEEN TWO ATTRIBUTES.....	5
MULTIPLE REGRESSION ANALYSIS.....	6
A USEFUL EXERCISE	8
PROBLEMS WHEN USING THE STATISTICAL TOOL.....	8

INTRODUCTION TO THE STATISTICAL TOOL

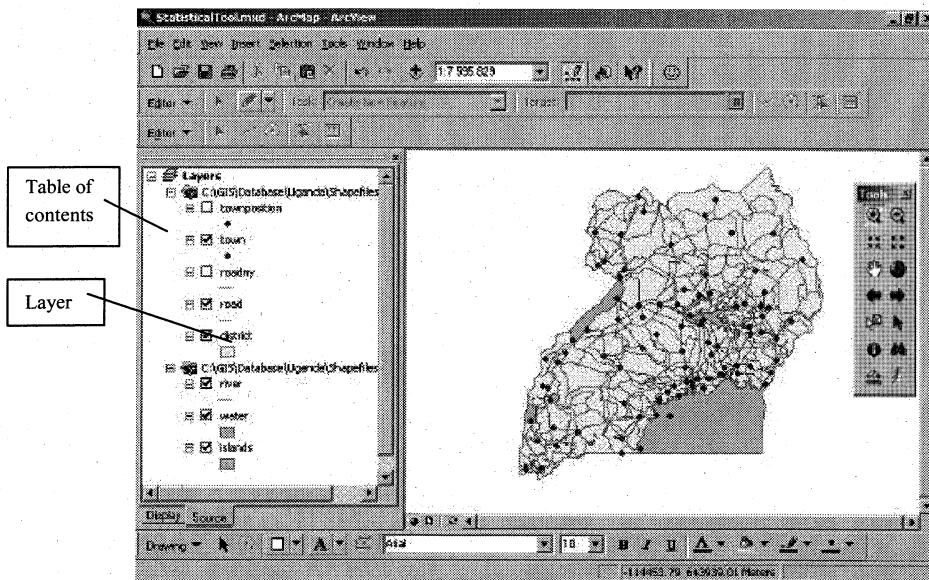
This Statistical tool contains functionalities for regression analyses with results shown as parameters and graphs. With these regression functionalities it is possible to analyse geographical data including data within social sciences. It is also possible to create new geographical data, such as distances between cities. The Statistical tool is found inside the program ArcMap.

STARTING ARCMAP

In this exercise you will learn how to open the GIS program ArcMap and get familiar with some elementary functions of the program.

- Double-click on the icon ArcMap on the desktop or click **Start > ArcGIS > ArcMap**.
- In the menu that shows after the program have started, mark the check-box **An existing map**.
- Open the map containing the Statistical tool by double-clicking on the text **Browse for maps**. Find the project in *C:\GIS\StatisticalTool\StatisticalTool.mxd* then click **Open**. (If the file *StatisticalTool.mxd* already appears in the menu window you can open the map by clicking on the file once and then click **Open**.)

You can now see a map of Uganda (to the right) and the table of contents (to the left). The picture below shows an example of the program. The layers in the map are presented in the table of contents and consist of for example roads, towns and rivers.



To see the statistical data you need to open the attribute table connected to the layer.

- Right-click on the layer *district*. Choose **Open attribute table**.

You can now see the attribute table containing district data. The figure below shows an example.

Appendix C

DISTRICT	REAL AREA	COUNTIES
Koido	13208	Dodoth, Jie, Labwor
Kitgum	16136	Agago, Aruu, Chua, Lamwo
Moyo	5006	East Moyo, West Moyo, Obongi
Arua	7830	Aringa, Ayivu, Koboko, Madi-Dkollo, Maracha, Terego, Wurra, Ar
Adjumani	1872	none
Gulu	11735	Aswa, Kilak, Nowoya, Omoro, Gulu Municipality
Moroto	14113	Bokora, Kadani, Matheniko, Pian, Upe, Morote Municipality
Nebbi	2891	Jonam, Okoro, Padyere
Lira	7251	Dokolo, Erute, Kyoga, Otuke, Moroto, Lira Municipality
Apac	6488	Kole, Kwania, Maruzi, Oyam
Katakwi	4905	Usuk, Amuria, Kapelebyong
Masindi	9326	Bujenje, Bulisa, Buruli, Kibanda
Soroti	10060	Amuria, Kaberamaido, Kalaki, Kapelebyong, Kastro, Serere, Soro
Hoima	5492	Bugahya, Buhaguzi
Nakasongola	3250	Buruli

To view the rest of the fields and rows you can use the scrollbars. To close the attribute table, click on the cross in the upper right corner.

