Teaching spatio-temporal analysis and efficient data processing in open source environment

by

Giuseppe Amatulli[1], Stefano Casalegno[2], Remi D’Annunzio[3], Reija Haapanen[4], Pieter Kempeneers[5], Erik Lindquist[3], Anssi Pekkarinen[3], Adam M. Wilson[1] and Raul Zurita-Milla[6]

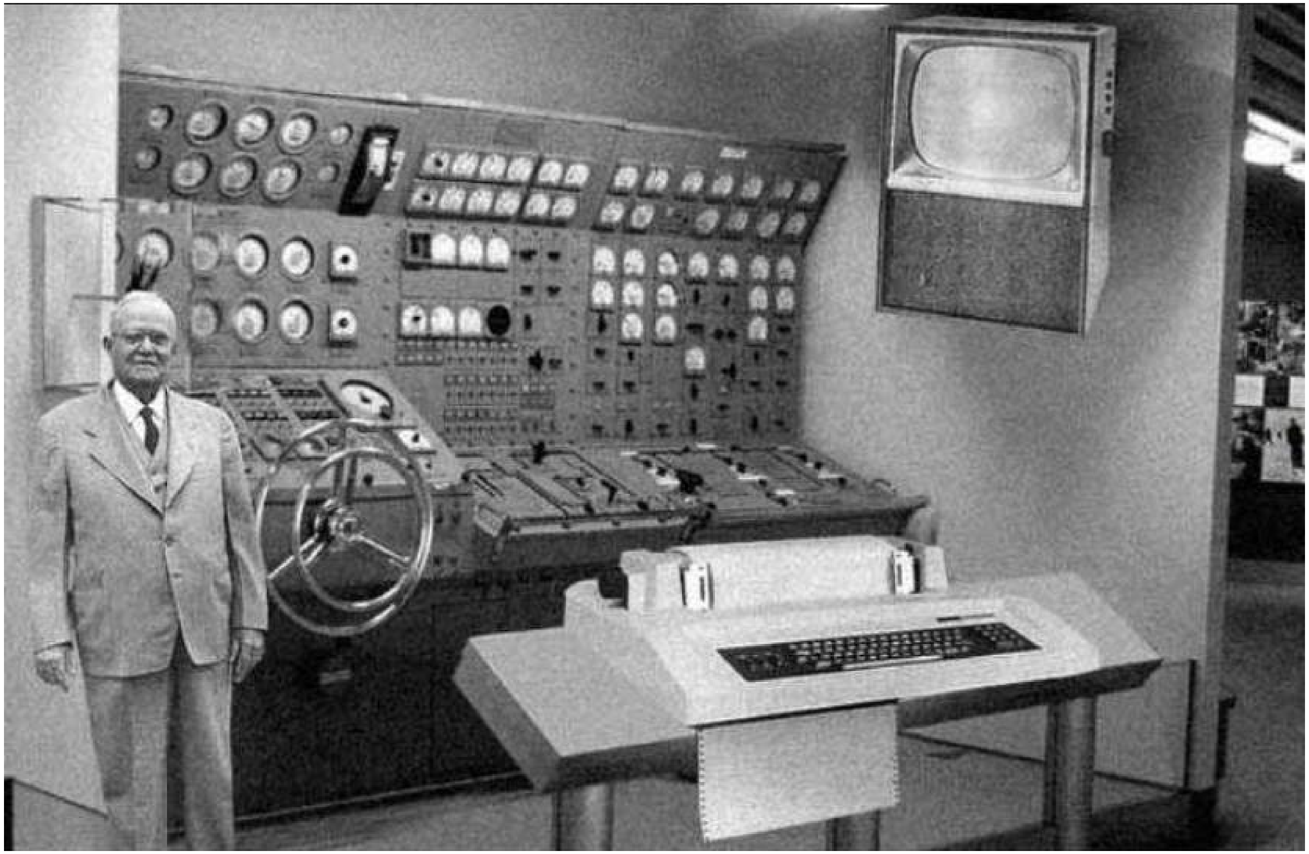
1 Department of Ecology and Evolutionary Biology, Yale University
2 University of Exeter, Environment and Sustainability Institute – spatial-ecology.net
3 FAO Forestry Department, Rome, Italy
4 Haapanen Forest Consulting, Vanhakylä, Finland
5 Flemish Institute for Technological Research (VITO), Mol, Belgium
6 University of Twente (NL), Faculty of Geo-Information Science and Earth Observation (ITC)

Corresponding author: giuseppe.amatulli@gmail.com

...

Introduction

- ✔Geo-data are getting larger and complex
- ✔To use these data, new capabilities in processing and new skills in analysis are needed



✔Many of the potential users have limited access to ICT equipment

✔Free and open-source software under the Linux Operating System (OS) can address this gap, as they can be used in desktop PCs, laptops and low-cost hardware such as the RaspberryPi

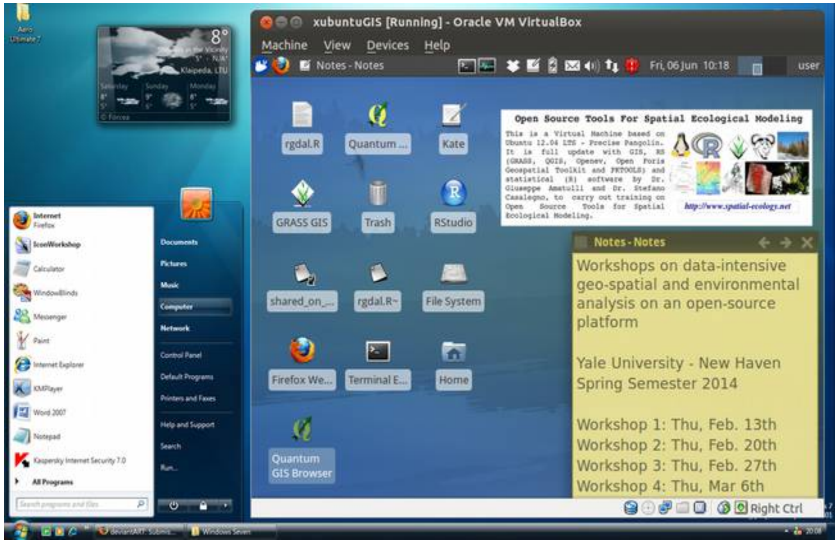
Introduction

- ✔We describe the effectiveness of teaching several open source programming languages to analyze spatio-temporal data

- ✔Experiences cover teaching in academic institutions, research centres, and national or international organizations in developed and developing countries.

