A comparison between free/open-source and proprietary geospatial software tools, based on a case study

Federica Migliaccio, Daniela Carrion, Cynthia Zambrano

DIIAR – Sezione Rilevamento, Politecnico di Milano, Italy, federica.migliaccio@polimi.it

Abstract

The objective of this study is to highlight advantages and disadvantages of the use of proprietary and free/open-source software for applications involving geospatial data. Based on a case study, we tried to compare some characteristics of the two categories of software, such as easy implementation and good usability (also based on training or tutorial material availability), analytical and display capability and functionality. The case study regards the realization of a GIS software toolbox in the frame of a project concerning the study of CO emissions caused by biomass burning on a global scale, for which it was requested to develop GIS tools in order to facilitate the management and spatial-temporal analysis of the data. Data consultation and processing was automated and the comparisons between different emission products was facilitated also for users who are not familiar with GIS environment. In fact, these tools allow the users to perform calculations in an intuitive and automatic way, and to define input and output parameters by means of a simple interface. At first, the software toolbox had been developed using ArcGIS 9.3 integrated with Python 2.5 language and the Matplotlib library. Then it was decided to migrate to a completely free/open source environment, namely GRASS version 6.4 developing new GRASS tools in C language. Future developments of the work should include comparison of ArcGIS to other free/open source GIS software, such as Open Jump or gvSIG, or comparison of open source GIS packages with one another.

1. Introduction

In this chapter the frame of the research presented in this paper will be outlined. The work has been developed within the INTERMEDE BBSO project *(Intercomparison of methods to derive global burnt biomass from satellite observations)*. The Politecnico di Milano co-operated with some of the participants to the project, designing and developing GIS tools, in order to allow automatic computation of parameters representing the statistical properties of the CO emission datasets and indices for the comparison of such datasets.

One must recall that biomass burning is a global phenomenon that occurs in any season, involving many types of vegetation and producing large quantities of gases and particles that play an important role in global change climate (Seiler and Crutzen, 1980). Many studies on this phenomenon have been

carried out over the years, and now the overall geographical and temporal distribution of fires can be observed in a systematic way and with good approximation, thanks to Remote Sensing technology.

The crucial point is represented by the fact that each research institution independently implements its algorithms for the computation of CO emissions, thus generating datasets and maps that are not immediately inter-comparable. The INTERMEDE inter-comparison exercise was especially set up for the purpose of highlighting differences and similarities between the various estimation models used to compute CO emissions starting from data collected from satellite sensors.

The results will hopefully help to understand which methods are better suited to derive maps of CO emissions, particularly on a local scale (i.e. geographical windows over continents or even smaller areas), for different periods (dry/wet season) of the year, also taking into account the correlation between the CO datasets and the land cover data. For this aim, the land cover map used as a reference during the project is GLC2000 - Global Land cover 2000 (Bartholomé and Belward, 2005).

The CO datasets considered in this project are five, all referring to the year 2003 and derived from observations collected by different satellite sensors. They are:

- ATSR (Kasischke et al., 2003), obtained from the Along Track Scanning Radiometer sensor on board the ERS-1 and ENVISAT satellites;
- VGTCOR (Michel et al., 2005), obtained from data coming from seven years of observations of SPOT – VEGETATION sensor;
- MODIS (Giglio et al., 2006), Moderate Resolution Imaging Spectroradiometer being a sensor on board the Terra and Aqua satellites; the derived dataset combines information from both active fires and burnt areas;
- ITO-PENNER (Ito and Penner, 2004), including data from the AVHRR (Advanced Very High Resolution Radiometer) sensor, on board the NOOA polar satellites;
- PETRON (Petron et al., 2004), including data from the MOPITT (Measurements of Pollution in the Troposphere) sensor, on board the Terra satellite (launched in 1999).

The datasets had been archived as raster grids with different spatial (1° × 1° or 0.5° × 0.5°) and temporal resolutions (daily or monthly), so in order to make it possible to compare them, they have been converted to a uniform spatial resolution of 1° × 1° and to a uniform temporal monthly resolution. Since the amount of data considered in the project is quite large, before the inter-comparison analysis and interpretation it was also necessary to synthesize the information and characterize the different datasets by computing the main statistical parameters on the basis of their spatial and temporal distribution and for each land cover, both at global and continental scale.