Questions: How many rows does the district table contain? What is the name of the last field?

USING THE STATISTICAL TOOL

With the Statistical tool you can perform analyses on the data in the attribute tables. The aim of this exercise is to learn how to use the functions of the Statistical tool.

Notice that you cannot view the attribute tables at the same time as you use the Statistical tool. If you need to view the attribute tables after starting the Statistical tool you have to end the function by clicking **Cancel** in the Statistical tool menu.

If you want to know more about the theory behind the functions or how to use them you can always click **Help** in any of the menus of the Statistical tool.

- Start the Statistical tool by clicking the "smiley".



You can now see the different functions you can choose to perform.

- Click **Help** and read about the different functions. To close the help window, click the cross in the upper right corner.

New geographical attributes

This function creates two different new attributes - shortest distance and travelling distance.

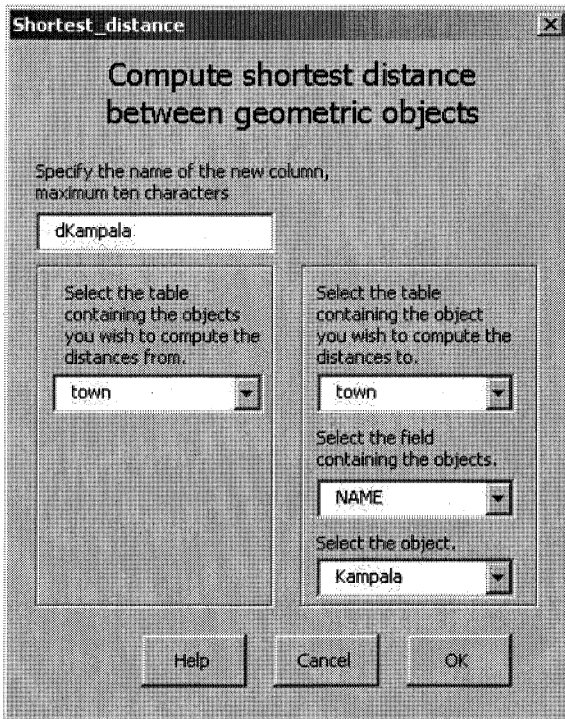
Shortest distance:

This function computes the shortest distance between several geometric objects and one geometric object. The function can for example be used for computing distances between several cities and one city or between several villages and a hospital. The result is shown in a

Appendix C

new field in the attribute table containing all these point attributes. In the following example you will compute the distance between Kampala and other cities in Uganda.

- Click **New geographical attributes**.
- In the next menu, click **Compute shortest distance between geometric objects**.
- In the next menu, start with giving the result field a suitable name. Notice that the name only can contain maximum 10 characters.
- In the left box all the attribute tables related to point objects are listed. Click the arrow and choose *town*. This table contains the cities from which the distances will be computed.
- In the first box to the right the same tables are listed. Choose *town*. This table contains the city (that you will choose later) to which the distances will be computed.
- In the second box to the right the fields of the chosen table are listed. Choose *Name*. This field contains the names of the cities.
- In the last box the names of the cities in the table *town* are listed. Scroll down and choose *Kampala*.
- Click OK.



You will now see a message that the distances are computed. After reading the message, click **OK**. The result is now found in the field you named in the table *town*.

- Right-click on the layer *town*. Click on **Open attribute table**. Look at the last field of the table.