Introduction

Installed Linux OS, Virtual Machine (VM) and/or LiveUSB are used



Open Source Tools For Spatial Ecological Modeling

This is a Virtual Machine based on Ubuntu 12.04 LTS - Precise Pangolin. It is full update with GIS, RS (GRASS, QGIS, Openev, Open Foris Geospatial ToolKit and PKTOOLS) and statistical (R) software by Dr. Giuseppe Amatulli and Dr. Stefano Casalegno, to carry out training on Open Source Tools for Spatial Ecological Modeling.

http://www.spatial-ecology.net

Notes - Notes

Workshops on data-intensive geo-spatial and environmental analysis on an open-source platform

Yale University - New Haven
Spring Semester 2014

Workshop 1: Thu, Feb. 13th
Workshop 2: Thu, Feb. 20th
Workshop 3: Thu, Feb. 27th
Workshop 4: Thu, Mar 6th

Learning programming skills to develop geo-spatial applications

- ✔Designing and implementing complex geo-spatial workflows require programming skills
- ✔Graduate students in env. sciences are rarely trained in these
- ✔In addition to actual problem solving, programming skills would:
  - ✔Enable an interactive approach to learning: algorithms and new ideas can be immediately explored and tested
  - ✔Train the student in scientific thinking

Learning programming skills to develop geo-spatial applications

- ✔ Combining expert knowledge on environmental sciences with versed programming is difficult

- ✔ One must be able to:
  - ✔ Conceptualise the problem
  - ✔ Master abstract programming concepts and techniques to actually write the code
  - ✔ Assess the results and
  - ✔ Assess the general data trend based on expertise on the actual subjects

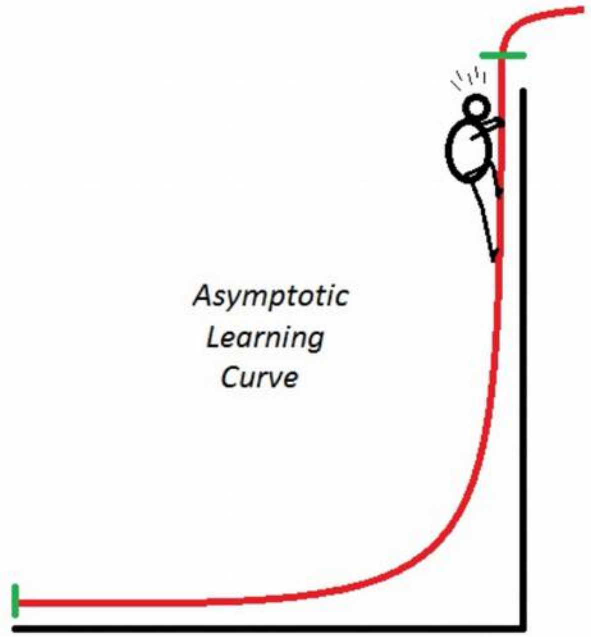
Learning programming skills to develop geo-spatial applications

✔Many gradute students may have the impression that programming is difficult to learn

✔This may unconsciously prevent students from getting started

✔Besides the psychological part, users may encounter problems in learning an open-source language, including:

1) Shifting to a new OS (e.g. Linux),
2) Learning new patterns and commands,
3) Installing necessary software,
4) Software updating



Developing open source geo-spatial programming course

- ✔We began organizing intensive training courses for graduate and postgraduate researchers and technicians

- ✔Courses target those who already possess basic knowledge in geospatial sciences

  → Focus is on applied processing with no introductory GIS/RS lessons

Developing open source geo-spatial programming course

- ✔Primary tool in these programming courses is a standardised Linux OS environment with:
  - ✔Pre-installed software packages,
  - ✔Readily available tutorials and
  - ✔Exercises

- ✔Documentation and exercises are freely accessible (through wikis www.spatial-ecology.net and www.openforis.org/wiki).

OPENFORIS

SPATIAL ECOLOGY

Developing open source geo-spatial programming course

- ✔Data analysis integrates multiple programming languages such as AWK, BASH, Python, R, GRASS to build workflows

- ✔Courses focus on automated processing and teach basic programming concepts for using command-line utilities to process large data sets.

- ✔With simple scripts, we show how to:
  - ✔Automate necessary tasks
  - ✔Adapt existing programs for specific problems

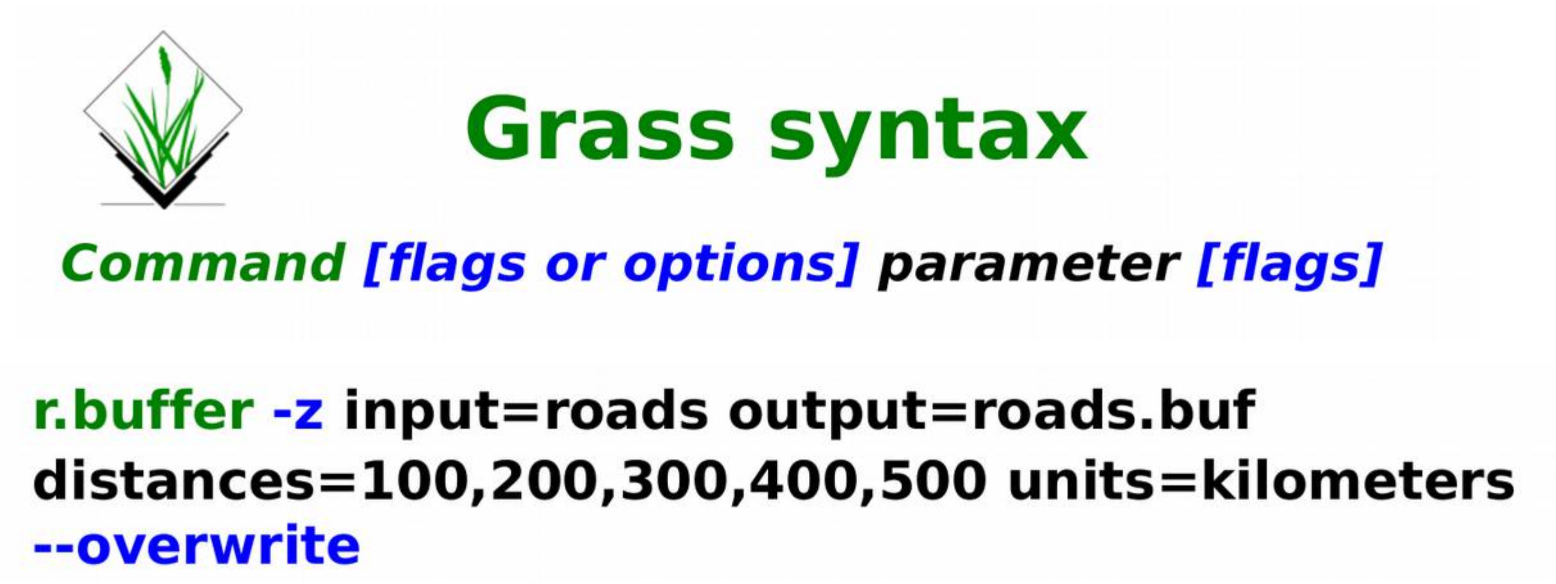
    ...to gain higher efficiency and improved performance

