

Geospatial analysis in Scala

Roxana Tesileanu

roxana.te@web.de
INCDS, Romania

October 2017

Contents

1	Introduction	1
1.1	Geoprocessing using GDAL	1
1.2	Overview of the software development process of Scala projects .	2
1.2.1	Using VIM as an editor for Scala code	3
1.2.2	Using SBT to create and manage Scala projects	6
1.2.3	Managing dependencies with SBT	9
2	Geoprocessing vector and raster data	12
2.1	Types of spatial data	12
2.2	Reading vector data	13
2.3	Georeferencing data	15
2.3.1	Spatial reference systems in OGR	16
2.3.2	Creating OGR geometries and vector layers	23
2.3.3	Reprojecting OGR geometries and vector layers	29
2.4	Overlay and proximity analyses	29
2.5	Writing vector data	30
2.6	Reading raster data	30
2.7	Pixels resizing	30
2.8	Moving window analyses	30
2.9	Map algebra	30
3	Spatial data analysis	30
3.1	Introduction into spatial data analysis	30
3.2	Using LAPACK for building multivariate statistical models . . .	32
3.3	From datasets to probability models	36
3.4	Commonly used probability density functions	41
3.5	Computing statistical matrices for multivariate analyses	44

4	Variable-based multivariate analyses	44
4.1	Predicting with Multiple Linear Regression (MLR) analysis . . .	44
4.2	Examining group differences with Multivariate Analysis of Vari- ance (MANOVA)	44
4.3	Vizualizing MANOVA residuals: spatial residual map	44
4.4	Classifying with Discriminant Function Analysis (DFA)	45
5	Object-based multivariate analyses	45
5.1	Implementing the k-Nearest Neighbors classifier	45
6	Combined approaches: the MANOVA-KNN pipeline	46
6.1	The learning pipeline for reducing error	46
6.2	Dealing with local variation: the adaptive KNN	47
7	Conclusions	47
	Appendices	51
	Appendix A Point coordinates for code snippet 3	52
	Appendix B Scala source file: readPointCoord.scala	55
	Appendix C Scala source file: createGeoms.scala	56
	Appendix D Scala source file: createVLayer.scala	57
	Appendix E Scala source file: usingLapack2.scala	58

1 Introduction

The aim of the present paper is to investigate how the Geospatial Data Abstraction Library (GDAL) [1] can be used in Scala [2], for performing the main geospatial analysis tasks: manipulating vector and raster data (geoprocessing) and geospatial data analysis.

1.1 Geoprocessing using GDAL

Using a programming language for geospatial analysis allows you to customize your analyses instead of being limited to what the software user interface allows. This is one of the most important advantages of open source software [3]. This work uses the GDAL open source library [1] developed at the Open Source Geospatial Foundation (OSGeo) (www.osgeo.org), and, the Scala programming language developed at École polytechnique fédérale de Lausanne (EPFL) (<http://www.scala-lang.org/>) to perform geospatial data analyses. GDAL was written in C and C++ and has bindings for several languages (Java, Perl and Python) [3].

In order to use GDAL, you need to install it on your machine, and for its import in Scala you need to install its Java bindings along with it. For installation details you can look at the GDAL homepage <http://www.gdal.org/>, download GDAL and follow the instructions for building from source, which might not be an easy task, depending on your operating system. Thanks to the efforts of the UbuntuGIS team (<https://wiki.ubuntu.com/UbuntuGIS>), on Ubuntu, the installation procedure of GDAL and its bindings is done rapidly. Firstly, you need to add the ubuntuGIS PPA, which offers the official stable UbuntuGIS packages, to your system (<https://launchpad.net/ubuntu/+archive/ubuntu/ppa>). This is done with the commands:

```
sudo add-apt-repository ppa:ubuntugis/ppa
sudo apt-get update.
```

Next, you install GDAL on your machine with the commands [4] [5] (<http://www.sarasafavi.com/installing-gdal-gr-on-ubuntu.html>, <https://packages.ubuntu.com/source/trusty/gdal>):

```
sudo apt-get install libproj-dev, gdal-bin, libgdal-dev, libgdal-doc
sudo apt-get update.
```

Finally, you add the Java bindings to your GDAL package (<https://launchpad.net/ubuntu/+source/proj>):

```
sudo apt-get install libgdal-java, libproj-java.
```

In order to import GDAL in Scala, you have to add its jar to the project's

classpath. An easy way of managing dependencies of a Scala project is to use SBT (<http://www.scala-sbt.org/>) [6] [7]. In this way you can take advantage of the most convenient way to place the gdal jar to the project's classpath, namely, to place a copy of it into the lib directory of the Scala project, now that the actual installation has already taken place.

Section 1.2 offers an introduction in SBT and into the software development process of Scala projects in general, in order to get ready to start with Section 2 which will offer a background in geoprocessing: manipulating vector data (reading and writing files of different vector data formats, georeferencing data, performing overlay and proximity analyses) and continuing with manipulating raster data (reading and writing files of different raster data formats, resizing pixels, performing moving window analyses and map algebra). Then, in Sections 3, 4, 5 and 6 the main approaches of spatial data analyses will be presented, based on variable- and object-based multivariate analyses.

1.2 Overview of the software development process of Scala projects

This subsection offers an introduction into the software development process of Scala projects pointing to some of the available learning resources ¹.

In order to implement software development projects, one should consider which tools to choose for the different components of the development system, which according to Rehman and Paul (2003) [8] are the following:

- the hardware platform,
- the operating system,
- editors,
- compilers, assemblers and debuggers,
- version control system, and,
- bug tracking.

For implementing Scala projects, working with the Simple Build Tool (SBT) (<http://www.scala-sbt.org>) condenses the list of the components of a development system, as it unites different steps in one tool (compiling, building, testing and debugging). Another advantage of using SBT is the opportunity of working interactively in Scala REPL, a tool for evaluating expressions in Scala (<https://docs.scala-lang.org/overviews/repl/overview.html>) within a SBT session with all dependencies and project classes on the classpath, to develop bits

¹Section 1.2 is taken from [6]

of code which can then be inserted in the editor of choice.

It is also important to understand that software development "is not just writing code", but rather a more comprehensive process [8]. Rehman and Paul (2003) describe it as comprising the following steps: requirement gathering, writing functional specifications, creating architecture and design documents, implementation and coding, testing and quality assurance, software release, documentation, and, support and, development and release of new features. Each project starts with a requirement analysis which investigates the real-world need of the final product. Which functions should the new software carry in real-world problem solving within the specified domain? Further, functional specifications are declared to state the functionality of a software product at an abstract level "defining its input/output behavior". On the basis of the functional specifications, an architecture of the product is created. The architecture "defines the different components of the product and how they interact with each other", without providing the explicit details on how they should be implemented to reach the desired functionality. This happens at the design stage, when design documents are created which define each individual component to the level of functions and procedures. Using the design documents and development tools (SBT and editor) the code is then implemented and tested. There are more types of testing: unit testing (testing one part or one component of the product using test cases to test functionality of this part of the software), sanity testing to check if all components compile, regression or stress testing to check the long-term behavior of the product when used continuously over a period of time, and functional testing using test cases built on functional specifications. If a bug (an anomaly) is found it must be reported and fixed. The documentation includes: technical documentation developed during the development process, technical documentation prepared for technical support staff, and end-user manuals and guides. Finally, the last stage of the life cycle of a software development project is the support and release of new versions depending on requirements.

The present subsection aims at introducing the reader to the tools needed to put in motion the development system of Scala projects on Linux, with emphasis on the Ubuntu distribution. The following subsubsections will cover brief introductions in using the VIM editor as an editor for Scala code and using SBT to create and manage Scala projects, also addressing the issue of dependency management.

1.2.1 Using VIM as an editor for Scala code

VIM is an editor which can be used from within the Linux command line and can be adapted to support the editing of Scala code by installing the vim-scala package of Derek Wyatt found at <https://github.com/derekwyatt/vim-scala>. Under Linux you can install VIM using the package tools at hand (like e.g. dpkg, apt-get, aptitude, rpm or yum) depending on the installed distribution. In Ubuntu the command for installing VIM is: "\$ sudo apt install vim". Afterwards, you

open the command line and launch VIM by typing "\$ vim <ENTER>". The following instructions will make the first VIM session easy to pass through:

- because VIM starts in command mode you can change to insert mode with the "i" key;
- to exit insert mode and return to command mode, press the <ESC> key;
- to write the open file to the hard drive use the command ":w filename" (in command mode) or just ":w" if the file already has a name;
- to save the changes type ":" in command mode;
- to exit VIM type ":wq" to quit and save changes, or ":q!" to discard changes (both in command mode);
- pressing <ESC> will not just place you in command mode but also cancel an unwanted and partially completed command in command mode.

The following paragraphs will help you navigate through the VIM file, and, process text (append, delete, copy, paste, search and replace, edit multiple files and get help).

To move the cursor you use: "l" (or right arrow), "h" (or left arrow), "j" (meaning going down one line), or "k" (meaning going up one line). Moving inside the current line is possible with "0" (to go to the beginning of the current line) and "\$" (to go to the end of the current line). To go to the end of the file, i.e. to the last line of the file you use "<SHIFT>g" and to go to the beginning of the file, i.e. to the first line you use "gg". To go to a specific line you use "numbergg" (for example: 2gg to go to the second line).

When you are in command mode, you can start inserting text by using the "i" key to place you in insert mode, but there are also other commands which can place you in insert mode so that you can start appending text. One of them is the "a" key, which lets you append text right at the spot where the cursor is placed, or the "A" key which lets you append text directly at the end of the line. You can also use the "o" key to open a line below the current line or "O" to open a line above the current line.

You can undo changes by using the "u" key and redo changes the using "<CTRL>r".

Deleting text in command mode goes with (but not only) the following commands:

- the "x" key deleted the current character; "3x" will delete the current character and the next two characters
- the "dd" command deletes the current line; "6 d d" will delete the current line and the next five lines

- the "dG" command will delete to the end of the file
- the "d\$" command will delete to the end of the line
- the "d0" command will delete to the beginning of the line.

The commands based on "d" not just delete text but also copy it to a paste buffer, which can be later recalled with the "p" command to paste the contents of the buffer after the cursor or the "P" command to paste the contents of the buffer before the cursor.

Cutting, copying and pasting text is done more traditionally with the "y" command (which stands for "yank", i.e. copy). To copy the current line you use "yy" ("6yy" means copy the current line and the next five lines). To copy to the end of the line use "y\$" and to the beginning of the line use "y0". You can join lines with "J".

Searching and replacing is done within the line and within the whole file. Searching within the line is done with "f" from "find" (for example "fa" will move the cursor to the next occurrence of "a"). You can also use the substitute command for a line (for example: while in command mode "s:caar/car<ENTER>" means substitute "caar" with "car", and replaces the first occurrence of the searched word). The ":#,#s/big/small/g" replaces "big" with "small" within a range of lines (":" starts the command, "#" stands for a line number and "g" stands for "global"). To search the entire file use "/" followed by the searched word and the <ENTER> key. You can go to the next occurrence of the searched word with the "n" (from "next") command. To search and replace over the entire file, i.e. globally, you can use the command ":%s/you/YOU/g" (":" starts the command, "%s" find and substitute, "g" globally). The search can also be done with options. For example entering ":set ic" will allow finding also combinations with capitals. To disable ignoring case enter ":set noic". You can also enable the "is" option (with the ":set is" command), which can be disabled with ":set nois".

To edit multiple files you can open them together from the beginning (for example "\$ vim file1, file2, file3) or you open additional files after you started VIM with the command ":e additional_file". The command "buffers" displays a list of files under editing. You can switch between files with the command ":buffer 1" for example to go to the first file listed by the ":buffers" command, or ":buffer 2" to go to the second file listed by the ":buffers" command. To copy content from one file to another you copy with yank, switch with buffer and paste. This works only if the files were opened together or with the ":e" command. Opening different VIM windows will not allow copying between files in this way. Instead, you have to use <SHIFT><CTRL><C> to copy from one file and <SHIFT><CTRL><V> to paste into another file. The <SHIFT><CTRL><V> command lets you paste text from all kind of sources (from documents, web sites, etc.). To insert an entire file into another you can use the ":r other_file_name_to_be_inserted_into_the_current" command, where "r"

stands for "retrieving".

To select text to paste you can also use the visual mode by typing "v" in command mode and then move the cursor to the end of your selection. The selected text will be highlighted. Then press "y" to yank it and then "p" to paste it.

A useful debugging command is the matching parentheses command. You can find out if the corresponding parenthesis or bracket is missing with "%" by moving the cursor on the first parenthesis or bracket and typing "%".

To execute an external command, i.e. a command as if you were in the command line, with the ":! your_command" command. For example the ":! ls <ENTER>" command will list the objects of the working directory.

Finally for getting help in VIM you type ":help <ENTER>". To close the help window use ":q". To find help on a specific command use ":help some_command", for example type ":help user-manual" to get to the user manual. For an interactive tutorial for VIM you can also use the "VIM Tutor" initially created by Pierce and Ware, which can be started in the command line with the command "\$ vimtutor <ENTER>". A short introduction in VIM can be also found in Shotts (2009) [9], a lecture which gently introduces readers to the Linux command line in general.

1.2.2 Using SBT to create and manage Scala projects

In Section 1.2.2 the reader will find out how to install SBT and use it to create and manage Scala projects. Furthermore, the issue of integrating external (managed) and internal (unmanaged) dependencies and of forking the Java Virtual Machine (JVM) will be discussed.

Installing SBT You can use the Scala REPL either directly from within the command-line, initiating a REPL session with the command "\$ scala", or, by launching a SBT-session with the command "\$ sbt" and then from within SBT launch the Scala REPL using the command "> console". On Ubuntu you can install the Scala language using the apt command ("sudo apt install scala") with no additional commands. Check the package management tool you use to get the same result. You can open the Scala REPL and type in some expressions; to exit type ":q" (VIM is useful after all). For guidance, check the book of Jason Swartz "Learning Scala" [10] and the book of Mark Lewis "Introduction to Programming and Problem Solving Using Scala" [11] for a great introduction to the Scala language.

Installing SBT requires a look at the SBT homepage (<http://www.scala-sbt.org/>) in order to get the four commands needed to install it using the command line. At the time the present document is written the commands for Ubuntu are:


```
$ echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a /etc/apt/sources.list.d/sbt.list
```

```
$ sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 2EE0EA64E40A89B84B2DF73499E82A756
```

```
$ sudo apt-get update
```

```
$ sudo apt-get install sbt
```

If you are using rpm distributions visit the SBT download page (<http://www.scala-sbt.org/download.html>) and look for the appropriate installing commands.

Next, open a Scala REPL session in SBT using the commands introduced at the beginning of this section and type in some expressions. You exit the console with `":q"` and the command `"> exit"` gets you out of the SBT and back to the command-line.

Creating a Scala projet Now, we will create a Scala project called `Scala_Playground` in order to continue the playground series of Linux introductions started by William Shotts (2009) in his book "The Linux Command Line".