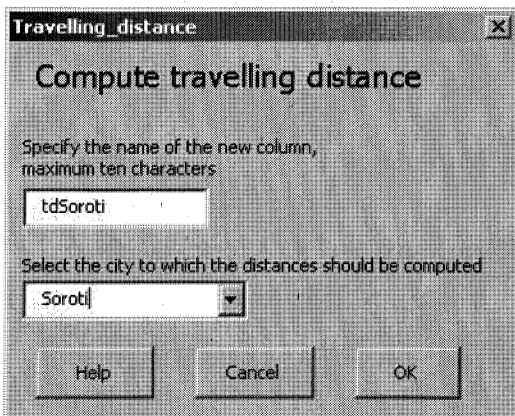
Question: How far is it from Busia to Kampala in meters?

Appendix C

Travelling distance:

This function computes the travelling distance between the district capital cities and one chosen city in Uganda. Travelling distance means the distance along roads. The result is shown in a new field in the *District_Town* table. In the following example you will compute the travelling distance between Soroti and the district capital cities in Uganda.

- Start the Statistical tool by clicking the "smiley".
- Click **New geographical attributes**.
- In the next menu, click **Compute travelling distance between cities**.
- In the next menu, start with giving the result field a suitable name, e.g. *tdSoroti*. Notice that the name only can contain maximum 10 characters.
- In the box the names of the cities in the table *town* are listed. Scroll down and choose *Soroti*.
- Click **OK**. This will now take about 5-10 minutes. You cannot use the program during this time.



You will then see a message that the distances are computed. After reading the message, click **OK**. The result is now found in the field you named in the table *District_Town*.

- Right-click on the layer *District_Town*. Click on **Open attribute table**. Look at the last field of the table.

Question: How far is it from Gulu to Soroti? What is the difference in meters between the travelling distance and the shortest distance between Soroti and Kampala?

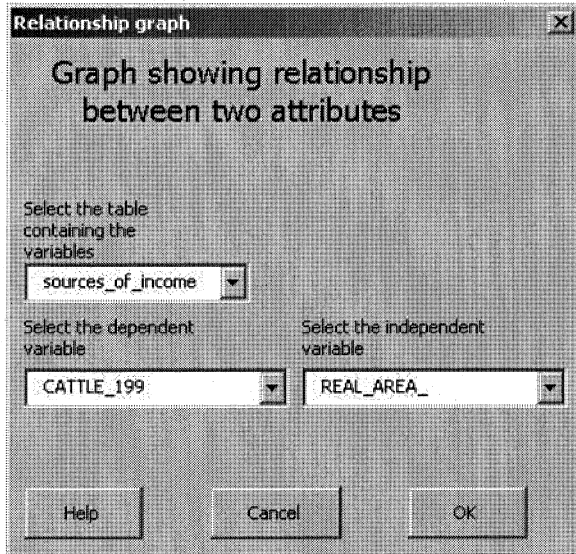
Graph showing relationship between two attributes

This function shows the relationship between two attributes by plotting the values of the two attributes in a graph and showing two statistical parameters; regression coefficient and standard error of regression coefficient (for an explanation of the parameters, use the help function of the Statistical tool). Be aware of that the two attributes must be found in the same table before using this function. In the following example we will examine the relationship between the area of each district and the number of cattle in each district.

- Start the Statistical tool by clicking the "smiley".
- Click **Graph showing relationship between two attributes**.

Appendix C

- In the next menu you first choose the table containing the attributes. Scroll down and choose *sources_of_income*.
- In the second box to the left you choose the dependent variable, i.e. the attribute you wish to explain. Choose *CATTLE_199*.
- In the box to the right you choose the independent variable, i.e. the explaining attribute. Choose *REAL_AREA_*.
- Click **OK**.