- ✔Scientific aspects are integrated as soon as the participants start to use the tools to solve specific analysis tasks

Developing open source geo-spatial programming course

Each language/tool is explained using a common structure:

1) Syntax, including specification of various flags and options,



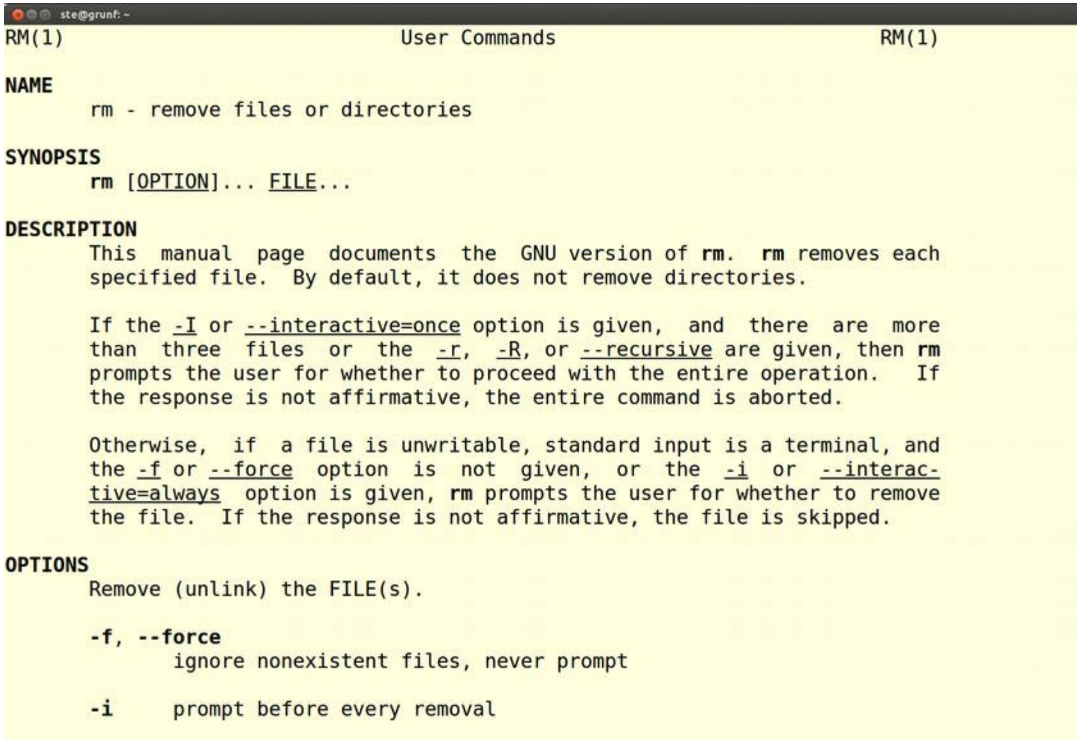
Grass syntax

Command [flags or options] parameter [flags]

```
r.buffer -z input=roads output=roads.buf
distances=100,200,300,400,500 units=kilometers
--overwrite
```