Open the terminal with `<CTRL-ALT-t>`. It starts directly in your home directory. Next, create a directory called `Scala_Playground` (`"$ mkdir Scala_Playground <ENTER>"`). Check if the directory was created with the command `"$ls <ENTER>"` which lists the contents of the directory in which you are placed. Further, switch to the `Scala_Playground` directory (`"$ cd Scala_Playground <ENTER>"`) and create a part of the inner structure of the current directory by adding another directory called `src` with its subdirectory `main`, with its subdirectory `scala`. This is done with the command `"$ mkdir -p src/main/scala"` which creates the entire chain of directories. Check it with `"$ls"` and `"$cd"` commands, return to the `Scala_Playground` directory with `"cd -"`.

Next, in the `Scala_Playground` directory, create a second directory called `project`. The `ls` command should print now for the `Scala_Playground` two contents: `project` and `src`.

The final step in creating the Scala project is to create two files: one directly in the `Scala_Playground` directory called `build.sbt` and the second in the `Scala_Playground/project` directory called `build.properties`. Open VIM when you are in the `Scala_playground/project` directory and type in:

```
sbt.version = 0.13.15
```

Go in the command mode with <ESC> and save the changes and give the open file a name: `:w build.properties`. Exit vim with `:q`.

The build.properties file sets the version of SBT.

Next, when you are in the Scala_Playground directory open VIM and edit the basic settings of the Scala project: project name, Scala version used, and the dependencies you need. For the moment we will ignore dependencies and create the basic Scala project. Type in the file opened in VIM the following two lines:

```
name := "Scala_Playground"
scalaVersion := "2.11.10"
```

The above two lines are not in Scala, they are in SBT's own language [7]. Now, get into the command line mode with <ESC> and save the changes and give the file the name build.sbt (`:w build.sbt`). Exit VIM with `:q`. You should now have three objects in the Scala_Playground directory: the build.sbt file, the project directory and the src directory.

You can now launch SBT while you are in the Scala_Playground directory using the `"$ sbt"` command. You can see the SBT's output as it uses the information from the two edited files and creates the desired Scala project. When the output is finished you get to see the SBT prompt `">"`. The project was created successfully. Now, launch the Scala REPL with the command `"> console"`. When the REPL is open you get to see the Scala prompt `"scala>"`. You can type in some expressions to see if it works. Type `:q` to get back to SBT and `"exit"` to get back to the command-line.

You can see that the Scala_Playground directory has gained some additional directories: the Scala_Playground/target directory and the Scala_Playground/project/target directory. These are created by SBT when it creates the Scala project, and serve as locations for the output of SBT tasks (e.g. output following compilations or packaging actions).

In the src/main/scala directory you should place your Scala codes in form of scala files.

Some basic SBT tasks Open SBT. When you see the SBT prompt `">"` type in your first SBT command: `"help"`. This generates an output which you can use to get further information about SBT and about the configuration of your project (command `"about"`, command `"settings"`). The command `"reload"` reloads the project in the current directory, which is useful after you've changed

the project settings (for example after adding dependencies) or after you've changed or added source files. It does the sanity check as it compiles the source files and outputs the errors encountered in the compilation process, accompanied by messages indicating possible reasons. It is very often helpful to try to read them in order to debug the code.

The command "tasks" is also listed in the help output. The SBT tasks show much of the functions SBT can take over. Type in "tasks" and check the default list of SBT tasks:

- starting the Scala REPL with the command "console"
- compiling source files with the command "compile"
- running an application with the command "run"
- testing the code executing all tests
- and many more.

For a detailed introduction in SBT you should check the SBT documentation at <http://www.scala-sbt.org/documentation.html> and read the book of Joshua Suereth and Matthew Farwell (2016) "SBT in action: the Simple Scala Build Tool".

1.2.3 Managing dependencies with SBT

Depending on the field in which a software developer is active, it is often necessary to use field-specific third-party Scala libraries which offer additional functionality. For data analysis, for example, the best known libraries are Spark (<http://spark.apache.org/>) which is a general engine for large-scale data processing (see [12] for an introduction in Spark), ScalaNLP (<http://www.scalanlp.org/>) which is a suite of Machine Learning and numerical computing libraries, or Figaro (<https://www.cra.com/work/case-studies/figaro>) which is a probabilistic programming language for probabilistic modeling based on Scala (see [13] for an introduction in Figaro).

For GUI programming, the best known libraries are Scala.swing (<http://www.scala-lang.org/api/2.9.1/scala/swing/package.html>) which is actually a Scala package but needs to be added to the dependencies list in order to be imported and used in Scala projects, and ScalaFX (<http://www.scalafx.org/>) which is based on Scala and sits on top of JavaFX [11] (see [11] for a general introduction in Scala and GUI programming using Scala).

For many software developers, the main reason of using SBT is because it offers an easy management of dependencies from repositories. Also the packaging and integration of libraries in form of .jar files into the project is possible. Internal libraries in form of .jar files are called unmanaged dependencies and you should

place them into the project's lib directory, which should be created inside the project's directory if you intend to use libraries in form of .jar files for a Scala project. External libraries are found on repositories and are called managed dependencies, as they are configured in the build definition and downloaded automatically from repositories [14].

Using external Scala libraries The first step towards integrating the functionality of these external libraries into our Scala code is to find out their repository coordinates for SBT use. One way to find out this piece of information is to search for them in the Maven Repository (<https://mvnrepository.com/>). The next step is to add the Maven dependency of the external library to the Scala project [12]. This is done by including an additional setting called "libraryDependencies" in the project's build.sbt file. This setting can take the form of a Sequence for including more than one external library dependency and in general looks like the following:

```
libraryDependencies ++= Seq(  
  groupId %% artifactID % revision,  
  groupId %% artifactID % revision  
)
```

where the exact library coordinates are found in the Maven Repository.

Figure 1 presents the contents of the build.sbt file of a Scala project called Scala_Playground (created in section 3.2) which includes the managed dependencies for enabling the use of Spark's MLlib (Spark's library for Machine Learning) and of scalaFX (library for GUI programming) and, the unmanaged dependency for enabling the use of javaFX package needed for GUI programming using scalaFX.

```

fork := true

name := "Scala_Playground"
scalaVersion := "2.11.10"

libraryDependencies ++= Seq(
  "org.apache.spark" % "spark-core_2.11" % "2.1.1" % "provided",
  "org.apache.spark" % "spark-mllib_2.11" % "2.1.1" % "provided",
  "org.scalafx" %% "scalafx" % "8.0.92-R10"
)

unmanagedJars in Compile += Attributed.blank(
  file("/usr/lib/jvm/java-8-openjdk-amd64/jre/lib/ext/jxrt.jar")
)

```

Figure 1: The contents of the build.sbt file of a Scala project with three external dependencies and one internal dependency (if more than one .jar file is needed, then they should be placed in the project's lib directory instead of specifying them in the build.sbt file).

Using internal Scala libraries Sometimes it is necessary to use internal libraries in form of .jar files. Jar files are automatically created from compiled source files by setting the variable "exportJars" of the build.sbt file to true (exportJars := true). The process of .jar creation is called packaging. The newly created .jar file is placed in the project's target/scala-2.11 directory. If you want to make use of an internal library in form of a .jar file in your Scala code, it has to be added to classpath so that it becomes available during compilation, testing, running, and when using Scala REPL ([14]). This can be achieved by overriding the settings variable called "unmanaged Jars" and indicating the path to the .jar file you want to use:

```
unmanagedJars in Compile := Attributed.blank(file("mypath/*.jar")).
```

The above setting is equivalent to the use of "-cp" when starting the Scala interpreter from outside SBT ("\$ scala - cp mypath/*.jar ").

If more than one .jar file is used, then you should create a new directory called lib inside the Scala project's directory in which you can place the necessary jars. The lib directory will serve as common location and will make the jars available to the compiler and interpreter, all the jars added to it being placed on the project classpath ([14]).

Using a forked JVM Running applications with the run task or using the REPL in SBT with the console task disposes the code to use the same JVM instance as SBT ([14]). There are some situations though, when running user code in the same JVM as SBT causes problems, and demand that the code is run in a different instance of JVM. In SBT this is called forking the JVM [7]. You can enable forking by setting the variable called "fork" in your build.sbt file to true ("fork := true"). The situations requiring JVM forking appear either when the user code employs System.exit (which shuts down the JVM), or when it starts multiple threads (some of which may not terminate until JVM terminates) like in the case of creating a GUI [14]. You can read more on forking in Suereth and Farwell (2016) and in the SBT documentation (<http://www.scala-sbt.org/1.x/docs/Forking.html>). The build.sbt file in fig. 1 enables forking in order to ensure the save close of GUIs created by user code.

2 Geoprocessing vector and raster data

The following subsections will offer a background in geoprocessing, starting with manipulating vector data (reading and writing files of different vector data formats, georeferencing data, and, performing overlay and proximity analyses), and continuing with manipulating raster data (reading and writing files of different raster data formats, resizing pixels, performing moving window analyses and map algebra).

2.1 Types of spatial data

Spatial data are divided in two categories: vector data and raster data. Vector data provide information about distinct features in space, i.e. different distinct items of interest, and are made up of points, lines and polygons [3]. The features of interest could be for example:

- roads, rivers, road networks, hidrological networks, country boundaries, city boundaries as examples of features represented by lines,
- mountain peaks, volcano peaks, weather stations, restaurants, as examples of features represented by points, and
- lakes, oceans, ownership status as examples of features represented by polygons.

Features have attributes attached to them such as the name of the individual observations (for example the wheather stations's name) and other recorded variables (like for example different concentrations of air pollutants, temperature or wind regime for each individual weather station). As it can be noticed, the multiple attributes which can be attached to features, can be of different types, and they actually represent different types of recorded variables (they might be dicrete or continuous numerical variables or categorical variables).

On the other hand, raster data provide information about characteristics of interest which take the form of a continuum like gradients, with no distinct boundaries. They are represented as two- or three-dimensional arrays of data values which form grids of values [3]. Because they can cope well with gradients, they capture local variation more easily than vector geometries, and are used in digital elevation models (DEMs). Also because the data source is pixel-based (e.g. aerial photos, satellite imagery) they can be used in vegetation mapping.

2.2 Reading vector data

The main objective of vector data analysis is to investigate relationships between features, by overlapping them on another or measuring distances between them [3]. A typical example for vector analyses is the investigation of GPS-collared wildlife to see the direction of travel, distances covered and how they interact with man-made features like roads [3].

In order to perform such vector-based analyses, we need to be able to read, edit and write vector data. This kind of functionality is offered by the OGR Simple Features Library for geoprocessing vector data, which is included in GDAL.

At this point it is noted that the Scala code relating to using the GDAL functionality introduced in this document has its origins in the Python code written by Chris Garrard in her book "Geoprocessing with Python" (2016). The main reason for the transition towards using Scala for geospatial analysis is the use of Scala's functional nature for the further processing of geodata, by using higher-order functions.

There are many different types of vector data formats. Among the most widely used ones are: the ESRI shapefile, the GeoJSON file, or the SpatiaLite or PostGIS databases. The ESRI shapefile format requires a minimum of three binary files, each of which serves a different purpose: geometry information is stored in .shp and .shx files, and attribute values are stored in a .dbf file. You need to make sure they are all grouped in the same folder, because they work together [3]. The GeoJSON format is used mainly for web-mapping applications and is a plain text file which can be easily examined. The GeoJSON format consists of a single file. Vector data can also be stored in relational databases with spatial extensions. The most widely used spatial extensions are SpatiaLite (for SQLite databases) and PostGIS (for PostgreSQL). You can check other vector data formats supported by GDAL at http://www.gdal.org/ogr_formats.html.

The OGR package of GDAL contains the classes used for geoprocessing vector data. The OGR Java Application Programming Interface (API) [1] (<http://gdal.org/java/>) lists them all. Among them there are the classes: Driver, DataSource, Layer, Geometry and Feature. In order to handle vector geodata with OGR, we need to understand how the geospatial information is organized in OGR. The spatial vector data is stored in a data source (for example a shapefile, a GeoJSON file, or a SpatiaLite or PostGIS database). This data source object

can have one or more layers, one for each dataset contained in the data source. Many vector formats, such as shapefiles can only contain one dataset, thus have one layer, but others like SpatiaLite can contain multiple datasets, thus have multiple layers. Each layer contains a collection of features, which holds the geometries (like for example points, lines, polygons) and their attributes [3].

The first step in accessing any vector data is to open the data source. For this you need to use a driver specific to the data format. Each vector data format has its own driver, which is used to read and write a particular format [3]. In order to make the configured OGR drivers available we call the `RegisterAll()` function at the beginning of the analysis. The `RegisterAll()` function is placed in the package `OGR` of the GDAL library, in the class `ogr`, so we need to call it using `org.gdal.ogr.ogr.RegisterAll()` (or, we can also import the class to save typing, but it remains unclear where the function resides inside GDAL).

Code snippet 1: accessing a shapefile using OGR in Scala You can find a shapefile on the internet at one of the sources indicated under <http://gisgeography.com/best-free-gis-data-sources-raster-vector/> and try the following snippet with your shapefile using the Scala REPL in SBT while you are in the Scala project's directory:

```
org.gdal.ogr.ogr.RegisterAll()

val dataSource = org.gdal.ogr.ogr.Open("example.shp")
dataSource: org.gdal.org.DataSource = org.gdal.ogr.DataSource@305c6b70

dataSource.GetLayerCount
res1: Int = 1

val lyr = dataSource.GetLayer(0)
lyr: org.gdal.ogr.Layer = org.gdal.ogr.Layer@305ca330

lyr.GetName
res2: String = example

lyr.GetFeatureCount
res3: Long = 927

val feat0 = lyr.GetFeature(0)
feat0 : org.gdal.ogr.Feature = org.gdal.ogr.Feature@3164b9a0

dataSource.delete()
```


Code snippet 1: Explanation In order to access the vector geodata, we make the OGR drivers available with `RegisterAll()`. Then, we create a variable called `dataSource` in which we store the `DataSource` object. We obtain it by opening the shapefile (or any other OGR file format) with `Open()`. We check how many layers (i.e. datasets) are available in the `DataSource` object by calling `GetLayerCount` on it. Further, we retrieve the first layer (which has the index 0), with `GetLayer(0)`, obtaining a variable called `lyr` in which we store it. You can check its name or the number of features it contains, with `GetName` or `GetFeatureCount`. You can see the available functions on the `lyr` with the help of autocompletion. Autocompletion works for example by typing `lyr.` followed by a `<TAB>`. This lists all the methods available for that object. We continue, by retrieving the first feature (which has the index 0) of the layer called `lyr` with `GetFeature(0)`. You can check with autocompletion the methods available on it. At the end of the code snippet we close the data source.

Geodata would actually be "normal" data without georeferencing. In the next subsection, we learn how to deal with spatial reference systems of already georeferenced features, and how to construct our own geometries and features and how to georeference them.

2.3 Georeferencing data

- investigate georeferenced data
- construct new geometries and features
- georeference new features

In order to be able to locate some coordinates on a map, you need to know what spatial reference system is used for the coordinates and what spatial reference system uses the map. If they are not the same, you need to perform transformations from one spatial reference system to another. Georeferencing the data means adding the information regarding the spatial reference system used when defining the features.