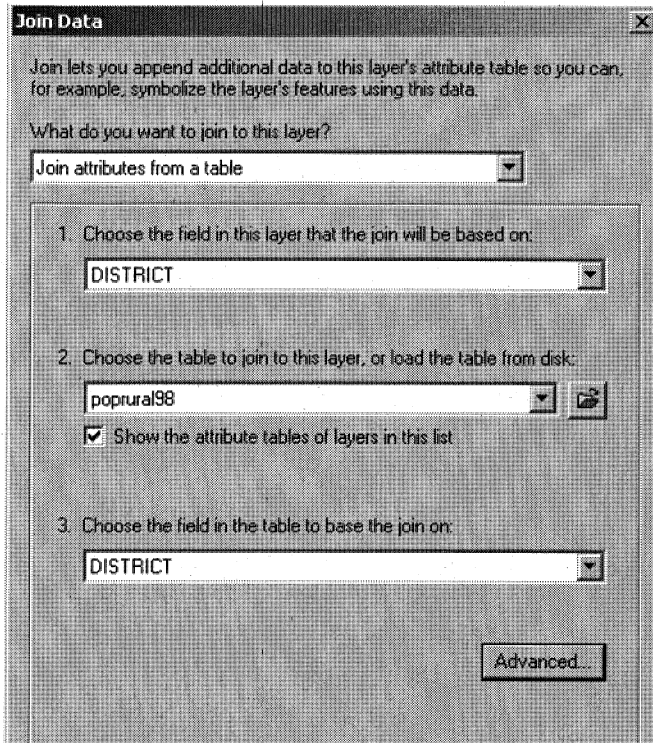
You can now see the result in a graph. You can make the graph window bigger by clicking a corner and dragging to a suitable size. Close the graph window by clicking the cross in the upper right corner. To know more about the theory and how to interpret the results you can start the Statistical tool again and use the Help function.

Joining tables - Graph showing relationship between two attributes

If the attributes you wish to investigate the relationship between are found in two different tables, these tables need to be joined first. When two tables are joined the information from them is gathered in one of them. In the following example we will first join the tables *sources_of_income* and *poprural98* and then examine the relationship between *CATTLE_199* and *RURAL98_1* (rural population).

- Right-click the layer *sources_of_income*. This is the table where the information from the two tables will be gathered. Choose **Joins and relates > Join...**
- At the top of the menu, choose **Join attributes from a table** (if it is not already chosen).
- In the next box you choose the field (from the table *sources_of_income*) that the join should be based on. Choose *DISTRICT*.
- In the next box you choose the table you wish to join with. Choose *poprural98*.
- In the last box you choose the second field that the join should be based on. Choose *DISTRICT*. This field has to contain the same data as the field chosen above.
- Click **OK**.

Appendix C



The information from the two tables is now gathered in the table *sources_of_income*. Open this table to see the result. You can see that the names of the fields have changed. They consist of two words separated by a dot. The first word is the name of the table where the information comes from. The second word is the old name of this field.

Question: Which table does the last field originally come from?

You can now investigate the relationship between e.g. *CATTLE_199* and *RURAL98_1* by using the graph function.

- Start the Statistical tool by clicking the "smiley".
- Click **Graph showing relationship between two attributes**.
- In the next menu you first choose the table containing the attributes. Scroll down and choose *sources_of_income*.
- In the second box to the left you choose the dependent variable. Choose *CATTLE199*.
- In the box to the right you choose the independent variable. Choose *RURAL98_1*.
- Click **OK**.

From the graph you can see that there is a weak statistical relationship between the two attributes.

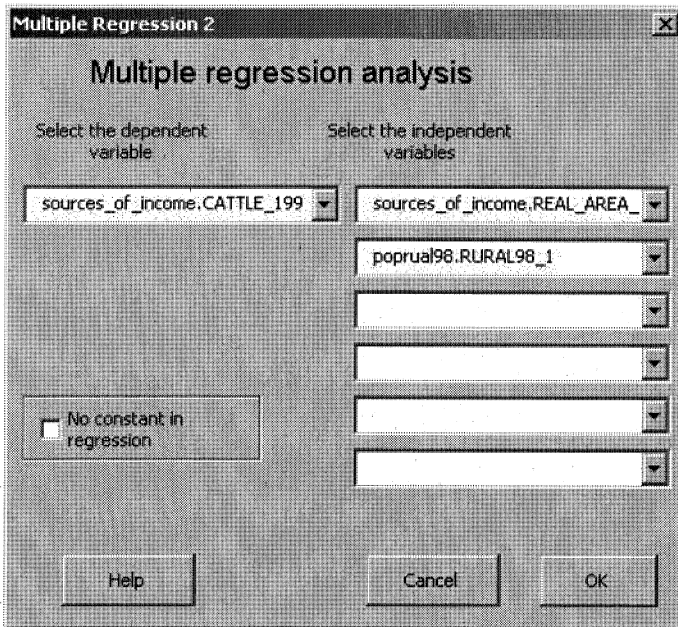
Multiple regression analysis

This function investigates the relationship between one attribute and several other attributes by computing the statistical parameters: regression coefficient, standard error of regression

Appendix C

coefficient, chi-square, r^2 , expectation value of residuals, and constant (for an explanation of the parameters, use the help function of the Statistical tool). Be aware of that the attributes must be found in the same table. If they are not, you need to join the tables (see example above). In the following example we will examine the relationship between the dependent variable *CATTLE_199* and the independent variables *RURAL98_1* and *REAL_AREA_* (area of each district).