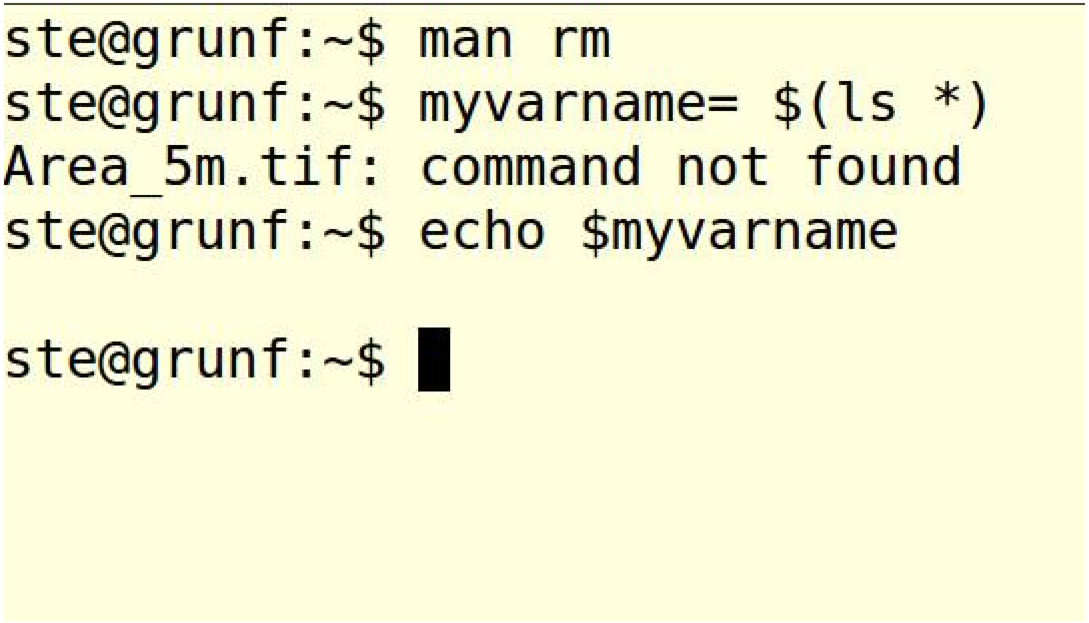
2) Accessing available documentation

!!!! Read the FANTASTIC manual



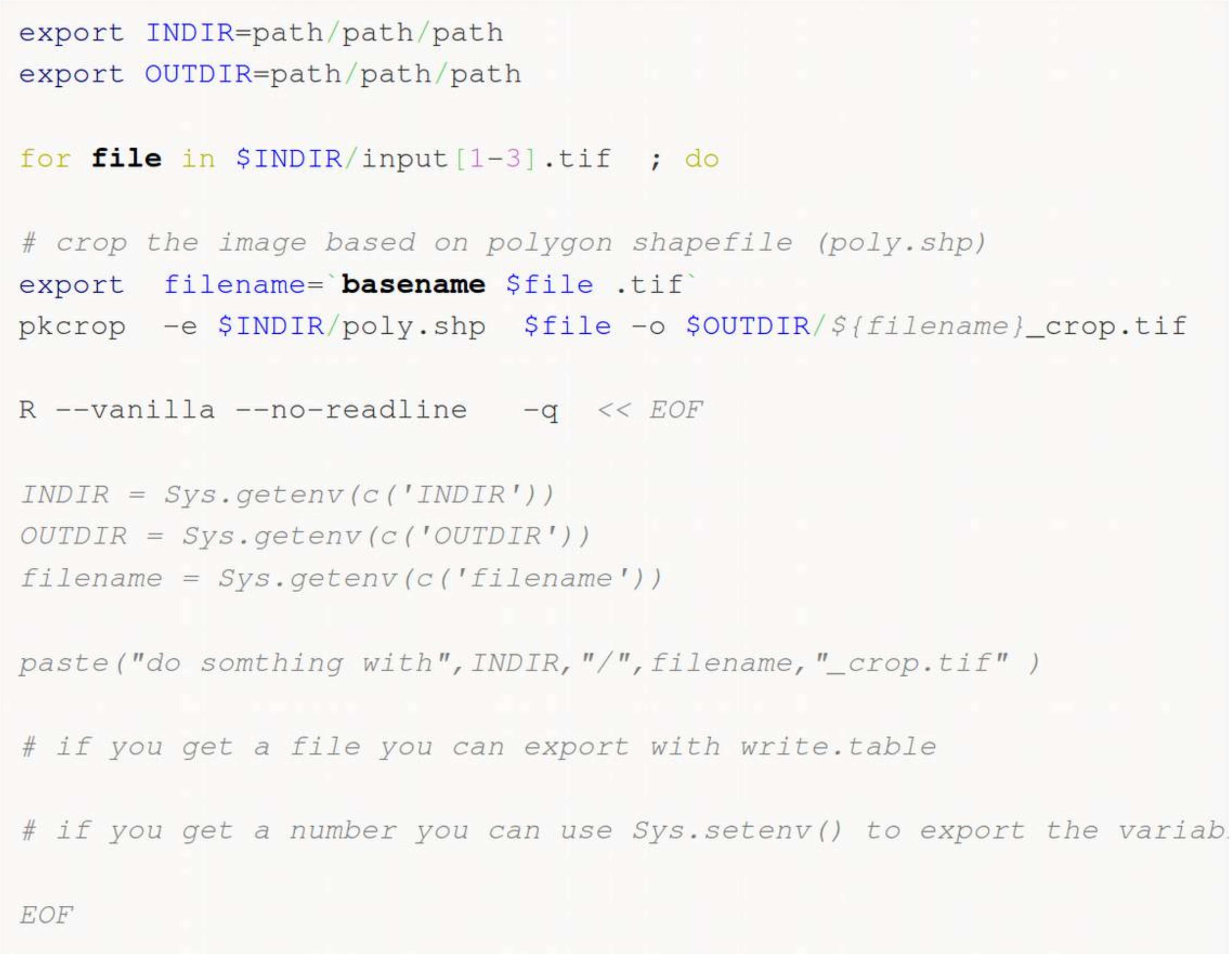
```
RM(1)                          User Commands                          RM(1)

NAME
       rm - remove files or directories

SYNOPSIS
       rm [OPTION]... FILE...

DESCRIPTION
       This  manual  page  documents  the  GNU version of rm.  rm removes each
       specified file.  By default, it does not remove directories.

       If the -I or --interactive=once option is given,  and  there  are  more
       than  three  files  or  the  -r,  -R, or --recursive are given, then rm
       prompts the user for whether to proceed with the entire operation.   If
       the response is not affirmative, the entire command is aborted.

       Otherwise,  if  a file is unwritable, standard input is a terminal, and
       the -f or --force  option  is  not  given,  or  the  -i  or  --interac-
       tive=always  option is given, rm prompts the user for whether to remove
       the file.  If the response is not affirmative, the file is skipped.

OPTIONS
       Remove (unlink) the FILE(s).

       -f, --force
              ignore nonexistent files, never prompt

       -i     prompt before every removal
```

3) Identifying and explaining the typical problems in data processing and the procedures to solve them (debugging)



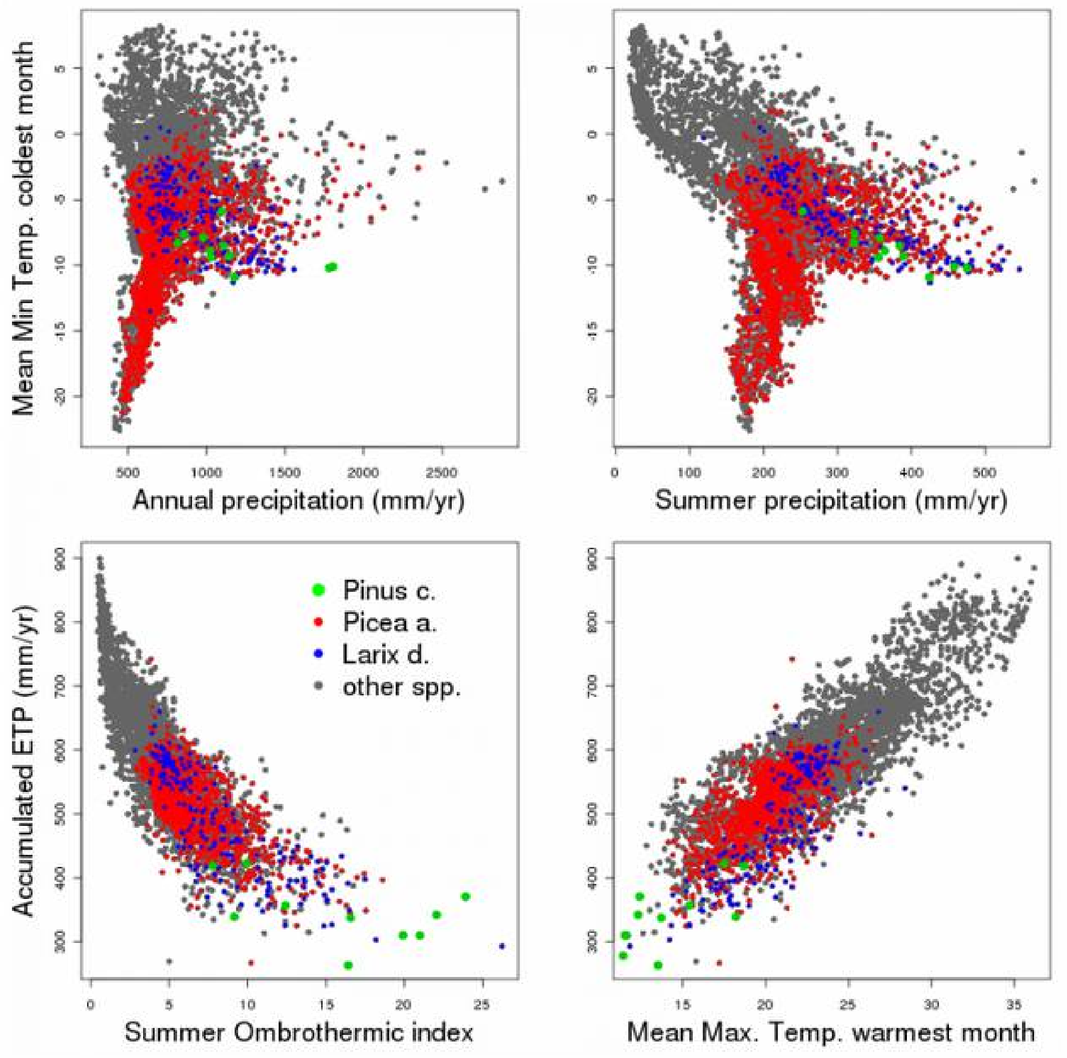
```
ste@grunf:~$ man rm
ste@grunf:~$ myvarname= $(ls *)
Area_5m.tif: command not found
ste@grunf:~$ echo $myvarname

ste@grunf:~$ █
```