A spatial reference system is made of three components [3]:

- a coordinate system,
- a datum and,
- a projection.

The set of coordinates is provided by a coordinate system, the datum specifies the model used to represent the curvature of the earth and a projection is used to transform the three-dimensional globe to a two-dimensional map. A set of coordinates is represented as set of two pieces of information: the latitude and the longitude. Positive latitude values are north of the equator, and positive longitudes are east of the prime meridian. Multiple methods exist for specifying latitude and longitude coordinates: decimal degrees (DD), degrees decimal minutes (DM) and degrees minutes seconds (DMS) [3]. But, if you don't specify

the datum used, the same set of latitude and longitude coordinates can refer to slightly different locations, because different datums represent different ellipsoids of different shapes. One of the most widely used datums is the World Geodetic System, last revised in 1984. This datum, also called WGS84, is also used by the Global Positioning System (GPS). WGS84 has a global coverage. But most datums model the curvature of the earth in a more localized area (a continent or a country). A datum designed for an area will not work well elsewhere [3]. Sometimes the difference between two datums can be of hundreds of meters for the same set of coordinates.

Projections convert the coordinates of a point on the globe to the coordinates of a point on a two-dimensional plane. In doing so, it creates distortions. The type of distortion depends on how the conversion is done. If you would like to preserve local shapes, you should use conformal projections, like for example Universal Transverse Mercator (UTM) [3]. To keep the amount of area the same, you should use equal-area projections, like for example the Lambert equal-area or Gall-Peters projections. Thus, sometimes using geographic coordinates (lat/lon), is not appropriate, depending on the aims of your analysis, and you need to choose a projection instead. Also, note that projections are not tied to specific datums, so knowing the projection of your data is not enough. You also have to know the datum used.

You can search for spatial reference systems on the epsg.io website under <http://epsg.io/>. The SRSs for Romania for example can be found under the link: <http://epsg.io/?q=Romania>.

2.3.1 Spatial reference systems in OGR

OGR provides ways of storing and converting the information regarding the spatial reference system (SRS) used for vector data in its package called OGR Spatial Reference (OSR). You can find out the SRS used by a georeferenced layer, by calling the `GetSpatialRef` function on it. If the layer is not georeferenced it returns null.

```
lyr.GetSpatialRef
res10: org.gdal.osr.SpatialReference =
GEOGCS["GCS_WGS_1984",
  DATUM["WGS_1984",
    SPHEROID["WGS_84",6378137,298.257223563]],
  PRIMEM["Greenwich",0],
  UNIT["Degree",0.017453292519943295],
  AUTHORITY["EPSG","4326"]]
```

```
lyr.GetSpatialRef
res14: org.gdal.osr.SpatialReference =
PROJCS["ETRS_1989_LAEA",
  GEOGCS["GCS_ETRS_1989",
```

```

DATUM["European_Terrestrial_Reference_System_1989",
    SPHEROID["GRS_1980",6378137.0,298.257222101]],
PRIMEM["Greenwich",0.0],
UNIT["Degree",0.0174532925199433]],
PROJECTION["Lambert_Azimuthal_Equal_Area"],
PARAMETER["False_Easting",4321000.0],
PARAMETER["False_Northing",3210000.0],
PARAMETER["longitude_of_center",10.0],
PARAMETER["latitude_of_center",52.0],
UNIT["Meter",1.0]]

```

The first spatial reference from above is not a projected SRS, because it has a GEOCS entry only, without a PROJCS one. But, we can still see that the datum used is WGS1984, the spheroid used is WGS84, the unit used for the set of coordinates is degree and the authority giving the ID code is EPSG (short for European Petroleum Survey Group). The second spatial reference from above is a projected one. It has a PROJCS entry, the projection used is the LAEA (Lambert Azimuthal Equal-Area projection). The datum is ETRS 1989 (European Terrestrial Reference System 1989). The SRS is thus the ETRS1989/LAEA. The unit of measure is 1.0 m.

The function `GetSpatialRef` called on a layer returns a `SpatialReference` object. Georeferenced data already have such an object defined, if not the `GetSpatialRef` function returns null. In order to create a `SpatialReference` object for your layers and geometries you need to firstly create an empty `SpatialReference` object, then you have to import into the empty `SpatialReference` object the information on the SRS you want to use, turning it into a valid `SpatialReference` object.

Code snippet 2: creating a `SpatialReference` object using OGR in Scala In this code I will use the ETRS1989/LAEA, the Pulkovo1942(58)/Stereo70 and the WGS84/Pseudo-Mercator SRS. The first one is the SRS for all Europe and it is used for statistical mapping at all scales and other purposes where true area representation is required (<http://spatialreference.org/ref/epsg/3035/>), the second SRS is for Romania and is used in large and medium scale topographic mapping and engineering surveys (<http://spatialreference.org/ref/epsg/3844/>), and the third is used for rendering maps in GoogleMaps, OpenStreetMap, Bing a.o. (<http://epsg.io/3857>). Google Earth uses WGS84 with geographic coordinates (lat/lon), unprojected, with EPSG code 4326 (<https://gis.stackexchange.com>) (see the first SRS example at page 7, section 2.3.1).

```

val newSR = new org.gdal.osr.SpatialReference()
newSR: org.gdal.osr.SpatialReference =

```

```

newSR.ImportFromEPSG(3035)
res1: Int = 0

newSR
res2: org.gdal.osr.SpatialReference =
PROJCS["ETRS89 / LAEA Europe",
  GEOGCS["ETRS89",
    DATUM["European_Terrestrial_Reference_System_1989",
      SPHEROID["GRS 1980",6378137,298.257222101,
        AUTHORITY["EPSG","7019"]],
      TOWGS84[0,0,0,0,0,0],
      AUTHORITY["EPSG","6258"]],
    PRIMEM["Greenwich",0],
    AUTHORITY["EPSG","8901"]],
    UNIT["degree",0.0174532925199433,
      AUTHORITY["EPSG","9122"]],
    AUTHORITY["EPSG","4258"]],
  PROJECTION["Lambert_Azimuthal_Equal_Area"],
  PARAMETER["latitude_of_center",52],
  PARAMETER["longitude_of_center",10],
  PARAMETER["false_easting",4321000],
  PARAMETER["false_northing",3210000],
  UNIT["metre",1,
    AUTHORITY["EPSG","9001"]],
  AUTHORITY["EPSG","30...

val newSR2 = new ogr.gdal.osr.SpatialReference()
newSR2: org.gdal.osr.SpatialReference =

newSR2.ImportFromProj4("+proj=laea +lat_0=52 +lon_0=10 +x_0=4321000
+y_0=3210000 +ellps=GRS80 +units=m +no_defs ")
res3: Int = 0

newSR2
res4: org.gdal.osr.SpatialReference=
PROJCS["unnamed",
  GEOGCS["GRS 1980(IUGG, 1980)",
    DATUM["unknown",
      SPHEROID["GRS80",6378137,298.257222101],
      TOWGS84[0,0,0,0,0,0],
      PRIMEM["Greenwich",0],
      UNIT["degree",0.0174532925199433]],
    PROJECTION["Lambert_Azimuthal_Equal_Area"],
    PARAMETER["latitude_of_center",52],
    PARAMETER["longitude_of_center",10],
    PARAMETER["false_easting",4321000],

```

```

    PARAMETER["false_northing",3210000],
    UNIT["Meter",1]]

val newSR3 = new org.gdal.osr.SpatialReference()
newSR3: org.gdal.osr.SpatialReference =

newSR3.ImportFromWkt("""PROJCS["ETRS89 / ETRS-LAEA",
|   GEOGCS["ETRS89",
|   DATUM["European_Terrestrial_Reference_System_1989",
|   SPHEROID["GRS 1980",6378137,298.257222101,
|   AUTHORITY["EPSG","7019"]],
|   AUTHORITY["EPSG","6258"]],
|   PRIMEM["Greenwich",0,
|   AUTHORITY["EPSG","8901"]],
|   UNIT["degree",0.01745329251994328,
|   AUTHORITY["EPSG","9122"]],
|   AUTHORITY["EPSG","4258"]],
|   UNIT["metre",1,
|   AUTHORITY["EPSG","9001"]],
|   PROJECTION["Lambert_Azimuthal_Equal_Area"],
|   PARAMETER["latitude_of_center",52],
|   PARAMETER["longitude_of_center",10],
|   PARAMETER["false_easting",4321000],
|   PARAMETER["false_northing",3210000],
|   AUTHORITY["EPSG","3035"],
|   AXIS["X",EAST],
|   AXIS["Y",NORTH]]""")
res9: Int = 0

```

```

newSR3
res10: org.gdal.osr.SpatialReference =
PROJCS["ETRS89 / ETRS-LAEA",
  GEOGCS["ETRS89",
    DATUM["European_Terrestrial_Reference_System_1989",
      SPHEROID["GRS 1980",6378137,298.257222101,
        AUTHORITY["EPSG","7019"]],
        AUTHORITY["EPSG","6258"]],
      PRIMEM["Greenwich",0,
        AUTHORITY["EPSG","8901"]],
        UNIT["degree",0.01745329251994328,
          AUTHORITY["EPSG","9122"]],
          AUTHORITY["EPSG","4258"]],
        UNIT["metre",1,
          AUTHORITY["EPSG","9001"]],
        PROJECTION["Lambert_Azimuthal_Equal_Area"],

```

```

    PARAMETER["latitude_of_center",52],
    PARAMETER["longitude_of_center",10],
    PARAMETER["false_easting",4321000],
    PARAMETER["false_northing",3210000],
    AUTHORITY["EPSG","3035"],
    AXIS["X",EAST],
    AXIS["...

val roSR1 = new org.gdal.osr.SpatialReference()
roSR1: org.gdal.osr.SpatialReference =

roSR1.ImportFromEPSG(3844)
res11: Int = 0

roSR1
res12: org.gdal.osr.SpatialReference =
PROJCS["Pulkovo 1942(58) / Stereo70",
  GEOGCS["Pulkovo 1942(58)",
    DATUM["Pulkovo_1942_58",
      SPHEROID["Krassowsky 1940",6378245,298.3,
        AUTHORITY["EPSG","7024"]],
      TOWGS84[2.329,-147.042,-92.08,0.309,-0.325,-0.497,5.69],
      AUTHORITY["EPSG","6179"]],
    PRIMEM["Greenwich",0,
      AUTHORITY["EPSG","8901"]],
    UNIT["degree",0.0174532925199433,
      AUTHORITY["EPSG","9122"]],
      AUTHORITY["EPSG","4179"]],
    PROJECTION["Oblique_Stereographic"],
    PARAMETER["latitude_of_origin",46],
    PARAMETER["central_meridian",25],
    PARAMETER["scale_factor",0.99975],
    PARAMETER["false_easting",500000],
    PARAMETER["false_northing",500000],
    UNIT["metre",1,
    A...

val roSR2 = new org.gdal.osr.SpatialReference()
roSR2: org.gdal.osr.SpatialReference =

roSR2.ImportFromWkt("""PROJCS["Pulkovo 1942(58) / Stereo70",
|   GEOGCS["Pulkovo 1942(58)",
|   DATUM["Pulkovo 1942(58)",
|     SPHEROID["Krassowsky 1940",6378245.0,298.3,
|     AUTHORITY["EPSG","7024"]],
|     TOWGS84[33.4,-146.6,-76.3,-0.359,-0.053,0.844,-0.17326243724756094],

```

```

|     AUTHORITY["EPSG", "6179"]],
|     PRIMEM["Greenwich", 0.0,
|     AUTHORITY["EPSG", "8901"]],
|     UNIT["degree", 0.017453292519943295],
|     AXIS["Geodetic latitude", NORTH],
|     AXIS["Geodetic longitude", EAST],
|     AUTHORITY["EPSG", "4179"]],
|     PROJECTION["Oblique Stereographic",
|     AUTHORITY["EPSG", "9809"]],
|     PARAMETER["central_meridian", 25.0],
|     PARAMETER["latitude_of_origin", 46.0],
|     PARAMETER["scale_factor", 0.99975],
|     PARAMETER["false_easting", 500000.0],
|     PARAMETER["false_northing", 500000.0],
|     UNIT["m", 1.0],
|     AXIS["Northing", NORTH],
|     AXIS["Easting", EAST],
|     AUTHORITY["EPSG", "3844"]]]")
res16: Int = 0

roSR2
res17: org.gdal.osr.SpatialReference =
PROJCS["Pulkovo 1942(58) / Stereo70",
GEOGCS["Pulkovo 1942(58)",
DATUM["Pulkovo 1942(58)",
SPHEROID["Krassowsky 1940", 6378245.0, 298.3,
AUTHORITY["EPSG", "7024"]],
TOWGS84[33.4, -146.6, -76.3, -0.359, -0.053, 0.844, -0.17326243724756094],
AUTHORITY["EPSG", "6179"]],
PRIMEM["Greenwich", 0.0,
AUTHORITY["EPSG", "8901"]],
UNIT["degree", 0.017453292519943295],
AXIS["Geodetic latitude", NORTH],
AXIS["Geodetic longitude", EAST],
AUTHORITY["EPSG", "4179"]],
PROJECTION["Oblique Stereographic",
AUTHORITY["EPSG", "9809"]],
PARAMETER["central_meridian", 25.0],
PARAMETER["latitude_of_origin", 46.0],
PARAMETER["scale_factor", 0.99975],
PAR...

val googleMapsSR = new org.gdal.osr.SpatialReference()
googleSR: org.gdal.osr.SpatialReference =

googleMapsSR.ImportFromEPSG(3857)

```

```
res18: Int = 0
```

```
googleMapsSR
res19: org.gdal.osr.SpatialReference =
PROJCS["WGS 84 / Pseudo-Mercator",
  GEOGCS["WGS 84",
    DATUM["WGS_1984",
      SPHEROID["WGS 84",6378137,298.257223563,
        AUTHORITY["EPSG","7030"]],
        AUTHORITY["EPSG","6326"]],
      PRIMEM["Greenwich",0,
        AUTHORITY["EPSG","8901"]],
        UNIT["degree",0.0174532925199433,
          AUTHORITY["EPSG","9122"]],
          AUTHORITY["EPSG","4326"]],
        PROJECTION["Mercator_1SP"],
        PARAMETER["central_meridian",0],
        PARAMETER["scale_factor",1],
        PARAMETER["false_easting",0],
        PARAMETER["false_northing",0],
        UNIT["metre",1,
          AUTHORITY["EPSG","9001"]],
        AXIS["X",EAST],
        AXIS["Y",NORTH],
        EXTENSION["PROJ4","+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0
+lon_0=0.0 +x_0=0.0 +y...
```

Code snippet 2: Explanation In code snippet 2 we've created new empty SpatialReference objects, in which we've imported information on the used SRS using more sources: the EPSG code, the PROJ4 string, and the WKT string. There are also more ways of importing. You can inspect them with autocompletion on an empty SpatialReference object.