- Start the Statistical tool by clicking the "smiley".
- Click **Multiple regression analysis**.
- In the next menu you first choose the table containing the attributes. Scroll down and choose *sources_of_income*. Click **OK**.
- In the next menu, to the left, you choose the dependent variable, i.e. the attribute you wish to explain. Choose *sources_of_income.CATTLE_199*.
- In the boxes to the right you choose the independent variables, i.e. the explaining attributes. Choose *sources_of_income.REAL_AREA_* in the first box and *poprural98.RURAL98_1* in the second.
- Click **OK**.



You will now see the results of the multiple regression. To know how to interpret the results, click **Help**. Click **OK** when you finished studying the results.

Note that a constant is normally used in regression within social sciences but if you wish to perform a regression without a constant you can check the **No constant in regression** box.

Notice! After using this function it is very important that you remove the join to restore the tables. The reason for this is to avoid unnecessary large and complicated tables.

- Right-click the layer of the table where the joined information is gathered (*sources_of_income*). Choose **Joins and relates > Remove join(s) > Remove all joins**.

Appendix C

A useful exercise

- Compute the shortest distance between *Entebbe* and the other cities in the table *town*, using the function **Compute shortest distance**. Name the new field *dEntebbe*.
- Join the tables *econom_ind_98* and *town*. Make sure that the information is gathered in the table *econom_ind_98*.
- Then investigate if there is a relationship between *dEntebbe* and *GDP_INDEX_* using the function **Graph showing relationship between attributes**.
- After seeing the result, remove the join from the table *econom_ind_98*!

Problems when using the statistical tool

If an error message, such as "Default" or Run-time error, appears click **End** and then end the Statistical tool. The reason to this message was probably that the data could not be used in this kind of operation. Control that data is used correctly and try again.

Appendix D: Test results for the statistical analysis tool

To assure that the functions in the statistical tool give a correct result different tests are performed on the functions. The calculations used for comparison are computed with Excel or by hand. The test cases and the test results are presented below.

New geographical attribute – Compute shortest distance

Test data is taken from the Uganda database. Results from the tests are compared with results computed by hand. As the exact coordinates for a town is hard to access, approximate coordinates from the map is used in the calculations.

Test case

Compute the shortest distance from Kampala to the cities Apac and Jinja in Uganda.

Results

From Kampala to Apac:

Result (tool): **185752** m

Result (by hand): $\sqrt{(356830 - 364041)^2 + (419039 - 233425)^2} = 185754m$

From Kampala to Jinja:

Result (tool): **71847** m

Result (by hand): $\sqrt{(434598 - 364041)^2 + (246979 - 233425)^2} = 71847m$

The tests show that the function works correctly (the small differences in the results are due to the collection of the coordinates done by hand).

New geographical attribute – Compute traveling distance

Test data is taken from the Uganda database. Results from the tests are compared with results computed by summarizing the line objects of the roads by hand.

Test case

Compute traveling distance from Kampala to the district capital cities Bundibugyo and Mukono.

Results

From Kampala to Bundibugyo:

Result (tool): **341493** m

Result (by hand):

$6785+7396+18981+28980+2367+63587+17364+14253+12193+65118+43277+3522+28469+29201=341493$

From Kampala to Mukono:

Result (tool): **21742** m

Result (by hand): $5274+4784+11684=21742$

The tests show that the function works correctly.

Graph showing relationship between two attributes

Test data is made up or taken from the Uganda database. Results from the tests are compared with results computed in Microsoft Excel analysis tool. Notice that the entire result table from Excel is not presented in this context.