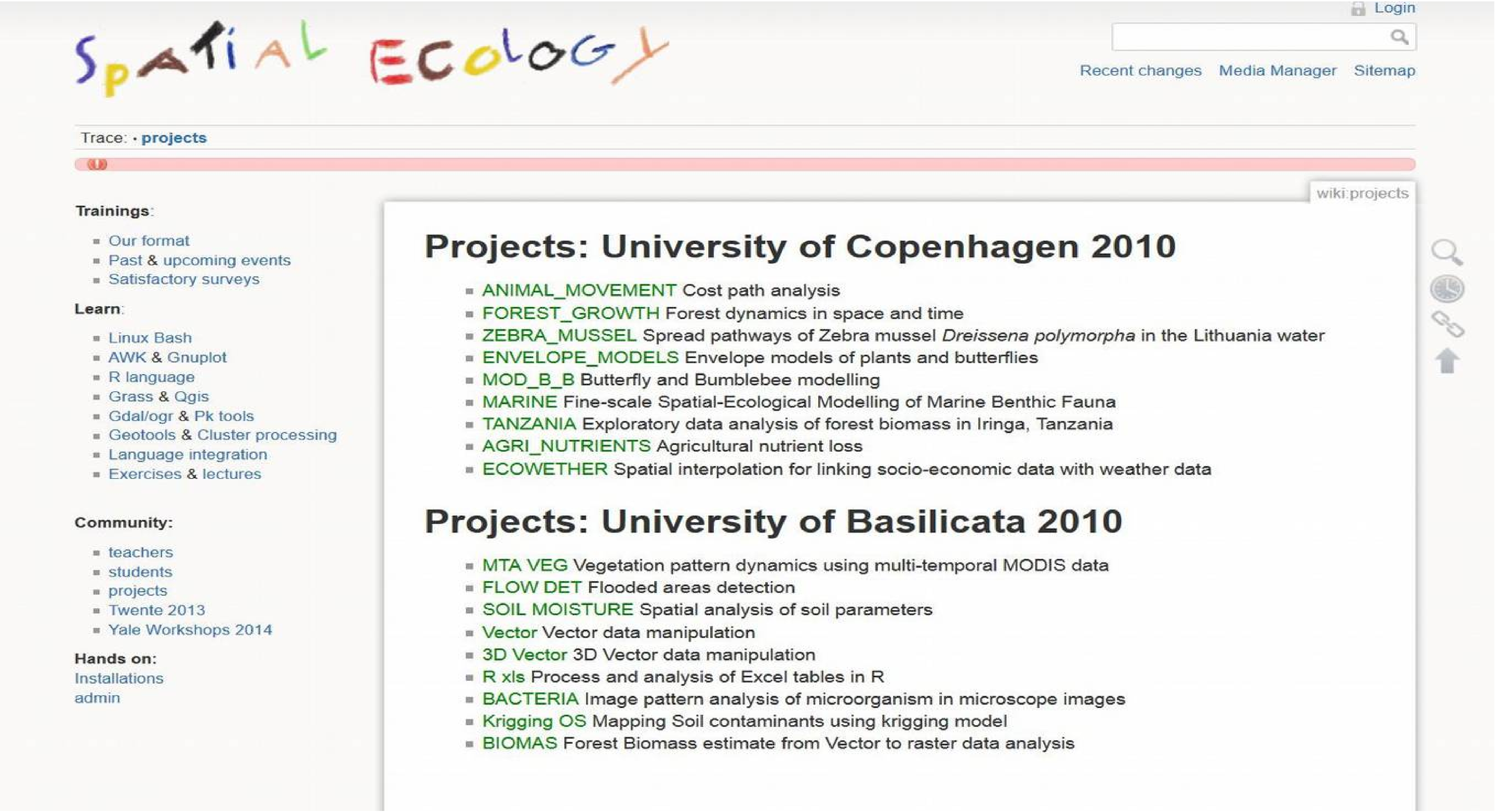
4) Structuring a script to connect various tools and/or languages



```
export INDIR=path/path/path
export OUTDIR=path/path/path

for file in $INDIR/input[1-3].tif  ; do

# crop the image based on polygon shapefile (poly.shp)
export  filename=`basename $file .tif`
pkcrop  -e $INDIR/poly.shp  $file -o $OUTDIR/${filename}_crop.tif

R --vanilla --no-readline   -q  << EOF

INDIR = Sys.getenv(c('INDIR'))
OUTDIR = Sys.getenv(c('OUTDIR'))
filename = Sys.getenv(c('filename'))

paste("do somthing with",INDIR,"/",filename,"_crop.tif" )

# if you get a file you can export with write.table

# if you get a number you can use Sys.setenv() to export the variab

EOF
```

5) Working with the output and
6) Evaluating the analysis results





Trainings:

- Our format
- Past & upcoming events
- Satisfactory surveys

Learn:

- Linux Bash
- AWK & Gnuplot
- R language
- Grass & Qgis
- Gdal/ogr & Pk tools
- Geotools & Cluster processing
- Language integration
- Exercises & lectures

Community:

- teachers
- students
- projects
- Twente 2013
- Yale Workshops 2014

Hands on:

Installations
admin

Projects: University of Copenhagen 2010

- ANIMAL_MOVEMENT Cost path analysis
- FOREST_GROWTH Forest dynamics in space and time
- ZEBRA_MUSSEL Spread pathways of Zebra mussel *Dreissena polymorpha* in the Lithuania water
- ENVELOPE_MODELS Envelope models of plants and butterflies
- MOD_B_B Butterfly and Bumblebee modelling
- MARINE Fine-scale Spatial-Ecological Modelling of Marine Benthic Fauna
- TANZANIA Exploratory data analysis of forest biomass in Iringa, Tanzania
- AGRI_NUTRIENTS Agricultural nutrient loss
- ECOWETHER Spatial interpolation for linking socio-economic data with weather data