In this section we've seen it is important to know the SRS used for your geodata. You can create a SRS using an empty SpatialReference object and importing the information needed using its EPSG code, WKT or PROJ4 string. The important fact to know, if your data comes from the GPS, is that this system uses WGS84 unprojected (EPSG code 4326). If you use Google Earth (EPSG code 4326) and try to map them on Google Maps (EPSG code 3857) you don't have the same SRS. You need to be able to make transformations between different SRSs. We will do this in the following sections, after we learn how to create individual geometries and entire layers.

2.3.2 Creating OGR geometries and vector layers

In this section I will use points looked up on Google Maps (so, using WGS84 / Pseudo-Mercator) and stored in a .csv file to create OGR geometries like Points, Lines or Polygons and their multi-versions. Then I will create a new empty georeferenced vector layer by means of a driver which will create the data source in which the empty vector layer will be stored. Then I will create the attribute fields for the empty layer, which will be stored in the layer definition. Then, I will create the features which will contain the previously created geometries and their attribute fields. Finally, I will insert the features into the new layer.

Code snippet 3: Reading point coordinates from a .csv file Suppose you have looked up your point coordinates on Google Maps or Google Earth and know what features you will have and what geometries you should use for your project. You should now create a .csv file (from converting either an EXCEL or a LIBRE OFFICE CALC file with "Save As" and choosing .csv format), in which you have three columns: the first one is the longitude (E/W), the second one is the latitude (N/S), and the third one represents the ID number of your features (from 1 to the last feature). I've created such a file with point coordinates from Google Maps [15] which we will use for this code snippet. You can inspect it at https://github.com/RoxanaTesileanu/multivariate_analyses/blob/master/DeepLearning/pointcoord.csv. Now, according to the point coordinates in the pointcoord.csv file, we're going to create 8 features for our project:

- the first group of points (ID 1) is for the first polygon which will be the representation of the first habitat patch (habitatPatch1) and is the Tampa Hill in Brasov,
- the second group of points (ID 2) is for the first road which will be the representation of a road in Racadau Area in Brasov,
- the third group of points (ID 3) is for the second road which will be the representation of a road in Carpatilor Area in Brasov,
- the fourth group of points (ID 4) is for the third road which will be the representation of a road in Noua Quarter in Brasov,
- the fifth group of points (ID 5) is for the second polygon which will be the representation of the second habitat patch (habitatPatch2) and is the Noua Forest Area,
- the sixth group of points (ID 6) is for the representation of a GPS-track for an imaginary radio-collared bear individual,
- the seventh group of points (ID 7) is for a polygon which will be the representation of a third habitat patch (habitatPatch3) toward Postavaru Peak in Poiana Brasov,

- the eighth group of points (ID 8) is for the fourth polygon which will be the representation for our study area.

Of course, these features are created for educational purposes. I haven't included as many points as necessary for a detailed representation of the items. Also, if you include more features (additional roads, more GPS-tracks, more habitat patches, etc.) then the results should be useful for further research purposes. The aim of this tutorial is to show how you can construct your features from point coordinates and how to perform spatial analyses on them whatever number of features you consider appropriate for your own purposes and whatever number of point coordinates you use for their representation.

After this general presentation of the data we're going to use, we can continue and read the points into Scala [2] (parse them):

```
import scala.io._
val source = Source.fromFile("pointcoord.csv")
val data = source.getLines.map(_.split(",")).toArray
val dataHP1 = data.filter(_(2) == "1")
dataHP1.length
val dataRoad1 = data.filter(_(2) == "2")
dataRoad1.length
val dataRoad2 = data.filter(_(2) == "3")
dataRoad2.length
val dataRoad3 = data.filter(_(2) == "4")
dataRoad3.length
val dataHP2 = data.filter(_(2) == "5")
dataHP2.length
val dataGPSTrack = data.filter(_(2) == "6")
dataGPSTrack.length
val dataHP3 = data.filter(_(2) == "7")
dataHP3.length
val dataStArea = data.filter(_(2) == "8")
dataStArea.length
val pointsHP1 = dataHP1.map(i => (i(0).toDouble, i(1).toDouble))
pointsHP1.length
val pointsRoad1 = dataRoad1.map(i => (i(0).toDouble, i(1).toDouble))
val pointsRoad2 = dataRoad2.map(i => (i(0).toDouble, i(1).toDouble))
val pointsRoad3 = dataRoad3.map(i => (i(0).toDouble, i(1).toDouble))
val pointsHP2 = dataHP2.map(i => (i(0).toDouble, i(1).toDouble))
val pointsHP3 = dataHP3.map(i => (i(0).toDouble, i(1).toDouble))
val pointsGPSTrack = dataGPSTrack.map(i => (i(0).toDouble, i(1).toDouble))
val pointsStArea = dataStArea.map(i => (i(0).toDouble, i(1).toDouble))
```

Code snippet 3: Explanation We read files in Scala with `scala.io.Source`. Because of that we import `scala.io._` at the beginning of the parsing code. We then create a value called `source` to store the data source into. Then, we create a value called `data` in which we access the lines of the `.csv` file. Each row of the `.csv` is a line. Because the `.csv` file is comma delimited, we split each line at `" , "` and then we transform each line to an `Array` of strings. For each group of points we create a value in which we store them (i.e. `dataHP1`, `dataRoad1`). This is done by means of applying a filter on the whole data, to get only the points with the ID number we want. We then check the length of each group of data calling `length` on it (which shows how many `Arrays` (i.e. point coordinates) it contains. Because the data is provided as `Array[String]` we must convert the groups of data to `Array[Double]`, which is done using the `map` function. If you encounter big problems in following this code snippet you can grasp to the book of Jason Swartz "Learning Scala" [10] and then build up your skills with the book of Marc Lewis "Introduction into Programming and Problem Solving Using Scala" [11] which is also accompanied by videos on youtube (<https://www.youtube.com/playlist?list=PLLMXbkbDbVt9MIJ9DV4ps-trOzWtphYO>).

Code snippet 4: Creating points and multipoints, lines and multilines, and, polygons and multipolygons In this code snippet I will use the values of the point groups (i.e. `pointsRoad1`, `pointsHP1`, etc.) from code snippet 3 to build OGR geometries. In order to use the the point groups, I've stored the code from snippet 3 in an object called `ReadPointCoordFromFile` found in the `readPointCoord.scala` file available at the link https://github.com/RoxanaTesileanu/multivariate_analyses/blob/master/DeepLearning/readPointCoord.scala. You should download the `.scala` file and store it in the Scala project's directory, to make it available in REPL. Reload your project in SBT, restart the console to get to the Scala REPL and try out the following code lines:

```
:load readPointCoord.scala
import ReadPointCoordFromFile._
pointsRoad1
val currentPosition = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPoint)
currentPosition.AddPoint(pointsRoad1(0)._1, pointsRoad1(0)._2)
currentPosition
currentPosition.GetX
currentPosition.GetY
val multiPointHP1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiPoint)
val geomsForMP = for (p<- pointsHP1) yield new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPoint)
val zippedGeomsPointsHP1 = geomsForMP.zip(pointsHP1)
for (z <- zippedGeomsPointsHP1) ( (z._1).AddPoint((z._2)._1, (z._2)._2))
for (z <- zippedGeomsPointsHP1) multiPointHP1.AddGeometry(z._1)
```

```

pointsRoad1
val road1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad1) road1.AddPoint(p._1, p._2)
road1.GetGeometryCount
road1.AddPoint(pointsRoad1(0)._1, pointsRoad1(0)._2)
road1.GetGeometryCount
road1.IsEmpty
road1.GetPointCount
val road2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad2) road2.AddPoint(p._1, p._2)
road2.GetPointCount
val road3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad3) road3.AddPoint(p._1, p._2)
road3.GetPointCount
val gpsTrack1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsGPSTrack) gpsTrack1.AddPoint(p._1, p._2)
val multiLineLines = Array(road1, road2, road3, gpsTrack1)
val multiLineEx = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiLineString)
for (l <- multiLineLines) multiLineEx.AddGeometry(l)
val habitatPatch1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p <- pointsHP1) habitatRing.AddPoint(p._1, p._2)
habitatRing.GetPointCount
habitatPatch1.AddGeometry(habitatRing)
habitatPatch1.CloseRings()
habitatPatch1.IsValid
val habitatPatch2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p <- pointsHP2) habitatRing2.AddPoint(p._1, p._2)
habitatRing2.GetPointCount
habitatPatch2.AddGeometry(habitatRing2)
habitatPatch2.CloseRings()
val habitatPatch3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p <- pointsHP3) habitatRing3.AddPoint(p._1, p._2)
habitatRing3.GetPointCount
habitatPatch3.AddGeometry(habitatRing3)
habitatPatch3.CloseRings()
val stArea = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val ringStArea = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p <- pointsStArea) ringStArea.AddPoint(p._1, p._2)
stArea.AddGeometry(ringStArea)
stArea.CloseRings()
stArea.IsValid
val multiPolygonEx = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiPolygon)
val multiPolyPolys = Array(habitatPatch1, habitatPatch2, habitatPatch3, stArea)

```

```
for (poly <- multiPolyPolys) multiPolygonEx.AddGeometry(poly)
```

Code snippet 4: Explanation In the above code we've created several OGR geometries:

- a point called `currentPosition`,
- a multipoint called `multiPointHP1`,
- a series of lines called `road1`, `road2`, `road3`, and `gpsTrack1`,
- a multiline called `multiLineEx`,
- a series of polygons called `habitatPatch1`, `habitatPatch2`, `habitatPatch3`, `stArea`, and
- a multipolygon called `multiPolygonEx`.

In general we can create an OGR geometry by creating a new empty geometry which will be populated with points by calling the `AddPoint()` function on them. The multi-geometry versions can be created by adding geometries to an empty multi-geometry by calling `AddGeometry()` on it. The polygons require a ring of points. In the above examples we only have one ring for each polygon. We could also create two rings (an inner and an outer ring) in order to create polygons with holes. For more details see [3].

Next, we will return to vector layers. In the following code snippet we will create a vector layer called `RoadsAndGPSTracks`, using the multiline geometry created in code snippet 4. For georeferencing we will use the WGS84 / Pseudo-Mercator SRS for which we've already created an instance in code snippet 2 called `googleMapSR`.

Code snippet 5: Creating a georeferenced vector layer Every new layer needs a name, a spatial reference and a geometry type. The new vector layer we will create in this code snippet will be called `RoadsAndGPSTracks`, it will be in WGS84 / Pseudo-Mercator and will be of type multiline. The layer `RoadsAndGPSTracks` will be created on a data source called `"rdgps"`, which will be on its turn created by means of a ESRI shapefile driver. After we create the empty vector layer we will populate it with features. Each feature will contain one geometry (so, one road) and its attribute fields (i.e. ID, urban quarter, ect.). For this we have to create the fields and set them to the features. Finally, we're going to insert the features into the new layer.

Because we're going to use all the pieces of code created up to this point, I've grouped them into a package called `GeospatialScala`. You can download the directory and place it into the `src/main/scala` directory of your Scala project

(https://github.com/RoxanaTesileanu/multivariate_analyses/tree/master/Deep Learning/src/main/scala/com/mai). The first line of the .scala source files specifies the package name. You must change it to "package GeospatialScala" if you have kept the src/main/scala directory structure.

```
import ReadPointCoordFromFile._
import CreateGeoms._
org.gdal.ogr.ogr.RegisterAll()
val driver = org.gdal.ogr.ogr.GetDriverByName("ESRI Shapefile")
val ds = driver.CreateDataSource("rdgps")
val googleMapSR = new org.gdal.osr.SpatialReference()
googleMapSR.ImportFromEPSG(3857)
val lyr = ds.CreateLayer("RoadsAndGPSTracks", googleMapSR,
org.gdal.ogr.ogr.Constants.wkbMultiLineString)
val fd1 = new org.gdal.ogr.FieldDefn("Type", org.gdal.ogr.ogr.Constants.OFTString)
val fd2 = new org.gdal.ogr.FieldDefn("ID", org.gdal.ogr.ogr.Constants.OFTInteger)
val newLyr = ds.GetLayer(0)
newLyr.CreateField(fd1)
newLyr.CreateField(fd2)
val usedDfn = newLyr.GetLayerDefn()
val feat1 = new org.gdal.ogr.Feature(usedDfn)
feat1.SetGeometry(road1)
feat1.SetField("Type", "road")
feat1.SetField("ID", 1)
newLyr.CreateFeature(feat1)
val feat2 = new org.gdal.ogr.Feature(usedDfn)
feat2.SetGeometry(road2)
feat2.SetField("Type", "road")
feat2.SetField("ID", 2)
newLyr.CreateFeature(feat2)
val feat3 = new org.gdal.ogr.Feature(usedDfn)
feat3.SetGeometry(road3)
feat3.SetField("Type", "road")
feat3.SetField("ID", 3)
newLyr.CreateFeature(feat3)
val feat4 = new org.gdal.ogr.Feature(usedDfn)
feat4.SetGeometry(gpsTrack1)
feat4.SetField("Type", "gpstrack")
feat4.SetField("ID", 4)
newLyr.CreateFeature(feat4)
newLyr.GetFeatureCount
ds.FlushCache
```

Code snippet 5: Explanation In the above code snippet, we've created an empty georeferenced layer called `lyr`. Next, we've created the field templates for two fields (type and ID), and inserted them into the layer stored by the data source. Afterwards, we've created the four features of to be inserted in the layer, namely: `feat1`, `feat2`, `feat3`, `feat4` in which we've stored the line geometries previously created (`road1`, `road2`, `road3`, `road4`, `gpsTrack1`). For each feature we've set not only the geometry used but also the field attributes, and inserted them into the extracted layer. We've written the changes to disk with `FlushCache()` called on the data source. You can see now that a new subdirectory appeared in your Scala project's directory called `rdgps` which contains the shapefile (a `.dbf` file, a `.prj` file, a `.shp` file and a `.shx` file). For more details see [3].

2.3.3 Reprojecting OGR geometries and vector layers

Let's suppose we want to reproject our previously created layer from WGS84 / Pseudo-Mercator (for which all points were in WGS 84/ Pseudo-Mercator) to WGS84 unprojected.

2.4 Overlay and proximity analyses

Using some examples the overlay and proximity tools of GDAL will be presented. The worked examples will deal with the following problems:

- 1) from the `RoadsAndGPSTracks` layer -> distances between roads and GPS track(s). I will create some additional imaginary tracks.

- 2) create the `HabitatPatches` multipolygon layer, then lay it over the `RoadsAndGPSTracks` layer, see which habitat polygons are used (intersection) and how intensively in time and space (does a GPS track for an animal show "circling" within a certain habitat patch, and where does the animal spends most of its time (for nighttime and daytime?). Include data on time and date. Eventually make a new layer only for GPS tracks.