2. The comparison of CO emissions raster maps

The idea of developing automatic GIS tools mainly ensued considering that the participants to the CO emissions comparison project were faced with a quite large amount of data, in particular from the point of view of the variety of CO emissions datasets. In other words, five distinct products were being

considered for the inter-comparison experiment, but it was considered very likely that in the next future more (and different) datasets could become available, coming from data of different sensors or sensors combinations or covering different time periods. Naturally this reflected a demand for software tools allowing to repeat calculations in an automatic, efficient and easy way. Besides, since the CO emission products are geo-referenced data and can usually be represented in terms of raster maps, it seemed quite fitting to set the tools in a GIS environment.

Given such a frame for the inter-comparison project, at the beginning quite simple statistical tools were implemented such as a global agreement index which can be easily mirrored in a raster map representing the level of "agreement" between CO emission products over single cells of 1° × 1° covering the whole earth surface (see Fig. 1): each cell of this raster map represents the value corresponding to the number of products that detect CO emission in that particular cell (for a reference, see Migliaccio and Pinto, 2009).

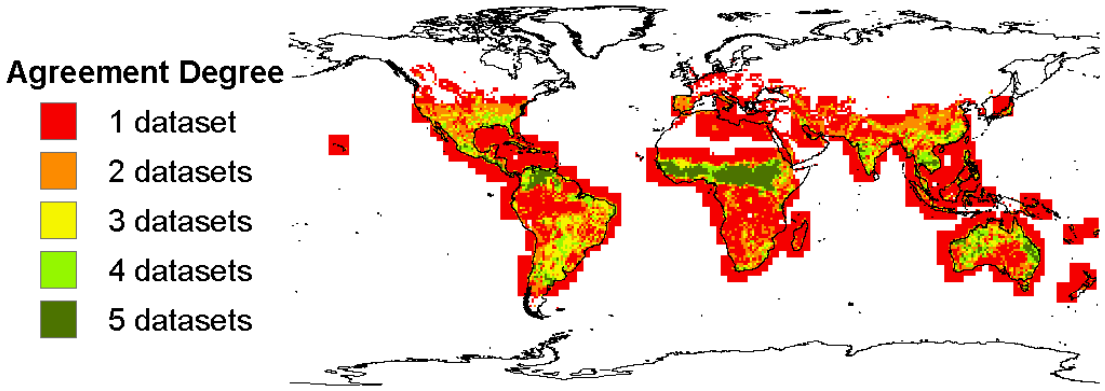

Figure 1. An "agreement map" showing the agreement between different CO emission datasets

Subsequently, statistical indices representing in other ways the agreement between couples of datasets were computed, both on a global or regional scale, always with temporal resolution of one month, starting from the very well known correlation coefficient, which measures the correlation between two datasets X and Y, σ_{XY} being the covariance and σ_X, σ_Y the standard deviations of the two datasets, see for example (Mood, Graybill and Boes, 1974):

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad [1]$$

The correlation coefficient is dimensionless and its value may range between -1 and 1.

Other indices which were singled out as suitable in the context of the research are described in the following.

The agreement coefficient suggested by (Ji and Gallo, 2006), which "measures" the agreement between two different datasets (or raster maps) X and Y, on the hypothesis that both datasets may be equally subject to measurement errors (in our case, also equally subject to errors due to the model used), X_i, Y_i being the values of corresponding pixels in the two datasets, $\overline{X}$ and $\overline{Y}$ being the mean values of the two datasets (note that this index is dimensionless and upper bounded by the value 1, which represents perfect agreement between the two datasets):

$$AC = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}\left(\left|\overline{X} - \overline{Y}\right| + \left|X_i - \overline{X}\right|\right)\left(\left|\overline{X} - \overline{Y}\right| + \left|Y_i - \overline{Y}\right|\right)} \qquad [2]$$