Projects: University of Basilicata 2010

- MTA VEG Vegetation pattern dynamics using multi-temporal MODIS data
- FLOW DET Flooded areas detection
- SOIL MOISTURE Spatial analysis of soil parameters
- Vector Vector data manipulation
- 3D Vector 3D Vector data manipulation
- R xls Process and analysis of Excel tables in R
- BACTERIA Image pattern analysis of microorganism in microscope images
- Krigging OS Mapping Soil contaminants using krigging model
- BIOMAS Forest Biomass estimate from Vector to raster data analysis

Developing open source geo-spatial programming course

- ✔ Providing a “ready-to-use” Linux OS including all necessary software and course materials ensures:
  1) minimum software installation effort
  2) standardized working environment
  3) easy portability to other computers, and
  4) gentle transition from the host OS to guest OS

- ✔ Participants are able to create their own workspace for data and scripts within the Linux OS VM/LiveUSB

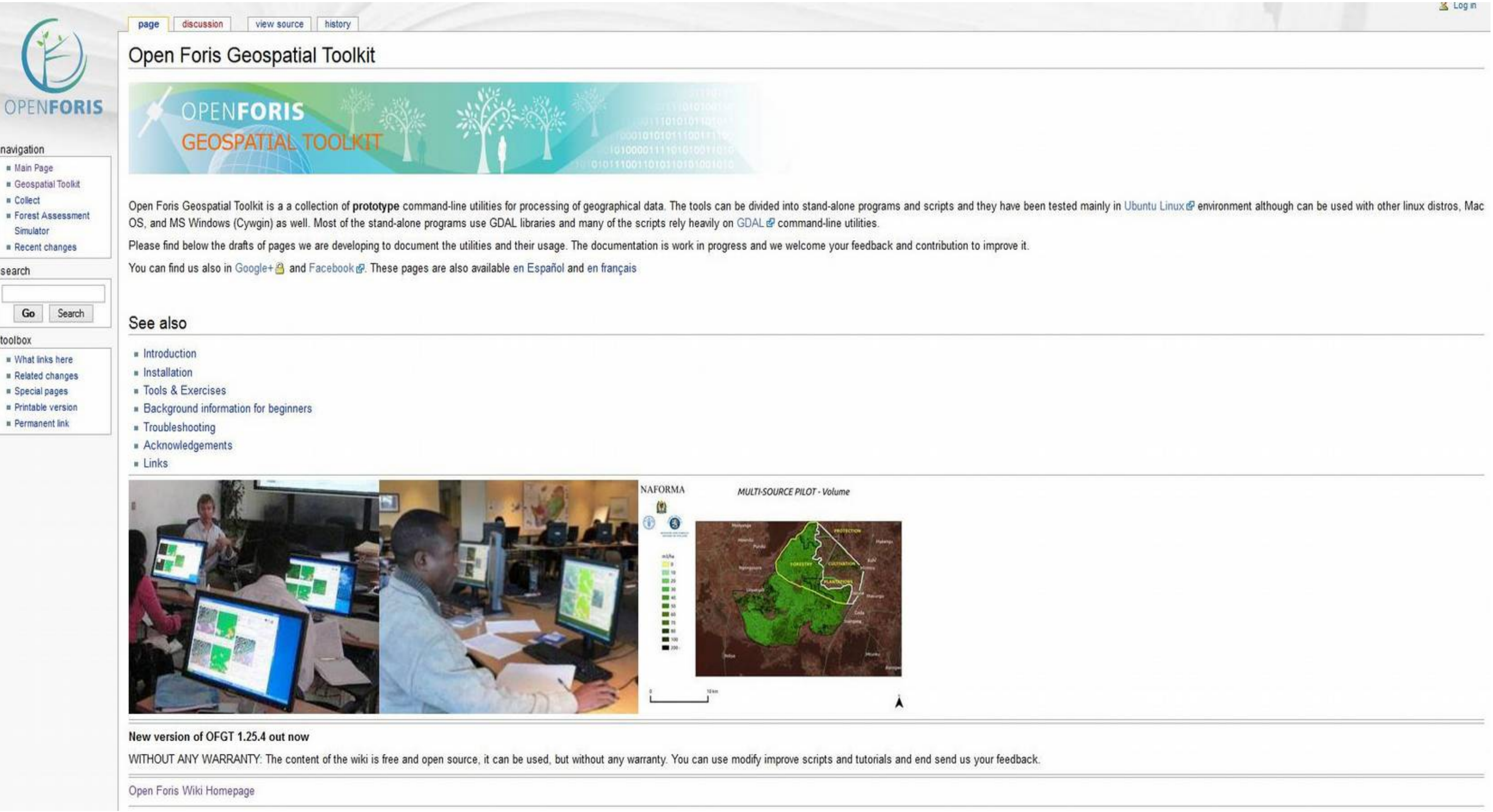
- ✔ A common cloned Linux OS facilitates group problem-solving, debugging, and testing commands during the training

Developing open source geo-spatial programming course

- ✔Selection of the teaching environment depends mainly on the available hardware resources:
  - ✔ If PCs with more than 4GB of memory are available, VM solutions can be efficiently applied
  - ✔ With older hardware and less memory, a Live system, in which the entire OS and required software is contained in USB or CD is used

Software tools and support material

- In addition to commonly used geo-spatial software such as GRASS and QGIS, we also incorporate several command-line utilities:
- Open Foris Geospatial Toolkit (OFGT)
  - Part of the FAO’s Open Foris initiative aimed at helping developing countries assess and monitor their forest resources
  - Includes image processing tools: radiometric normalization, image algebra, image segmentation, classification
  - As well as tools for change detection, extraction of image statistics, image filtering, feature extraction and gap-filling
  - Tools are divided into stand-alone C-programs and scripts (AWK, BASH, Python, perl, R)
  - Many of the stand-alone programs and scripts use GDAL/OGR libraries and command line utilities



navigation

- Main Page
- Geospatial Toolkit
- Collect
- Forest Assessment Simulator
- Recent changes

search

Go Search

toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

Open Foris Geospatial Toolkit

Open Foris Geospatial Toolkit is a a collection of **prototype** command-line utilities for processing of geographical data. The tools can be divided into stand-alone programs and scripts and they have been tested mainly in Ubuntu Linux environment although can be used with other linux distros, Mac OS, and MS Windows (Cywgin) as well. Most of the stand-alone programs use GDAL libraries and many of the scripts rely heavily on GDAL command-line utilities.