- 3) do an intersection of the EU and Romania borders and find the area of the polygons for RO of the salt-affected soils from the data set from the ESDAC-JRC (<https://esdac.jrc.ec.europa.eu/content/saline-and-sodic-soils-european-union>) -> intersections

2.5 Writing vector data

2.6 Reading raster data

2.7 Pixels resizing

2.8 Moving window analyses

2.9 Map algebra

3 Spatial data analysis

3.1 Introduction into spatial data analysis

In its essence geostatistics adds georeferenced spatial information to the vector of recorded variables of each individual observation, which represents a spatial location ². It further adds into models the space dependent random error. In this document I state that the classical multivariate techniques like multiple regression analysis or multidimensional scaling, can include this spatial perspective into modeling by treating the X and Y coordinates as "classical" variables like in variable-based multivariate statistics, or, by adding the X and Y differences between points to the Euclidean distance like in object-based multivariate techniques, thus reducing the space dependent random error to the "classical" random error of linear models. Another basic principle of geostatistics is the recognition of local variation as opposed to the global variation of the entire dataset. This principle can be incorporated into multivariate statistics by using combined approaches, i.e. approaches which deal with global variation as well as with local variation of a dataset. In this work I propose the MANOVA-KNN pipeline [17] for reducing error as a combined approach dealing with both types of variation.

Geostatistics is a departure from classical statistics just because of the sentence that "it takes spatial autocorrelation of observations into account when predicting values for new points". It is not more than that contributing to its differentiation from classical multivariate statistics. Maybe the most important fact is that geospatial analysis doesn't treat variables as we are used to in classical statistics but uses individual observations (i.e. individual points) and investigates the relationships between them from a spatial perspective [18], [19], [20], [21], [22]. It is adding space as a variable in the vector of recorded variables of each individual observation/point. The highlight of individual points is actually like in object-based classical multivariate statistics. It is treating space as an autocorrelated variable across a series of individual points, and letting all the other variables be "classical".

For me the most important moment in this introductory phase is to make you realize that until now in geostatistics we didn't talk about classical samples of

²Section 3.1 is taken from [16]

observations, where we concentrated on variables, but instead in geostatistics we concentrated mainly on pairs of observations (i.e. of points). We computed covariances for such pairs, not covariances between variables as in classical statistics. We didn't have weights for entire variables, we had weights for individual points caring those values of the variables studied. It is I believe very important the moment when you understand this. Spatial models treat individual points in ways similar to treating individual variables in classical statistics. But we must be aware these are individual points we are talking about, and we very much use distance measures for objects (like the Euclidean distance) like in the multivariate object-based classical statistics.

That being said I think I can use the same matrix calculations as in classical object-based multivariate analyses (where we use objects to predict values for variables), i.e. a n by n MATRIX OF DISSIMILARITIES BETWEEN OBJECTS by means of which we derive variables as LINEAR COMBINATIONS OF THE OBJECTS (Q-mode analyses) - see [23]. And of course classical variable-based multivariate analyses are equally possible - see [23] and [24]. The example from Quinn and Keough (2002) at the multiple regression chapter, where the study of Paruelo and Lauenroth is presented in which they've modeled the relative abundance of C3 plants against longitude and latitude is an implementation of this perspective [25]. If the spatial analysis includes a random error with spatial dependence, then why not include in the model the X and Y coordinates as two separate variables and make the random error spatial independent? Adding appropriate variables to a model is the approach used in multivariate statistics to reduce the unexplained variation [23] , [24].

Maybe spatial analysis is just classical multivariate analysis (variable- or object-based, or combined); the important thing is to include X and Y variables in the model and to check for eventual local variations within the dataset.

This doesn't mean I give up the "spatial perspective". I will still use the X-Y coordinate plane to inspect how the residuals from fitted linear models are located. Eventually, treat local variation by using machine-learning optimization procedures or by delineating more than one target population. And reevaluate the sampling design based on these preliminary conclusions.

I will also use the classification of Cressie (1993) [18] which delineates three types of geospatial analyses:

- on continuous surfaces (raster) (using variable-based multivariate techniques)
- on discrete spatial features based on multiple points (lines, polygons) (using variable- and object-based multivariate techniques)
- on discrete spatial features based on individual points (points) (using object-based multivariate techniques).

3.2 Using LAPACK for building multivariate statistical models

In order to start building the multivariate statistical models, a library is required which offers computing solutions for linear algebra problems like solving systems of simultaneous equations (for multiple linear regression models), and, QR or SVD decomposition of statistical matrices (for MANOVA). In this work I use LAPACK (Linear Algebra Package) and ATLAS (Automatically Tuned Linear Algebra Software). The later provides the Fortran interface to the portably efficient BLAS (Basic Linear Algebra Subprograms) implementation [26] on which LAPACK is based [27]. LAPACK and BLAS are provided by the University of Tennessee, Berkeley University of California, Denver University of Colorado and Nag Ltd. [28] [29]. They are both written in Fortran and are freely-available software packages (see <http://www.netlib.org/lapack/>).

On Ubuntu you can install ATLAS and LAPACK using the following commands [30] [31] [32]:

```
$ sudo apt-get install libatlas-base-dev libatlas-doc
$ sudo apt-get install liblapack-dev liblapack-doc-man liblapack-doc liblapack-pic liblapack3 liblapack-test
```

It might also be necessary to install the J LAPACK package and the Fortran-to-Java compiler in order to access the LAPACK routines. You can install it using [33] [34]:

```
$ sudo apt-get install libjlapack-java f2j
```

The J LAPACK package provides the LAPACK numerical subroutines translated from Fortran to Java. It uses a Fortran-to-Java translator, f2j [35].

In order to import LAPACK in Scala, you need to add the package containing it to the build.sbt file of the Scala project ("net.sourceforge.f2j" % "arpack_combined_all" % "0.1" [36]). Afterwards, you can import it in Scala REPL using the command:

```
import org.netlib.lapack
```

You can check the routines provided by LAPACK using the TAB-autocompletion on the lapack instance. Among these, the driver routines for solving systems of linear equations for a general matrix (xGESV) or for linear least squares problems (xGELS for QR factorization, or xGELSS for SVD factorization) [37]. Let's suppose for a moment there are no random errors around and your recorded variables perfectly catch the variation of the target population! You could for

example try out to solve the system of equations $A\mathbf{x} = \mathbf{b}$, where A is a coefficient matrix (each row of A represents an observation, and each cell within an observation represents the entry for a recorded explanatory variable), \mathbf{x} is a column vector of unknowns from the system of equations and \mathbf{b} is a column vector containing the entries of the response variable [38] [39] [40] [41] [42] [11]^{3 4}.

solve $A\mathbf{x} = \mathbf{b}$, where⁵

$$A = \begin{bmatrix} 3 & 1 & 0 \\ -1 & 2 & 2 \\ 5 & 0 & -1 \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} 6 \\ -7 \\ 10 \end{bmatrix}$$

Coding pitfall  When using dgesv make sure to use the transposed version of the intended Arrays! A row vector is actually interpreted as a column vector and vice-versa. A 2D Array like A is interpreted as A^T . The bellow code will clarify how this can be overcome.

Code snippet 6: using LAPACK dgesv to solve linear equation systems This code snippet implements the dgesv subroutine using A and \mathbf{b} from above:

```
import org.netlib.lapack
import org.netlib.util.intW._
val noArows = 3
val nrhs = 1
```

³Note that vectors are expressed with small bold characters and matrices with italic capitals

⁴The solution to the system of equations $A\mathbf{x} = \mathbf{b}$ is found using the inverse of A , i.e. A^{-1} , as following: $\mathbf{x} = A^{-1}\mathbf{b}$. This is why it is important to have an invertible matrix in order to be able to find a solution. The inverse of A is found using the row reduction method from the relationship: $AI = A^{-1}I$ or using the determinant of A also by row reduction or using more complex methods. Or, the solution to $A\mathbf{x} = \mathbf{b}$ can be found through the LU decomposition of A , but if A is not invertible, then the U matrix is a zero matrix and the solution to the system of equations cannot be found [43] [44]. A square matrix is invertible only if its determinant doesn't equal 0, so not all matrices are invertible (in this case they are called singular matrices). See [41] for more details. In statistics it is important to have invertible matrices, i.e. for which the determinant is not 0 [23].

⁵The solution for this system is the column vector $\mathbf{x} = (0.33, 4.99, -8.33)$. For our errorfree vision of the world you can use the equation $0.33a + 4.99b - 8.33c = \text{prediction}$ to perfectly predict new response values from new recorded variables. This is of course not statistically correct! I will introduce random errors in the next sections.

```

val matrixA : Array[Array[Double]] = Array(Array(3, 1, 0), Array(-1, 2, 2),
Array(5, 0, -1))
var aoffset = 0
val lda = 3
var ipiv : Array[Int] = Array.fill(noArows)(0)
var ipivoffset = 0
var b : Array[Double] = Array(6, -7, 10)
var boffset = 0
val ldb= 3
var info = new org.netlib.util.intW(0)
lapack.Dgesv.dgesv(noArows, nrhs, (matrixA.transpose).flatten, aoffset, lda, ipiv,
ipivoffset, b, boffset, ldb, info)
val result = (b, ipiv)
for (i<- result._1) println(i)
for (i<-result._2) println(i)

```

The scala source file for the above gdesv implementation is in Appendix E and on my GitHub profile under the multivariate_analyses repository (the DeepLearning directory).

Code snippet 6: explanation So, what exactly does the above piece of code? In order to implement the dgesv subroutine, I’ve first imported the needed packages, then prepared the 11 arguments needed by the dgesv function. The dgesv function solves the linear system $A\mathbf{x} = \mathbf{b}$ using the LU decomposition of A , i.e. using the relationship: $LU = A$, where L is a lower triangular matrix and U is an upper triangular matrix. Briefly, it takes the linear system $LU\mathbf{x} = \mathbf{b}$ and sets $U\mathbf{x} = \mathbf{y}$ and $L\mathbf{y} = \mathbf{b}$ [41] [38]. This two-steps procedure makes it easier to find the solution, as finding \mathbf{y} from $L\mathbf{y} = \mathbf{b}$ is easier because L is a lower triangular matrix, and then knowing \mathbf{y} makes it easier to solve for \mathbf{x} from $U\mathbf{x} = \mathbf{y}$ also because it is easier to solve equations with triangular matrices than with general square matrices [41] [38]. The LU decomposition (also called LU factorization) is a high-level algebraic implementation of the Gaussian elimination procedure, which improves the efficiency of matrix algorithms [38]. Given the fact that LU decomposition works using permutations (i.e. operations meant to implement the swapping of rows and columns of the A matrix within the Gaussian elimination), the decomposition of A can also be written as: $A = PLU$, where P is the permutation matrix [37] [38]. There are several ways to obtain a permutation matrix P : partial pivoting (used by dgesv [37]), complete pivoting, and Rook’s pivoting [38], where pivoting means swapping rows and columns in order to get 0s under and above a leading 1 (<https://people.richland.edu/james/lecture/m116/matrices/pivot.html>). Partial pivoting involves swapping of rows, complete pivoting involves swapping rows and columns [45] and Rook’s pivoting provides a more stable version of pivoting using a geometric analysis of Gaussian elimination [46].

Now, back to the implementation code of the dgesv function. As previously mentioned, this function takes 11 arguments:

- the number of rows of the coefficient matrix A ,
- the number of columns of the \mathbf{b} column vector (which is 1),⁶
- the flattened and transposed coefficient matrix A ,⁷
- the starting index of the flattened Array[Double] version of the A matrix,⁸
- the leading dimension of A which represents its number of rows,
- the Array[Int] which represents a dummy array which will be overwritten by dgesv and which will store the IPIV array (the array which stores the pivot indices that define the permutation matrix P),⁹
- the starting index of the IPIV array,
- the \mathbf{b} column vector with the constants of the linear system, i.e. the right hand side; during the dgesv routine, \mathbf{b} gets overwritten and will contain the solution of the system,
- the starting index of the \mathbf{b} column vector,
- the leading dimension of \mathbf{b} which is the no. of rows of the column vector,
- and the info argument which is an Int of type org.netlib.util.intW and which will be overwritten with the output information message of dgesv, and should be 0 on success [39].

Using the above code a linear system of type $A\mathbf{x} = \mathbf{b}$ was solved using LAPACK dgesv, and, at the same time I've made use of Scala types and also the output was represented as a Scala type. This is an important aspect in further development steps of the statistical algorithms, as it doesn't "overload" the family of types used.

⁶the \mathbf{b} is also called right hand side, shortly rhs.

⁷The dgesv function requires a flattened transposed coefficient matrix A because of the following reasons: (i) the dgesv requires 1D Arrays [35], so the flattened A matrix needs to be used, at the same time one must be aware that \mathbf{b} represents a column vector not a row vector, thus, (ii) a transposed Scala 2D Array is used (because of the fact that \mathbf{b} actually should represent a column vector but it is not, it is a row vector), in order to preserve the flow of the matrix operations.

⁸ In Scala the starting index of Arrays is 0 [42] [11]

⁹The IPIV(i) array shows which row was interchanged at the the i^{th} step row [39].

3.3 From datasets to probability models

Every statistical analysis is performed on a dataset. A dataset can be represented as a data matrix A with m rows (each observation being represented by a row) and n columns (each recorded variable being represented by a column). Thus the data matrix A contains m observations, where $m = 1, \dots, i$, and n variables, where $n = 1, \dots, j$. Each observation is represented by a row vector \mathbf{m}_i (forming a set M of observations) and each variable is represented by a column vector \mathbf{n}_j (forming a set N of variables) [47]. Each entry in A , a_{ij} , represents the recorded value for the i^{th} observation \mathbf{m}_i on the j^{th} recorded variable \mathbf{n}_j [23] [24].

If prior information helps in distinguishing between response and predictor variables¹⁰, then the dataset can be splitted in two or more subsets and supervised statistical analyses are performed based on this dataset partitioning. If no prior information is available, and no clear partitioning of the dataset is possible, then unsupervised statistical analyses are performed [47] [24]. Supervised learning is about predicting response variables from predictor variables [47], whereas unsupervised learning is descriptive and aims at finding patterns in the data [48].

Approaches of supervised learning to solve prediction tasks can be classified according to the type of variables used [47] and according to the dataset partitioning [24]. The individual entries a_{ij} of the data matrix A , contain the values of the recorded variables. These values may represent continuous, discrete or categorical variables [24]. If there is one response variable and more predictor variables, we could use multiple linear regression (if the response variable as well as the predictor variables are continuous) or multifactor ANOVA (if the response variable is continuous and the predictor variables are categorical). If we have more than one response variables, we have composites of response variables, and according to their type we could use multifactor MANOVA (if there are more continuous response variables and more categorical predictor variables) or canonical correlation (if we would like to investigate the correlation between one set of variables and another set of variables for the same objects) [23] [48]. In a general sense, supervised learning deals with either regression (continuous predictor variables) or classification (categorical predictor variables) to predict new outcomes [48].

Unsupervised learning primarily aims at dimensionality reduction, as in PCA which reduces the number of continuous variables to a smaller number of newly derived variables called "components" or at discovering clusters as in cluster analysis [24] [23] [48].

Furthermore, we can approach a dataset either by columns, choosing a variable-based multivariate technique (R-mode analyses), or by rows, choosing an object-

¹⁰In the statistical literature predictor variables are also called explanatory variables, independent variables, features, attributes or covariates. The response variables are also called dependent variables [47] [48].

based multivariate technique (Q-mode analyses) [23]. There is a broad range of multivariate techniques. In this book I aim at developing the basic understanding of the most commonly used ones.