Test cases

1. Standard regression on goat population 1998 as dependent variable and sheep population 1998 as independent variable. Data is missing for some districts.
2. Standard regression on made up variables x and y.

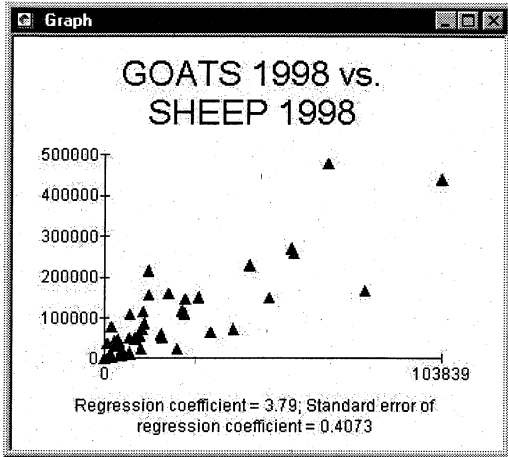
y	x
25	1
42	3
60	5
35	7
45	3
45	4
24	3
43	1
24	6

3. Standard regression on the area of the districts in Uganda as dependent variable and perimeter as independent variable.

Results from the tests

Test case 1:

Results from the statistical tool



Result from Excel

<i>Regression Statistics</i>	
Multiple R	0.830136
R Square	0.689125
Adjusted R Square	0.681154
Standard Error	61127.7
Observations	41

ANOVA

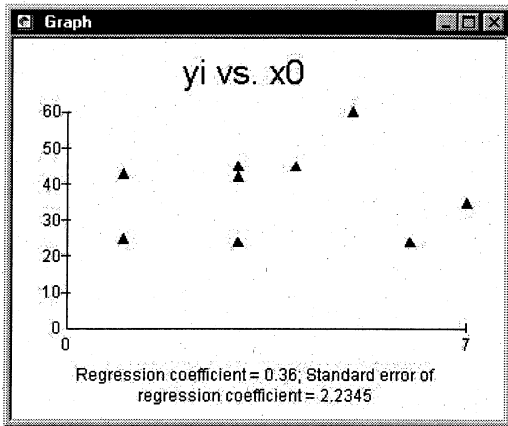
	<i>df</i>	<i>SS</i>
Regression	1	3.23E+11
Residual	39	1.46E+11
Total	40	4.69E+11

Coefficients Standard Error

Intercept	22495.97	12939.48
X Variable 1	3.787909	0.407391

Test case 2:

Results from the statistical tool



Results from Excel

<i>Regression Statistics</i>	
Multiple R	0.061241
R Square	0.00375
Adjusted R Square	-0.13857
Standard Error	13.02972
Observations	9

ANOVA

	<i>df</i>	<i>SS</i>
Regression	1	4.473856
Residual	7	1188.415
Total	8	1192.889

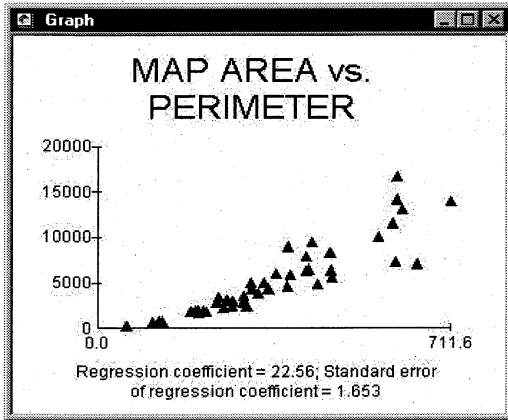
Coefficients Standard Error

Intercept	36.78105	9.273425
X Variable 1	0.362745	2.234578

Appendix D

Test case 3:

Results from the statistical tool



Results from Excel

Regression Statistics	
Multiple R	0.901355
R Square	0.812442
Adjusted R Square	0.80808
Standard Error	1730.177
Observations	45

ANOVA

	df	SS
Regression	1	5.58E+08
Residual	43	1.29E+08
Total	44	6.86E+08

	Coefficient	Standard Error
Intercept	-2667.98	642.8825
X Variable 1	22.56264	1.65321

The tests show that the function works correctly.