Please find below the drafts of pages we are developing to document the utilities and their usage. The documentation is work in progress and we welcome your feedback and contribution to improve it.

You can find us also in Google+ and Facebook. These pages are also available en Español and en français

See also

- Introduction
- Installation
- Tools & Exercises
- Background information for beginners
- Troubleshooting
- Acknowledgements
- Links

New version of OFGT 1.25.4 out now

WITHOUT ANY WARRANTY: The content of the wiki is free and open source, it can be used, but without any warranty. You can use modify improve scripts and tutorials and end send us your feedback.

Open Foris Wiki Homepage

Software tools and support material

- ✔pktools:
  - ✔Open source software toolbox (GPL v3)
  - ✔Implemented in similarly to the GDAL/OGR utilities
  - ✔pktools is more oriented towards remote sensing and image processing.
  - ✔Includes e.g. image filtering and composing multiple images according to different rule sets
  - ✔Also machine learning algorithms for image classification or data regression, including feature selection algorithms, artificial neural networks and support vector machines

- ✔After pre-processing of data with geo-spatial software, R and various BUGS languages are used to fit statistical models and produce summary output

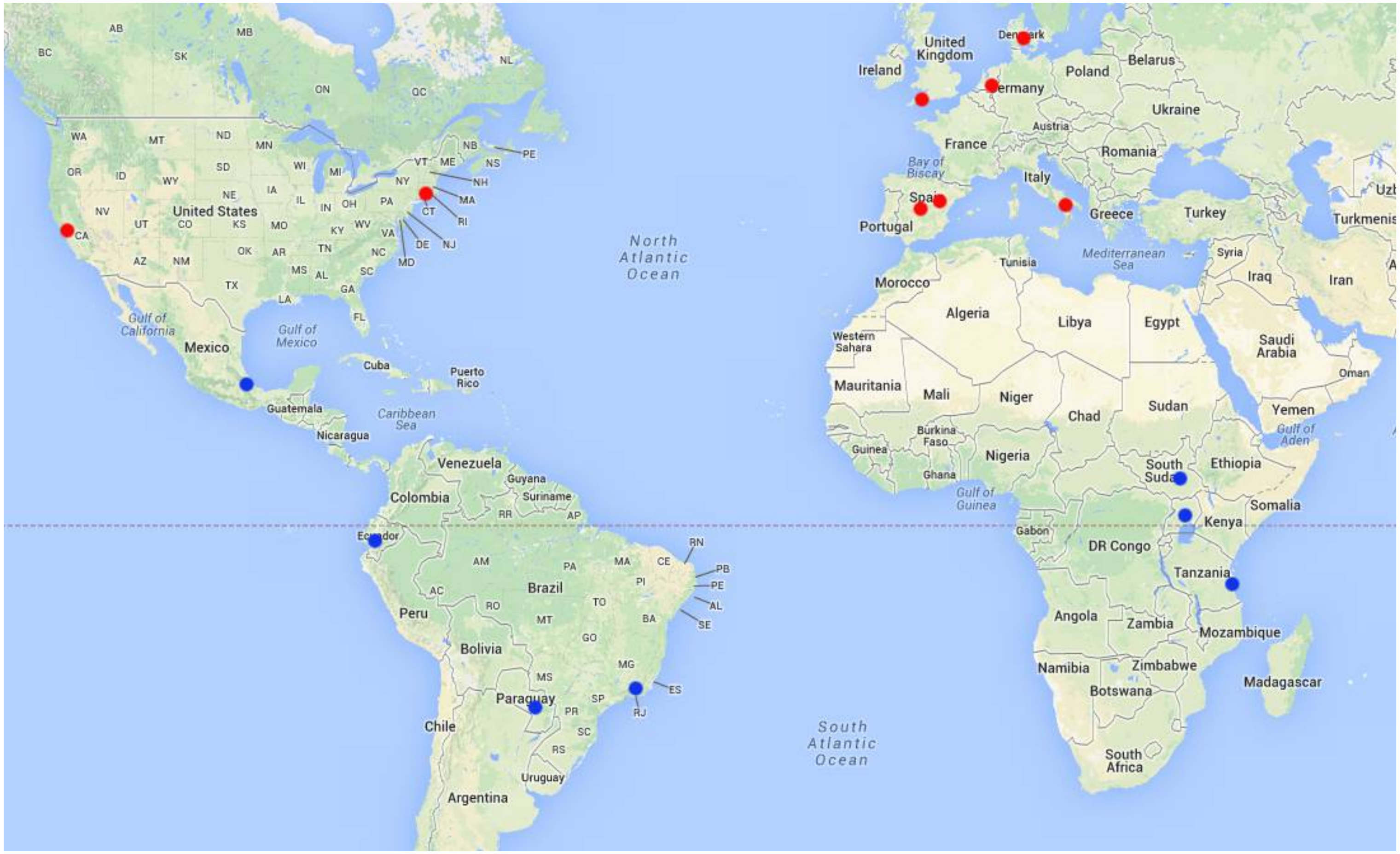
- ✔The use of a variety of new tools help students to understand that the same task can be carried out in many ways

Results and discussion: course evaluation

✔Since 2008, more than 400 students and technicians have participated in trainings organized by the authors of this paper

✔Trainings took place in Italy, Spain, Denmark, Peru, Brazil, Paraguay, Mexico, Ecuador, Uganda, the Republic of South Sudan, UK, the Netherlands, Tanzania and the USA

✔Background of course participants included some exposure to GIS, RS and/or database software but the Linux OS was mainly new



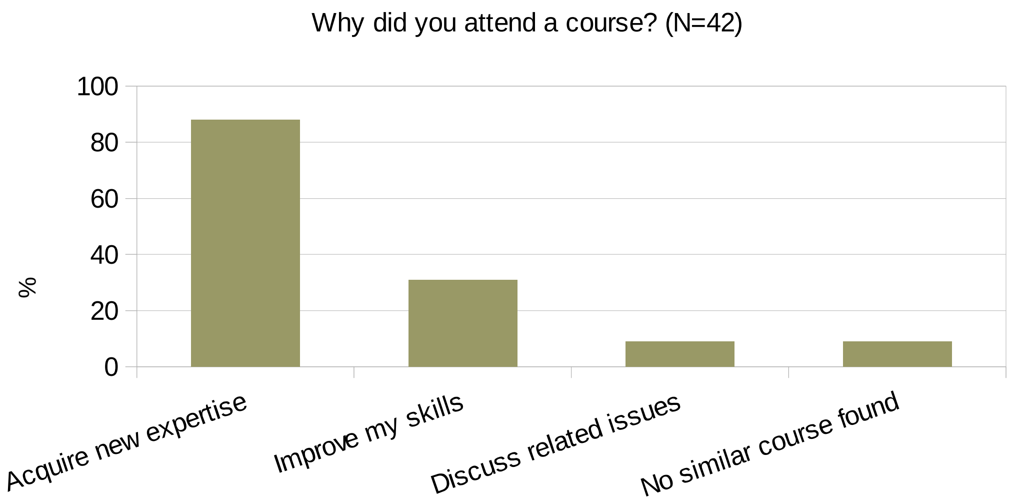
- ● Academic institutions
- ● National and international organizations