Now, how are probabilities related to a dataset variable? What is the variation of a dataset? And what is the uncertainty of a statistical model?

Suppose one of your dataset variables is the tree diameter at breast height (DBH) measured in cm. You've measured 255 trees, you use the diameter classes from 10 to 19, from 20 to 29, from 30 to 39, from 40 to 49, from 50 to 59, from 60 to 69, from 70 to 79 and from 80 to 89. Based on these classes you count how many trees fall in each class, and put this information in the below table.

Table 1: Counts of trees falling in each diameter class

class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8
2	9	38	67	74	50	12	3

From table 1 we can compute the probability of each class by dividing each class count by the total count 255^{11 12}, obtaining the information from table 2 [49]:

Table 2: Probabilities of each diameter class

class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8
0.0078	0.0352	0.1490	0.2627	0.2901	0.1960	0.0470	0.0117

Because each probability is associated with a single diameter class, table 2 practically defines a function [49].¹³ Such a function is called a probability function [49] or a probability model [13]. Table 2 represents a discrete probability function because it has a finite domain (class 1 to class 8 represent a discrete variable). Probability functions with an infinite domain within some real number interval (values like 43.25 cm DBH are possible and represent a continuous variable) are called continuous probability distributions. Discrete probability models are best visualized using a histogram. The bars of the histogram have the same width and their heights are determined by the probabilities of the possible values (i.e. class 1 to class 8). See fig. 2.

¹¹Negative probabilities are not allowed, a probability can only take values from 0 to 1.

¹²The sum of these probabilities is 1.00. Due to rounding errors, this is not the case for table 2, but using Arrays to store variables and to do calculations on them using the [Double] Scala type, overcomes this problem [11].

¹³A function is a rule that binds exactly one element from its domain to exactly one element of its codomain [50].



Figure 2: An empirical discrete probability function of a sample.

For a discrete probability function, the area of each bar gives the probability of a particular DBH class, each bar representing a rectangle with a width = 1 and a height of $P(X=x)$, where X is the discrete random variable (i.e. the DBH class) and x is a possible value of X [49]. The area of a rectangle is its width multiplied by its height. So, if we want to find out the probability that a tree falls in a class between class 4 and class 7, $P(class4 \leq classoftree \leq class7)$, we have to find the area of the bars involved in this calculation and add them up. Because it involves class 4, class 5, class 6 and class 7, we get $P(class4) + P(class5) + P(class6) + P(class7)$, i.e. $1(0.2627) + 1(0.2901) + 1(0.1960) + 1(0.0470)$, which is 0.7958. The sum of the probabilities of all possible values of our discrete random variable is 1.00, thus the total area under the bars is 1.00. This is an easy way of integrating under a function. When we have a discrete probability distribution, we use the geometry rules of finding areas of bars. When we have a continuous probability distribution (like the Gaussian distribution, for example) and its shape doesn't allow us to use such simple methods, we integrate using integration algorithms which find out the area, thus the probability for an interval of values $[a,b]$: $P(a \leq X \leq b) = \int_a^b f(x)dx$ [49]. Because a continuous probability function includes all possible outcomes/values of a continuous random variable, the sum of these probabilities is 1. This is the reason why the area under a continuous probability function is 1. Such functions are called **probability density func-**

tions. Continuous probability functions can be probability density functions if the integration over their range of possible values is 1 [49]. An important distinction between discrete probability distributions and probability density functions (which are continuous), is that in a discrete distribution, the probability that a random variable X takes one of the possible values is given for every possible value of X. In a probability density function, the probability that X equals a specific value say c is $\int_c^c f(x)dx = 0$, the area under a point is 0, thus, **for a probability density function, only probabilities of intervals can be found** [49]. Another way of expressing probabilities is using the **cumulative distribution function**, which gives the probability that a random variable X is less than or equal to an arbitrary value t. If f is a probability density function of a random variable X with a range of [a,b], then the cumulative distribution function is defined as: $F(x) = P(X \leq t) = \int_a^t f(x)dx$.

Every probability distribution is characterized using its location, spread and shape [23] [20] [19]. Measures of location aim at finding the center of the distribution and the commonly used one are the mean, median and mode. Measures of spread aim at describing how much a variable varies over its range of values and include variance, standard deviation, interquartile range. Finally, measures of shape describe the shape of the distribution and include the coefficient of skewness (describing the symmetry of a distribution) and kurtosis (describing the peakedness of a distribution).

For recalling these measures their formulae are provided below. For a recorded random continuous variable X with values x from 1 to i within its range, the interval [a,b], and having a distribution of values f(x):

MEAN
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

MEDIAN ¹⁴

$$x_{median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ odd} \\ (x_{n/2} + x_{(n/2)+1})/2 & \text{if } n \text{ even} \end{cases}$$

MODE The mode is the most frequent value in a distribution. The class with the highest probability on the histogram contains the mode. If in a histogram the class interval is small then the mid-value of the most frequent class may be taken as the mode.

¹⁴The median is the midpoint of the observed values if they are arranged in increasing order. Half of the values are below the median and half of the values are above the median. Once the data are ranked the formula for median can be applied.

VARIANCE $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

STANDARD DEVIATION ¹⁵ $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

INTERQUARTILE RANGE ¹⁶ $IQR = Q_3 - Q_1$

SKEWNESS ¹⁷ $skewness = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^3$

COEFFICIENT OF SKEWNESS ¹⁸ $g_1 = \frac{skewness}{s^3}$

KURTOSIS $kurtosis = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4$

COEFFICIENT OF KURTOSIS ¹⁹ $g_2 = \frac{kurtosis}{s^4} - 3$

In practice, even if we sample from a population and we have a distribution of sample values for a variable X, we are not able to construct a probability density function for X for every analysis [49]. Instead, we use one of the several probability density functions (like the Normal distribution, Uniform distribution, Poisson distribution a.o.) which are well known and investigated in detail.

¹⁵The standard deviation is simply the square root of the variance

¹⁶The interquartile range is the difference between the upper (Q_3) and lower (Q_1) quartiles. If the median splits the data into halves, the quartiles split the data into quarters. If the data are arranged in increasing order, then a quarter of the data (25%) falls below the lower or first quartile, and a quarter of the data falls above the upper or third quarter. The IQR catches the 50% of the data which is between the upper and lower quartile [20].

¹⁷Skewness is also called the third moment about the mean. Variance is also called the second moment about the mean. And the fourth moment about the mean is the kurtosis.

¹⁸The coefficient of skewness of symmetric distributions is 0.

¹⁹The coefficient of kurtosis relates mainly to the Normal distribution, whose coefficient of kurtosis is 0. Distributions which are more peaked than the Normal have a coefficient of kurtosis greater than 0, while those who are flatter have a coefficient of kurtosis smaller than 0 [19].

3.4 Commonly used probability density functions

- Uniform distribution
- Normal distribution
- Multivariate Normal distribution
- Beta distribution
- Binomial distribution
- Poisson distribution

UNIFORM DISTRIBUTION

The Uniform distribution is the simplest probability distribution. It implies the same probability for the entire range, $[a,b]$, of a random continuous variable X and is used to express the fact that no value of X is believed to be more possible than the other values of X for the interval $[a,b]$ [51] [49] [13]:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases} \quad (1)$$

See fig. 3 for an example of a uniform probability density function.

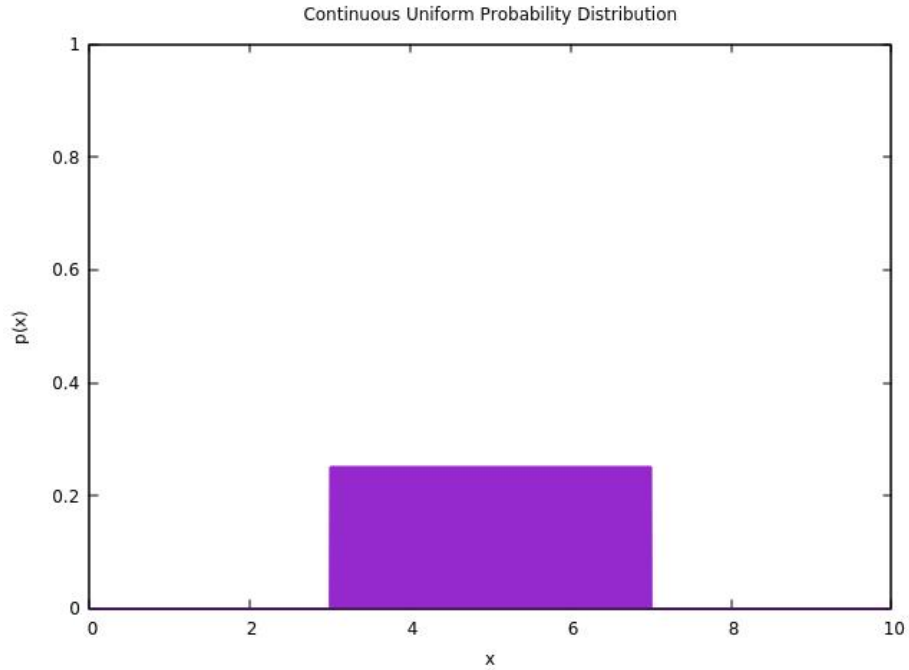


Figure 3: The continuous Uniform probability distribution for a continuous variable X taking values between 3 and 7 which all have the same probability of 0.25 (i.e. $1/(b-a) = 1/(7-3)$) of being the outcome.

We conventionally write the fact that X is uniformly distributed on the interval $[a, b]$ as $x \sim \text{Unif}(a, b)$ [51]. Like the sample distribution functions can be characterized by the mean, variance and shape, the probability density functions can also be characterized by these measures. The mean of a probability density function is called the expected value of $p(x)$ and is written as $E(x)$ and the variance of the probability density function is written as $V(x)$. If for discrete probability density functions the mean is calculated using the formula $E(x) = \sum_{i=1}^n x_i p_i$, for continuous probability density functions the mean is found by $E(x) = \int_a^b x p(x) dx = \mu$ [49]. With regard to the variance, for discrete probability functions $V(x)$ is calculated based on the expected value and the x_i values with their respective p_i as follows: $V(x) = \sum_{i=1}^n (x_i - E(x))^2 p_i$. For continuous probability density functions the variance $V(x)$ is calculated similarly but using integration again: $V(x) = \int_a^b (x_i - \mu)^2 p(x) dx = \sigma^2$. According to the above formulae for continuous probability functions, the $E(x)$ and $V(x)$ for the Uniform density are: $E(x) = \frac{b+a}{2} = \mu$ and $V(x) = \frac{(b-a)^2}{12} = \sigma^2$ ²⁰.

NORMAL DISTRIBUTION

The Normal distribution is one of the most important continuous probability density functions and it is used by many statistical techniques (for example linear regression or analysis of variance) involving random continuous variables which are supposed to be normally distributed implying that the deviations from the fitted model are also normally distributed [23] [49]. We conventionally write the fact that a random continuous variable X has a Normal density with mean $E(x) = \mu$ and variance $V(x) = \sigma^2$ as $x \sim N(\mu, \sigma^2)$ [49]. Its probability density function is given by:

$$p(x) = \left(1/\sigma\sqrt{2\pi}\right) e^{-(x-\mu)^2/2\sigma^2} \quad (2)$$

for $x \in (-\infty, +\infty)$, $\mu \in (-\infty, +\infty)$, and $\sigma^2 > 0$. From the previous subsection, it is known that integrating over the range of values of the continuous random variable X for a probability density function is 1. Thus, also for every Normal probability density $p(x)$ of a $x \sim N(\mu, \sigma^2)$ [49]²¹:

$$\begin{aligned} \int_{-\infty}^{+\infty} \left(1/\sigma\sqrt{2\pi}\right) e^{-(x-\mu)^2/2\sigma^2} dx &= 1 \\ E(x) &= \int_{-\infty}^{+\infty} x \left(1/\sigma\sqrt{2\pi}\right) e^{-(x-\mu)^2/2\sigma^2} dx = \mu \\ V(x) &= \int_{-\infty}^{+\infty} (x_i - \mu)^2 \left(1/\sigma\sqrt{2\pi}\right) e^{-(x-\mu)^2/2\sigma^2} dx = \sigma^2 \\ \sqrt{V(x)} &= \sigma^{22} \end{aligned}$$

²⁰See [49] for the integration steps needed to get to the $E(x)$ and $V(x)$ of the Uniform distribution.

²¹ μ and σ^2 are found using complex integration techniques beyond the scope of this text [49]

²² σ is called standard deviation of the mean.

Now, that the origins of μ and σ , the two parameters of every Normal distribution, are clear, we can see that there are infinitely many possible combinations of mean and variance, thus an infinite number of possible Normal distributions [23]. Each Normal probability distribution has an associated bell-shaped Normal curve. Every Normal curve is symmetric about a vertical line through the mean μ , and its inflexion points are found at X values $\mu - \sigma$ and $\mu + \sigma$. A Normal curve never touches the x -axis; it is asymptotic in both directions of the x -axis [49]. See fig. x (Normal curve for mean 50 and variance 100, standard dev=10). Because there is an infinite number of possible Normal distributions which might be used, calculating probability values from all these curves would imply too much computational effort each time an analysis is done, thus computer programs use the standard Normal distribution (z distribution) when performing calculations [49]. The z distribution is a Normal probability distribution with mean of zero and a variance of one [23]. Because the probability values of the z distribution can only be found iteratively through intensive computation, these values are taken from already existing tables [49].

z-Scores Theorem The x_i values of any Normal distribution are converted to z_i -scores, i.e. the x_i values of a z distribution, using the following relationship:

$$z_i = \frac{x_i - \mu}{\sigma}$$

The probability value to the left of any number x_i of any Normal curve corresponds to the probability value to the left of its associated z_i score of the standard Normal curve [49]. According to the above formula, the z -scores are actually standard deviation multiples (being obtained through division by σ). Thus, a z -score of 1.00 has a probability, i.e. area under the standard Normal curve to its left, of 0.8413 (read from z -table) and a z -score of -1.00 has a probability of 0.1587 (read from z -table) associated to it. We can thus say that between $\mu - \sigma$ and $\mu + \sigma$, i.e. between $-\sigma$ and $+\sigma$ of a standard Normal curve (because $\mu = 0$), we have $0.8413 - 0.1589 = 0.6826$ of area, i.e. probability [49]. Or, in other words, 68.26% of all possible X values of any Normal curve are between $\mu - \sigma$ and $\mu + \sigma$ [52]. Similarly, 95.44% of all possible X values are between two standard deviations below and above the mean, and 99.74% are between three standard deviations below and above the mean [52].