Multiple regression analysis

Test data is made up or taken from the Uganda database. Results from the tests are compared with results computed in Microsoft Excel analysis tool. When using the multiple regression function there is an option to use, or not use a constant in the calculation. Notice that the entire result table from Excel is not presented in this context.

Test cases

1. Multiple regression on three independent variables using a constant. Map area is used as dependent variable and population of cattle, goat and sheep 1998 as independent variables. Data is missing for some districts.
2. Multiple regression with three independent variables and not using a constant. y_i is used as dependent variable and x_0 , x_{1i} and x_{2i} as independent variables. The test data is made up.

y_i	x_0	x_{1i}	x_{2i}
25	1	5	6
42	3	7	7
60	5	12	9
35	7	5	8
45	3	8	5
45	4	7	5
24	3	4	8
43	1	7	7
24	6	4	2

Appendix D

3. Multiple regression with six independent variables not using a constant. The dependent variable is GDP index and the independent variables are map area, cattle, goats, sheep, perimeter and real area. All data is taken from the Ugandan database. Data is missing for some districts. Notice that the variables in the regression analysis not is realistic.
4. Multiple regression with six independent variables using a constant. y_i is used as dependent variable and $x_0, x_1, x_2i, x_3i, x_4i$ and x_5i as independent variables. The test data is made up.

y_i	x_0	x_1i	x_2i	x_3i	x_4i	x_5i
25	1	5	6	5	9	4
42	3	7	7	7	6	6
60	5	12	9	3	4	87
35	7	5	8	43	23	5
45	3	8	5	7	6	43
45	4	7	5	8	7	2
24	3	4	8	2	2	4
43	1	7	7	12	11	65
24	6	4	2	6	32	2

5. Multiple regression with two independent variables using a constant. Map area is used as dependent variable and GDP index and economic index as independent variables. The test is taken from the Ugandan database.

Results

Test case 1:

Results from the statistical tool

The screenshot shows a window titled "Result Multiple Regression" with the following data:

Statistics of the multiple regression analysis			
	Name	Coefficient	Standard error
Dependent parameter	MAP_AREA		
Independent parameter 1	CATTLE_199	2,040.57	7,316.07
Independent parameter 2	GOATS_1998	22,681.46	14,506.04
Independent parameter 3	SHEEP_1998	15,687.37	56,973.66
Independent parameter 4			
Independent parameter 5			
Independent parameter 6			
Constant		0.19	0.52

Chi-square	7.9926409970609	Help
r2	-0.52	
expectation value of residuals	4,711,876,542.04	OK

Results from Excel

Regression Statistics

Multiple R	0.2880697
R Square	0.0829841
Adjusted R Square	0.0065662
Standard Error	3.66E+09
Observations	40

ANOVA

	df	SS
Regression	3	4.36419E+19
Residual	36	4.82264E+20
Total	39	5.25906E+20

Coefficients Standard Error

Intercept	4.08E+09	838694761.2
X Variable 1	-3349.119	5789.961044
X Variable 2	15591.213	11362.47451
X Variable 3	-22993.96	44964.60817

Appendix D

Test case 2:

Results from the statistical tool

Result Multiple Regression			
Statistics of the multiple regression analysis			
	Name	Coefficient	Standard error
Dependent parameter	yt		
Independent parameter 1	x0	0.76	0.66
Independent parameter 2	x1t	4.84	0.64
Independent parameter 3	x2t	0.52	0.67
Independent parameter 4			
Independent parameter 5			
Independent parameter 6			
Constant			

Chi-square	117.53	Help
r2	0.90	OK
expectation value of residuals	4.4258709124118	

Results from Excel

Regression Statistics	
Multiple R	0.94945931
R Square	0.90147298
Adjusted R Square	0.70196398
Standard Error	4.42590446
Observations	9

ANOVA

	df	SS
Regression	3	1075.357107
Residual	6	117.5317815
Total	9	1192.888889

	Coefficients	Standard Error
Intercept	0	#N/A
X Variable 1	0.75777448	0.661278329
X Variable 2	4.83812238	0.636607399
X Variable 3	0.52189573	0.668792462

Test case 3:

Results from the statistical tool

Result Multiple Regression			
Statistics of the multiple regression analysis			
	Name	Coefficient	Standard error
Dependent parameter	econom_ind_98.GI		
Independent parameter 1	sources_of_income	0.00	0.00
Independent parameter 2	sources_of_income	0.00	0.00
Independent parameter 3	sources_of_income	0.00	0.00
Independent parameter 4	sources_of_income	0.00	0.00
Independent parameter 5	sources_of_income	0.00	0.00
Independent parameter 6	sources_of_income	0.00	0.00
Constant			

Chi-square	0.91	Help
r2	-4.57	OK
expectation value of residuals	0.1635992233549	

Results from Excel

Regression Statistics	
Multiple R	65535
R Square	-3.82775
Adjusted R Square	-4.56712
Standard Error	0.152514
Observations	40

ANOVA

	df	SS
Regression	6	-0.62704
Residual	34	0.790857
Total	40	0.163815

	Coefficients	Standard Error
Intercept	0	#N/A
X Variable 1	-6.1E-11	1.39E-11
X Variable 2	1.72E-07	2.43E-07
X Variable 3	-7.6E-07	4.89E-07
X Variable 4	3.82E-06	1.85E-06
X Variable 5	1.74E-06	2.36E-07
X Variable 6	5.94E-06	5.28E-06

Appendix D

Test case 4:

Results from the statistical tool

Result Multiple Regression			
Statistics of the multiple regression analysis			
	Name	Coefficient	Standard error
Dependent parameter	y1		
Independent parameter 1	x0	0.21	1.56
Independent parameter 2	x11	4.74	1.69
Independent parameter 3	x21	-1.38	1.72
Independent parameter 4	x31	0.33	0.24
Independent parameter 5	x41	-0.34	0.51
Independent parameter 6	x51	0.01	0.13
Constant		15.06	17.28

Chi-square	49.83	Help
r2	0.96	
expectation value of residuals	4.99	OK

Results from Excel

Regression Statistics	
Multiple R	0.97889243
R Square	0.95823039
Adjusted R Square	0.83292155
Standard Error	4.99131776
Observations	9

ANOVA

	df	SS
Regression	6	1143.062383
Residual	2	49.82650595
Total	8	1192.888889

	Coefficients	Standard Error
Intercept	15.0582245	17.28399856
X Variable 1	0.21201952	1.559948097
X Variable 2	4.73651509	1.89217332
X Variable 3	-1.3806319	1.722551812
X Variable 4	0.3344445	0.238444842
X Variable 5	-0.3387164	0.507930075
X Variable 6	0.01141865	0.134096986

Test case 5:

Results from the statistical tool

Result Multiple Regression			
Statistics of the multiple regression analysis			
	Name	Coefficient	Standard error
Dependent parameter	sources_of_income		
Independent parameter 1	econom_ind_98.GI	-14,571,92	8,026,201,1
Independent parameter 2	econom_ind_98.EI	107,606,72	56,296,946
Independent parameter 3			
Independent parameter 4			
Independent parameter 5			
Independent parameter 6			
Constant		7,096,779,1	4,153,035,1

Chi-square	5.6577170487946	Help
r2	0.18	
expectation value of residuals	3,670,252,699.48	OK

Results from Excel

Regression Statistics	
Multiple R	0.4190684
R Square	0.1756183
Adjusted R Square	0.1363621
Standard Error	3.67E+09
Observations	45

ANOVA

	df	SS
Regression	2	1.20527E+20
Residual	42	5.65772E+20
Total	44	6.86298E+20

	Coefficients	Standard Error
Intercept	7.097E+09	4153020227
X Variable 1	-1.457E+10	8026198559
X Variable 2	107605688	56297057.55

Appendix D

The results from these tests show that the tool does not work correctly in all cases. By these tests and further testing we have come to the conclusion that multiple regression does not work when the independent variables contain very large numbers. In test case 1 and 3 numbers larger than 100 000 are found in the independent variables. These tests give incorrect results. In test case 5 numbers larger than 100 000 are used as dependent variable but the independent variables contain small numbers. This test gives a correct result. The problem with the function is probably due to the communication between the C++ code and the VBA code. This problem needs to be further investigated.