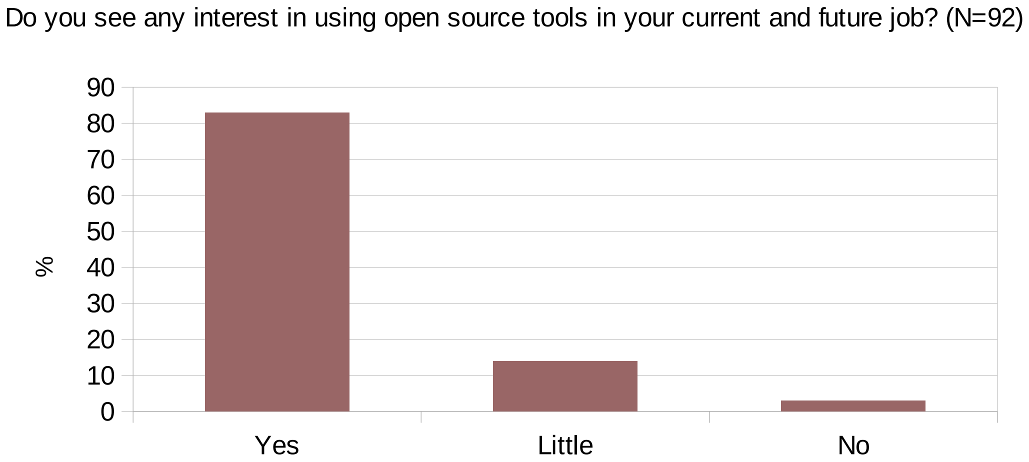
Results and discussion: course evaluation

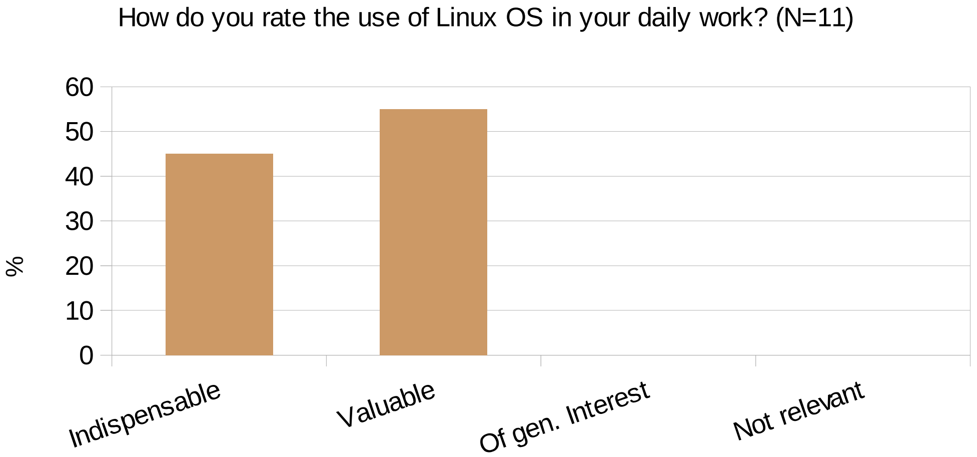
At the end of each course we asked the participants to fill in a questionnaire divided into four main parts:

1) Student’s motivation
2) Capacity to (independently) continue to develop skills in this area
3) Utility of using the new tools for their work
4) Course organization and effectiveness

Student's expectations and interest in learning open source geo-spatial programing







Results and discussion: course evaluation

Concerning the self learning attitude:

- ✔85% of the participants felt able to independently improve and learn more about open source tools
- ✔...in particular BASH (82%), probably due to its extensive usage during the exercises
- ✔72% of the participants felt they were able to independently improve their knowledge after the course
- ✔86% reported they would likely keep using the open-source command line

→ The course was able to transmit the importance and power of the command-line

Results and discussion: course evaluation

- ✔Concerning the suitability of OFGT and pktools to perform different image analysis tasks, we report a positive attitude among the course participants

- ✔Finally, concerning the pedagogic issues:
  - ✔78% participants felt the courses were well adapted to their needs, skills, and knowledge
  - ✔18% believed it was too difficult, and
  - ✔4% found the training boring and/or too basic

Results and discussion: course evaluation

- ✔Typical drawback has been the lack of time and therefore the participants wished to receive additional training on the subject

- ✔Thus, a relatively innovative course design was tested in the Netherlands in Dec 2013 - Jan 2014:
  - ✔A week of intensive face-to-face training (lectures and practicals) was followed by a month of self-study
  - ✔Participants could assimilate the course contents and apply them to datasets/problems related to their own research
  - ✔This put the participants in control of their own learning
  - ✔... and allowed them to focus on tools and techniques directly related to their work
  - ✔A second face-to-face training was organized to present specialized tools
  - ✔During this week, participants presented their works and obtained feedback

Results and discussion: course evaluation

- ✔Our long term follow-up reveals that the post-training impact depends on the institutional commitment

- ✔Students working in environments unfriendly to open-source software are less likely to adapt the tools for every-day problems
  - ✔One main factor being lack of software and hardware support

- ✔In other environments, it seems to be relatively easy to switch from proprietary software to open-source tools
  - ✔There the use of open source tools has also drastically increased among the colleagues of our students

Conclusion

✔Our experience shows that usage of open source geo-spatial tools on a much broader scale could be enhanced through increased formal course offerings as part of all geo-spatial science curricula

✔However, in the developing countries the situation is often challenging due to general lack of computing and internet infrastructure

✔Therefore, the selection of an appropriate approach (VM/Live OS) is a key for satisfactory user experience and a successful course

Conclusion

We are confident that with the continuing advances in technology, increased access to internet and the overall improvement in available computing resources, interest in open-source tools and computer programming will subsequently increase and, as such, will be a very important tool in the learning about and processing geospatial data.

Conclusion

We are confident that with the continuing advances in technology, increased access to internet and the overall improvement in available computing resources, interest in open-source tools and computer programming will subsequently increase and, as such, will be a very important tool in the learning about and processing geospatial data.

Thank you!