MULTIVARIATE NORMAL DENSITY

At the beginning of section 3.3 the multivariate data matrix A was introduced, with \mathbf{m}_i observations (forming a set M) and \mathbf{n}_j variables (forming a set N). The univariate Normal distribution can be extended by adding more dimensions, i.e. more variables, to an observation \mathbf{m}_i . If for a univariate Normal distribution, each observation \mathbf{x}_i of a variable X can be represented as a one-dimensional vector containing the value of one recorded variable, for bi- and multivariate Normal distributions the individual observations \mathbf{m}_i can be represented as two- or multidimensional vectors containing the values of two or more recorded variables (the set N). For a bivariate Normal distribution, with \mathbf{m}_i expressed as

$\mathbf{m}_i = (n_{i1}, n_{i2})$, the mean of the bivariate Normal distribution is expressed as a $n \times 1$ matrix $\boldsymbol{\mu} = [\mu_1, \mu_2]'$ ²³ and the variance $V(\mathbf{x})$ becomes a variance-covariance matrix $\boldsymbol{\Sigma}$ [51] [53] which will be implemented in Scala in section 3.5. For now, note that the elements of $\boldsymbol{\Sigma}$ are the covariances between the variables contained in the set N , and, that $\boldsymbol{\Sigma}$ represents the multivariate equivalent for $V(\mathbf{x})$ [53].

Conventionally, we write the fact that the observations of the set M are distributed according to a multivariate normal density as $\mathbf{m} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [51]. For a bivariate Normal distribution we can write:

$$\mathbf{m} = \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

BETA DISTRIBUTION

The Beta distribution is commonly used as a prior probability distribution for binary variables in Bayesian analyses (e.g. within the beta-binomial model [51], [13]) [23].

3.5 Computing statistical matrices for multivariate analyses

4 Variable-based multivariate analyses

4.1 Predicting with Multiple Linear Regression (MLR) analysis

4.2 Examining group differences with Multivariate Analysis of Variance (MANOVA)

4.3 Vizualizing MANOVA residuals: spatial residual map

MANOVA residuals are plotted on the X-Y plane of the spatial coordinates in a spatial residual map indicating their magnitude (i.e. dimensions). On this plot areas of high clumped residuals are delineated²⁴. Within each of these areas of local variation, the MANOVA residuals are investigated with the multiple regression technique. This will explain the magnitude of residuals (i.e. how large they are) using the X and Y spatial coordinates in which the data are given.

$$residual_{fromManova} = \beta_1 X + \beta_2 Y + \beta_0 + \varepsilon$$

From a simple slope analysis you can see how the residuals behave when X increases at a constant Y. The slope from this analysis gives the angle of the k-neighborhood for the adaptive KNN classifier for that area:

²³The $n \times 1$ matrix has the form of a column vector.

²⁴Section 4.3 is taken from [16]

$$\tan(\text{neighborhoodangle}) = \text{slope}$$

This angle gives the axis along which the neighborhood will be set.

4.4 Classifying with Discriminant Function Analysis (DFA)

5 Object-based multivariate analyses

5.1 Implementing the k-Nearest Neighbors classifier

The present subsection introduces the Scala functions used to run the kNN classification algorithm and then it delineates its role within a real-world classification pipeline for reducing error ²⁵.

In order to access the functions for the kNN classification, the scalaML package needs to be imported into the current REPL session. The directory with the Scala source files can be downloaded at https://github.com/RoxanaTesileanu/multivariate_analyses/tree/master/DeepLearning/src/main/scala/com/mai.

The kNN classifier reads CSV and text files for classification tasks, uses basic vector operations for handling variables and objects, runs the kNN algorithm on the data and computes the error rate of the kNN classifier.

Basic vector operations Because most statistical algorithms manipulate datasets which are actually collections of vectors, manipulating vectors is an essential task. The scalaML package offers this functionality, using the Scala main types while developing machine-learning algorithms. This makes the process of synchronizing types in further development stages easier. The vector operations are found in the BasicVectorOP.scala source file (see appendix x) and include: vector addition, vector subtraction, elementwise multiplication and dot product for any two vectors of type `Array[Double]`, and matrix multiplication for any two matrices of type `Array[Array[Double]]`, where each `Array[Double]` represents an observation (object), and all elements with the same index of inside arrays (i.e. all elements with the same index inside of all `Array[Double]` within the `Array[Array[Double]]`) represent a column (i.e. a variable) of the multivariate dataset [41], [42], [10], [54], [2].

INSERT HERE THE EXAMPLES FROM THE KNN DOCUMENTATION!!!

Reading files for classification tasks The ReadFile.scala source file (appendix x) of the scalaML package contains the function used to read CSV and text files for classification tasks. It returns a tuple with the information needed for the kNN algorithm (a data matrix of type `Vector[Array[Double]]`, the data labels in the form of a `Vector[Int]`, and the used classes in the form of a `Range[Int]`) [55], [42], [10], [2].

²⁵Section 5.1 is taken from [17]

INSERT HERE THE EXAMPLES FROM THE KNN DOCUMENTATION!!!

KNN classification function The kNN classification function (see appendix x) implements in Scala the pseudocode of the kNN algorithm [55], [42], [10], [2]. It starts by calculating the Euclidean distance between a new object (which is going to be classified by the algorithm) and every object of the dataset, it sorts the distances in increasing order, it then takes k objects with the lowest distances and finds the majority class among those k objects, returning it as the prediction for the class of the new object [55]. The variables can also be scaled from 0 to 1 before using the kNN classifier, if a normalization is required. For this purpose a normalizing function can be used (see appendix D) [23], [55], [42], [2].

INSERT EXAMPLES FROM THE KNN DOCUMENTATION!!!

Testing framework In order to test the accuracy of the kNN classifier, an error rate can be computed using the testing function provided in the TestFrameWorkClassif.scala source file (see appendix x) [55], [42], [2]. The error rate is given by the total number of errors (misclassified observations) divided by the total number of tested observations.

INSERT EXAMPLES FROM THE KNN DOCUMENTATION!!!

6 Combined approaches: the MANOVA-KNN pipeline

6.1 The learning pipeline for reducing error

Section 6.1. is taken from [16]. The machine-learning workflow for classification tasks for high-dimensional datasets (i.e. more than three variables recorded for each object) generally consists of the following steps: data collection, reading the data in Scala, or other programming language used for statistical computation, analyzing the data using multivariate analyses like (MANOVA, PCA, cluster analysis, etc.), applying and testing the classification algorithm (kNN, Bayes classifier, etc.), deciding on and argumenting the choice of the most appropriate model. Real-world workflows are iterative, requiring a dynamic configuration of the used algorithms and eventually the application of different algorithms to compare the results. Such workflows are called in machine-learning practice "learning pipelines" [12].

The present work proposes a combination of the MANOVA experimental design and analysis, with a further data classification step using the adaptive kNN classifier. The background idea is that, if two or more groups differ in their group centroids, they are, globally speaking, suitable for classification tasks. If, not, then we probably miss the variables which make a difference, or, we don't really have two different populations, and further classifications are possible but miss their point. In the first case, we further investigate (add new variables or remove current variables), and if the MANOVA test indicates group differences, we are able to find a first decision boundary using the linear combination with

the highest eigenvalue.

The decision boundary ideally separates the different groups. Though, variables overlap in the decision space, forming a "transition zone". This is where the residual variance comes from. KNN can then be further applied to investigate the situation occurring in the neighborhood of the objects located in the transition zone, and the best techniques to do this logically appear to be distance-based (i.e. based on dissimilarity measures between objects). Now, the question is how we delineate the transition zone? We can check the objects which produce the MANOVA residual (analysis of the residual) [23]. For high residuals we choose a higher k number, and for low residuals we choose a lower k number. This approach could represent a helpful investigation pipeline for spatial local variation and for diffusion processes in general, possibly requiring the use of alternative dissimilarity measures identifying the correct direction for neighbor selection, and certainly needs further investigation. Nevertheless, it promises a high practical relevance for geospatial and environmental data analysis.

6.2 Dealing with local variation: the adaptive KNN

From the spatial residual map I get the angle of the k-neighborhood for the adaptive KNN classifier. I still have to set the direction and the size of the neighborhood. For this I have to set boundaries along the X and Y axis. This means I have to delineate the zone of local variation, and set an extent of it. This information is also taken from the spatial residual map. Then from the left part of this spatial extent towards right I can choose the neighbors. For the size of the k-neighborhood I will use iterations (from [16]).

The idea of adapting the use of machine-learning algorithms to real-world data requirements is not new. Hastie and Tibshirani [56], describe the Discriminant Adaptive Nearest Neighbor (DANN) procedure as a locally adaptive form of nearest neighbors classification using a modified linear discriminant analysis (LDA) procedure to estimate an effective metric for computing neighborhoods [17]. The main idea of combining distance-based approaches with approaches based on the association between variables is driven by the fact that a multivariate dataset can be analysed in both ways [23]: calculating scores for the derived variables (components) for each object, and calculating dissimilarity measures (distances) for every possible pair of objects (from [17]).

7 Conclusions

Note

This document is "under construction". It contains older technical reports written by the author ([6], [17], [16]) and also new sections aiming at creating a

unified document which is easy to follow and util for a broad range of readers. Please download the current version from my GitHub profile under the multivariate_analyses project repository: https://github.com/RoxanaTesileanu/multivariate_analyses/blob/master/literature_analysis/geospatial_scala/geospatial_scala.pdf.

The statistical plots in this document were generated in Scala using the JavaPlot package developed by Panayotis Katsaloulis [57]. You can find the scala source files used for generating them under the link: https://github.com/RoxanaTesileanu/multivariate_analyses/tree/master/DeepLearning/src/main/scala/com/mai/scalaPlot.

The present document was edited using Latex [58] (<https://www.latex-project.org/>). The source .tex file of the present document is also available in the multivariate_analyses repository on my GitHub profile. Special thanks to Gustavo Mezzetti for the Latex halloweenmath package: <http://mirrors.concertpass.com/tex-archive/macros/latex/contrib/halloweenmath/halloweenmath-man.pdf>!

References

- [1] OSGeo, “GDAL Java API.” [Online]. Available: <http://gdal.org/java/>
- [2] EPFL, “Scala Standard Library 2.12.0,” 2017. [Online]. Available: <http://www.scala-lang.org/api/2.12.0/scala/index.html>
- [3] C. Garrard, *Geoprocessing with Python*. Shelter Island: Manning Publications Co., 2016.
- [4] S. Safavi, “Installing GDAL/OGR on Ubuntu,” 2015. [Online]. Available: <http://www.sarasafavi.com>
- [5] Canonical, “UbuntuGIS.” [Online]. Available: <https://wiki.ubuntu.com/UbuntuGIS>
- [6] R. Tesileanu, “Using Linux as a development platform for Scala projects,” 2017. [Online]. Available: https://www.researchgate.net/publication/319260791_Using_Linux_as_a_development_platform_for_Scala_projects
- [7] J. Suereth and M. Farwell, *SBT in action: the simple Scala Build Tool*. Shelter Island: Manning Publications Co., 2016.
- [8] R. U. Rehman and C. Paul, *The Linux Development Platform*. Pearson, 2003.
- [9] W. E. J. Shotts, *The Linux Command Line*. LinuxCommand.org, 2009.
- [10] J. Swartz, *Learning Scala*. Sebastopol: O’Reilly, 2015.

- [11] M. C. Lewis and L. L. Lacher, *Introduction to Programming and Problem Solving Using Scala*. CRC Press, 2017.
- [12] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark*. Sebastopol: O'Reilly, 2015.
- [13] A. Pfeffer, *Practical probabilistic programming*. Shelter Island: Manning Publications Co., 2016.
- [14] “SBT Documentation,” 2017. [Online]. Available: <http://www.scala-sbt.org/documentation.html>
- [15] Google, “Google Maps.” [Online]. Available: <https://maps.google.com/>
- [16] R. Tesileanu, “Geostatistics: classical multivariate statistics from a spatial perspective?” 2017. [Online]. Available: https://www.researchgate.net/publication/320798750_Geostatistics_classical_multivariate_statistics_from_a_spatial_perspective
- [17] R. Tesileanu, “Introduction to statistical computing in Scala: an implementation of the k-Nearest Neighbors classifier,” 2017, available at https://www.researchgate.net/publication/319505548_Introduction_to_Statistical_Computing_in_Scala-an_Implementation_of_the_K-Nearest_Neighbors_classifier.
- [18] N. A. C. Cressie, *Statistics for Spatial Data*. Wiley, 1993.
- [19] R. Webster and M. A. Oliver, *Geostatistics for Environmental Scientists*, second edition ed. Wiley, 2007.
- [20] E. H. Isaaks and R. M. Srivastava, *Applied Geostatistics*. New York: Oxford University Press, 1989.
- [21] T. Hengl, *A Practical Guide to Geostatistical Mapping*, 2009.
- [22] K. Johnston, Ver Hoef, Jay M., Krivoruchko, Konstantin, and Lucas, Neil, “Using ArcGis Geostatistical Analyst,” 2003.
- [23] G. Quinn and M. Keough, *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press, 2002.
- [24] J. D. Carroll and P. E. Green, *Mathematical Tools for Applied Multivariate Analysis*, revised ed. San Diego: Academic Press, 1997.
- [25] J. Paruelo and Laeunroth, W.K., “Relative abundance of plant functional types in grasslands and shrublands of North America,” *Ecological Applications*, vol. 6, pp. 1212–1224, 1996.
- [26] Sourceforge, “Automatically Tuned Linear Algebra Software ATLAS.” [Online]. Available: <http://math-atlas.sourceforge.net/>

- [27] Univ. of Tennessee, Berkeley Univ. of California, and Denver Univ. of Colorado, “LAPACK Working Note 81: Quick Installation Guide for LAPACK on Unix Systems,” 2007. [Online]. Available: <http://www.netlib.org/lapack/lawnspdf/lawn81.pdf>
- [28] Univ. of Tennessee and Berkeley Univ. of California and Denver Univ. of Colorado and Nag Ltd., “LAPACK,” 2017. [Online]. Available: <http://www.netlib.org/lapack/>
- [29] Univ. of Tennessee, Berkeley Univ. of California, Denver Univ. of Colorado, and Nag Ltd., “BLAS,” 2017. [Online]. Available: <http://www.netlib.org/blas/>
- [30] Canonical, “lapack package in Ubuntu.” [Online]. Available: <https://launchpad.net/ubuntu/+source/lapack>
- [31] Canonical, “atlas package in Ubuntu.” [Online]. Available: <https://launchpad.net/ubuntu/+source/atlas>
- [32] Canonical, “Ask Ubuntu.” [Online]. Available: <https://askubuntu.com>
- [33] Canonical, “liblapack-java package in Ubuntu.” [Online]. Available: <https://packages.ubuntu.com/trusty/liblapack-java>
- [34] Canonical, “f2j-java package in Ubuntu.” [Online]. Available: <https://packages.ubuntu.com/zesty/java/f2j>
- [35] D. M. Doolin, J. Dongarra, and K. Seymour, “JLAPACK - compiling LAPACK Fortran to Java,” *Scientific Programming*, vol. 7, pp. 111–138, 1999.
- [36] MvnRepository, “Maven Repository.” [Online]. Available: <https://mvnrepository.com/>
- [37] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, Du Croz, J., A. Greenbaum, Hammarling, S., McKenney, A., and Soresen, D., *LAPACK Users’ Guide*, 3rd ed. Philadelphia: SIAM, 1999.
- [38] G. H. Golub and Van Loan, C. F., *Matrix Computations*, 4th ed. The John Hopkins University Press, 2013.
- [39] Nag Ltd., “NAG Fortran Library Routine Document F07aaf (DGESV).”
- [40] K. Seymour, “Using the dgesv routine in Java - e-mail communication,” 2017.
- [41] P. Dawkins, *Paul’s notes on linear algebra*, ser. Math Tutorial Lamar University. Lamar University, 2005.
- [42] M. Odersky, L. Spoon, and B. Venners, *Programming in Scala*, 2nd ed. Walnut Creek: Artima, 2010.

- [43] Univ. of Tennessee, Berkeley Univ. of California, Denver Univ. of Colorado, and Nag Ltd., “Lapack java api,” 2017. [Online]. Available: <http://icl.cs.utk.edu/projectsfiles/f2j/javadoc/index.html>
- [44] Intel, “Intel Math Kernel Library LAPACK Examples.” [Online]. Available: https://software.intel.com/sites/products/documentation/doclib/mkl_sa/11/mkl_lapack_examples/index.htm#dgesv_ex.c.htm
- [45] Press, W.H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 1992, vol. 1.
- [46] Poole, G. and Neal, L., “The Rook’s pivoting strategy,” *Journal of Computational and Applied Mathematics*, vol. 123, pp. 353–369, 2000.
- [47] T. Hastie and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed. Springer, 2013.
- [48] K. P. Murphy, *Machine Learning A Probabilistic Perspective*. Cambridge: MIT Press, 2012.
- [49] M. Lial, R. Greenwell, and N. Ritchey, *Calculus with applications*, 10th ed. Pearson, 2012.
- [50] S. Takahashi and Trend-Pro Co. Ltd., *Manga Guide to Linear Algebra*. No Starch Press and Ohmsha Ltd., 2012.
- [51] S. Jackman, *Bayesian analyses for social sciences*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2009.
- [52] D. S. Moore, W. I. Notz, and M. A. Fligner, *The Basic Practice of Statistics*, 7th ed. MacMillan, 2015.
- [53] J. I. Marden, *Multivariate Statistics*. University of Illinois at Urbana-Champaign.
- [54] A. Trask, *Grokking deep learning*. Shelter Island: Manning Publications Co., 2017.
- [55] P. Harrington, *Machine learning in action*. Shelter Island: Manning Publications Co., 2012.
- [56] T. Hastie and R. Tibshirani, “Discriminant Adaptive Nearest Neighbor Classification and Regression,” *IEEE Trans. Pattern Anal. Match. Intell.*, vol. 18, no. 6, pp. 607–616, 1996.
- [57] P. Katsaloulis, “JavaPlot,” 2017. [Online]. Available: <http://javaplot.panayotis.com/>
- [58] C. Vellage, “LaTeX-Tutorial.com.” [Online]. Available: <https://www.latex-tutorial.com/>

Appendix A Point coordinates for code snippet 3

The following point coordinates were looked up on Google Maps (WGS84 / Pseudo-Mercator) and were used in section 2.3.2 to create the example vector layer. Please note they are for educational purposes.

@author: Roxana Tesileanu, INCDS BV, roxana.te@web.de

//points for Tampa Hill, BV, entered clockwise

45.631118, 25.583995
45.632863, 25.586184
45.633955, 25.586698
45.634345, 25.587706
45.636370, 25.589595
45.636730, 25.589423
45.639431, 25.593006
45.642450, 25.598735
45.642083, 25.603038
45.640676, 25.607061
45.639738, 25.607404
45.638925, 25.607131
45.636318, 25.604551
45.634559, 25.602963
45.633002, 25.602048
45.630514, 25.597860
45.630139, 25.596379
45.629801, 25.596057
45.629043, 25.596454
45.627693, 25.595639
45.627115, 25.594705
45.626380, 25.595060
45.625577, 25.594877
45.626220, 25.592055
45.626215, 25.589502
45.627325, 25.589545
45.628361, 25.589137
45.629576, 25.587421
45.631077, 25.583880

// road Valea Racadau, BV

45.612195, 25.588475
45.614445, 25.590527
45.620019, 25.594370

45.623190, 25.594629
45.625304, 25.595018
45.626392, 25.595363
45.626890, 25.595169
45.627222, 25.595623
45.628853, 25.597199
45.628929, 25.597285
45.629729, 25.599574
45.629442, 25.600114
45.630016, 25.601280
45.630590, 25.603785
45.631979, 25.606290
45.633806, 25.609572
45.637007, 25.611947
45.636780, 25.614906
45.636146, 25.618339

//road zona Carpatilor, BV

45.636539, 25.618857
45.633428, 25.620369
45.632975, 25.619160
45.631299, 25.620088
45.631888, 25.623327
45.630529, 25.624105
45.629789, 25.622442
45.627494, 25.622463
45.628023, 25.618209

//road zona spre Noua, BV

45.628627, 25.624364
45.624866, 25.629071
45.627011, 25.633951
45.623432, 25.635506
45.622359, 25.634275
45.619671, 25.631036
45.619626, 25.624752
45.620547, 25.622507

//points for Padurea Noua, BV, entered clockwise

45.612222, 25.589546
45.621849, 25.595210
45.626304, 25.597557
45.629890, 25.602105

45.632456, 25.609955
45.633898, 25.610253
45.635462, 25.612225
45.636364, 25.615346
45.635233, 25.616744
45.631935, 25.617357
45.629894, 25.621671
45.627750, 25.621804
45.628823, 25.617216
45.627631, 25.617099
45.626265, 25.618057
45.626477, 25.619801
45.625147, 25.621939
45.625546, 25.625259
45.623579, 25.626994
45.623868, 25.629202
45.622561, 25.632379
45.620556, 25.630997
45.620293, 25.626782
45.622734, 25.621884
45.621789, 25.618871
45.619452, 25.621083
45.618350, 25.625433
45.616194, 25.625733
45.613602, 25.628569
45.611492, 25.629284
45.602445, 25.618537
45.606375, 25.590345

//imaginary GPS-track

45.617797, 25.612945
45.620919, 25.618952
45.618454, 25.608328
45.622381, 25.619041
45.620180, 25.603591
45.617171, 25.585708
45.622910, 25.586664
45.633693, 25.596202

//points for Saua Gorita spre Postavaru

45.610059, 25.586784
45.621180, 25.560753
45.627951, 25.569719
45.627187, 25.577856


```
45.627582, 25.579325
45.630980, 25.583281
45.631979, 25.606290
45.629479, 25.587688
45.628346, 25.589346
45.626449, 25.589949
45.625579, 25.594658
45.615251, 25.590175
```

```
// points for the study area, entered clockwise
45.623683, 25.562561
45.643940, 25.599744
45.637329, 25.623779
45.622708, 25.636663
45.614645, 25.633913
45.599965, 25.615302
```

Appendix B Scala source file: readPointCoord.scala

@author: Roxana Tesileanu, INCDS BV, roxana.te@web.de

References:

M. Lewis - Introduction into Programming and Problem Solving in Scala

R. Tesileanu - Introduction into Statistical Computing in Scala - An Implementation of the k-Nearest Neighbors classifier

```
package com.mai.GeospatialScala
import scala.io._
object ReadPointCoordFromFile {

    val source = Source.fromFile("pointcoord.csv")
    val data = source.getLines.map(_.split(",")).toArray
    val dataHP1 = data.filter(_(2) == "1")
    dataHP1.length
    val dataRoad1 = data.filter(_(2) == "2")
    val dataRoad2 = data.filter(_(2) == "3")
    dataRoad1.length
    dataRoad2.length
    val dataRoad3 = data.filter(_(2) == "4")
    dataRoad3.length
    val dataHP2 = data.filter(_(2) == "5")
```

```

dataHP2.length
val dataGPSTrack = data.filter(_(2) == "6")
dataGPSTrack.length
val dataHP3 = data.filter(_(2) == "7")
dataHP3.length
val dataStArea = data.filter(_(2) == "8")
dataStArea.length
val pointsHP1 = dataHP1.map(i => (i(0).toDouble, i(1).toDouble))
pointsHP1.length
val pointsRoad1 = dataRoad1.map(i => (i(0).toDouble, i(1).toDouble))
val pointsRoad2 = dataRoad2.map(i => (i(0).toDouble, i(1).toDouble))
val pointsRoad3 = dataRoad3.map(i => (i(0).toDouble, i(1).toDouble))
val pointsHP2 = dataHP2.map(i => (i(0).toDouble, i(1).toDouble))
val pointsHP3 = dataHP3.map(i => (i(0).toDouble, i(1).toDouble))
val pointsGPSTrack = dataGPSTrack.map(i => (i(0).toDouble, i(1).toDouble))
val pointsStArea = dataStArea.map(i => (i(0).toDouble, i(1).toDouble))
}

```

Appendix C Scala source file: createGeoms.scala

@author: Roxana Tesileanu, INCDS BV, roxana.te@web.de

References: C. Garrard - Geoprocessing with Python

```

package com.mai.GeospatialScala
import ReadPointCoordFromFile._

```

```

object CreateGeoms {
val currentPosition = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPoint)
currentPosition.AddPoint(pointsRoad1(0)._1, pointsRoad1(0)._2)
currentPosition.GetX
currentPosition.GetY
val multiPointHP1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiPoint)
val geomsForMP = for (p<- pointsHP1) yield new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPoint)
val zippedGeomsPointsHP1 = geomsForMP.zip(pointsHP1)
for (z <- zippedGeomsPointsHP1) ( (z._1).AddPoint((z._2)._1, (z._2)._2))
for (z <- zippedGeomsPointsHP1) multiPointHP1.AddGeometry(z._1)
val road1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad1) road1.AddPoint(p._1, p._2)
road1.AddPoint(pointsRoad1(0)._1, pointsRoad1(0)._2)
val road2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad2) road2.AddPoint(p._1, p._2)
}

```

```

val road3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p <- pointsRoad3) road3.AddPoint(p._1, p._2)
val gpsTrack1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLineString)
for (p<- pointsGPSTrack) gpsTrack1.AddPoint(p._1, p._2)
val multiLineLines = Array(road1, road2, road3, gpsTrack1)
val multiLineEx = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiLineString)
for (l <- multiLineLines) multiLineEx.AddGeometry(l)
val habitatPatch1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing1 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p<- pointsHP1) habitatRing1.AddPoint(p._1, p._2)
habitatPatch1.AddGeometry(habitatRing1)
habitatPatch1.CloseRings()
val habitatPatch2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing2 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p<- pointsHP2) habitatRing2.AddPoint(p._1, p._2)
habitatPatch2.AddGeometry(habitatRing2)
habitatPatch2.CloseRings()
val habitatPatch3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val habitatRing3 = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p<- pointsHP3) habitatRing3.AddPoint(p._1, p._2)
habitatPatch3.AddGeometry(habitatRing3)
habitatPatch3.CloseRings()
val stArea = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbPolygon)
val ringStArea = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbLinearRing)
for (p <- pointsStArea) ringStArea.AddPoint(p._1, p._2)
stArea.AddGeometry(ringStArea)
stArea.CloseRings()
val multiPolygonEx = new org.gdal.ogr.Geometry(org.gdal.ogr.ogrConstants.wkbMultiPolygon)
val multiPolyPolys = Array(habitatPatch1, habitatPatch2, habitatPatch3, stArea)
for (poly <- multiPolyPolys) multiPolygonEx.AddGeometry(poly)
}

```

Appendix D Scala source file: createVLayer.scala

@author: Roxana Tesileanu, INCDS BV, roxana.te@web.de

References: C. Garrard - Geoprocessing with Python

```
package com.mai.GeospatialScala
```

```

import ReadPointCoordFromFile._
import CreateGeoms._

```

```

object CreateVLayer {

  org.gdal.ogr.ogr.RegisterAll()
  val driver = org.gdal.ogr.ogr.GetDriverByName("ESRI Shapefile")
  val ds = driver.CreateDataSource("rdgps")
  val googleMapSR = new org.gdal.osr.SpatialReference()
  googleMapSR.ImportFromEPSG(3857)
  val lyr = ds.CreateLayer("RoadsAndGPSTracks", googleMapSR, org.gdal.ogr.ogrConstants.wkbMultiLineStrin
  val fd1 = new org.gdal.ogr.FieldDefn("Type", org.gdal.ogr.ogrConstants.OFTString)
  val fd2 = new org.gdal.ogr.FieldDefn("ID", org.gdal.ogr.ogrConstants.OFTInteger)
  val newLyr = ds.GetLayer(0)
  newLyr.CreateField(fd1)
  newLyr.CreateField(fd2)
  val usedDfn = newLyr.GetLayerDefn()
  val feat1 = new org.gdal.ogr.Feature(usedDfn)
  feat1.SetGeometry(road1)
  feat1.SetField("Type", "road")
  feat1.SetField("ID", 1)
  newLyr.CreateFeature(feat1)
  val feat2 = new org.gdal.ogr.Feature(usedDfn)
  feat2.SetGeometry(road2)
  feat2.SetField("Type", "road") feat2.SetField("ID", 2)
  newLyr.CreateFeature(feat2)
  val feat3 = new org.gdal.ogr.Feature(usedDfn)
  feat3.SetGeometry(road3)
  feat3.SetField("Type", "road")
  feat3.SetField("ID", 3)
  newLyr.CreateFeature(feat3)
  val feat4 = new org.gdal.ogr.Feature(usedDfn)
  feat4.SetGeometry(gpsTrack1)
  feat4.SetField("Type", "gpstrack")
  feat4.SetField("ID", 4)
  newLyr.CreateFeature(feat4)
  newLyr.GetFeatureCount
  ds.FlushCache

}

```

Appendix E Scala source file: usingLapack2.scala

```

/*
This source file represents an implementation example of the LAPACK subrou-

```

tine dgesv in Scala.

@author: Roxana Tesileanu, INCDS BV, roxana.te@web.de

References:

K. Seymour - LAPACK F2J Javadoc

M. Odersky - Programming in Scala

M. Lewis - Introduction into Programming and Problem Solvin in Scala

Nag Ltd. - NAG Fortran Library Document F07AAF (DGESV)

*/

```
import org.netlib.lapack
import org.netlib.util.intW._

    object UsingLapack2{

val noArows = 3
val nrhs = 1
val matrixA : Array[Array[Double]] = Array(Array(3, 1, 0), Array(-1, 2, 2),
Array(5, 0, -1))
var aoffset = 0
val lda = 3
var ipiv : Array[Int] = Array.fill(noArows)(0)
var ipivoffset = 0
var b : Array[Double] = Array(6, -7, 10)
var boffset = 0
val ldb= 3
var info = new org.netlib.util.intW(0)
lapack.Dgesv.dgesv(noArows, nrhs, (matrixA.transpose).flatten, aoffset, lda, ipiv,
ipivoffset, b, boffset, ldb, info)
val result = (b, ipiv)
for (i<- result._1) println (i)
for (i<- result._2) println (i)
}
```