Mielke's measure of agreement, which is actually based on the measure of the mean square error between two datasets X and Y, X_i, Y_i being the values of corresponding pixels in the two datasets (note that this index is dimensionless and upper bounded by the value 1, which represents perfect agreement between the two datasets):

$$\rho = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - Y_j)^2} \qquad [3]$$

Robinson's coefficient of agreement, which measures the perpendicular distance between point (X_i, Y_i) and the $X = Y$ line, X_i, Y_i being the values of corresponding pixels in the two datasets and $\overline{Z}$ being the mean value of $\overline{X}$ and $\overline{Y}$ (note that this index is dimensionless and its value ranges between 0 and 1):

$$A = 1 - \frac{\sum_{i=1}^{n}(X_i - Z_i)^2 + \sum_{i=1}^{n}(Y_i - Z_i)^2}{\sum_{i=1}^{n}(X_i - \overline{Z})^2 + \sum_{i=1}^{n}(Y_i - \overline{Z})^2} \qquad [4]$$

A reference for Mielke's measure of agreement and Robinson's coefficient of agreement can also be found in (Ji and Gallo, 2006).

Now the concept was to develop a set of interactive geospatial tools, not only for the computation of the above described indices and of other statistical measures, but also for their graphical

representation and in general for the handling of CO emission raster data. One fundamental issue was that such geospatial tools had to be easily exploited even by novice GIS users.

3. The implementation of tools for CO emission maps comparison in a GIS proprietary environment

At the beginning, the packages used for the implementation of the tools were the ESRI ArcGIS suite of applications and the *open-source* language Python. In particular, it was decided to exploit the ESRI ArcGIS ModelBuilder application for designing and implementing the procedures, in all cases, when it was possible. ModelBuilder enables to create “models”, i.e. simple diagrams that describe the flow of procedures, containing the necessary information for the execution and creation of output data and providing helpful documentation. However, one drawback of this application is that in some cases these “models” are not enough customizable. An example of this assertion is represented by the management of iteration cycles, which are an essential issue in a case like the one presented in this paper, where computations must be repeated for each available dataset or couple of datasets, for each time period for which observations are delivered (usually, monthly periods), for different geographical windows and possibly for different land cover distributions. For this reason, it was necessary to integrate the ModelBuilder application with custom Python code scripts, which allowed to extend the default functions of ArcGIS. Besides, also functions from the open-source library Matplotlib were integrated, obtaining as a result the automatic generation of graphs at each iteration performed.

In total, 50 tools were realized and these instruments were collected in a toolbox in order to facilitate the portability among different users or computers. A simple graphic interface was prepared for each tool, to make the insertion of input data easier and also to suggest possible default values.

Finally, a window for the display of the operations in progress was provided during each tool run-time, and for each tool an extensive documentation in HTML format was included, containing text, pictures and diagrams, which are accessible to the user at any time.

4. The GIS tool implemented in an open-source environment

The choice described in the previous paragraph, namely to mainly develop the software tools in the proprietary ESRI ArcGIS environment, was driven by the requests of the first toolbox users. In fact, some of the researchers involved in the INTERMEDE project explicitly asked for a proprietary solution. However, the increasing interest of the participants in the project and some considerations about the usability of the GIS tools which were growing in number and in level of complexity, led to the decision to migrate to a free/open source environment. This could allow to distribute the CO emission comparison toolbox to a larger number of research groups, regardless of the fact that they owned or not an ESRI ArcGIS licence.

Several open-source GIS software packages were considered, taking into account not only their characteristics, but also their level of development and adoption within the scientific community. In the

end, the GRASS GIS 6.4.0 RC6 was chosen (Fig. 2), mainly because it can be run in many OS. Indeed GRASS is developed in a UNIX environment and is ported to many other systems like MS-Windows (NT/2000/XP) for the experimental winGRASS port, and MacOS X. Moreover quite advanced software libraries for the processing of raster data are available. Besides, software tools created in the GRASS environment will have the possibility to be subsequently also called in QuantumGIS, which is at the moment another largely popular free/open-source GIS software, having a “user friendly” interface.

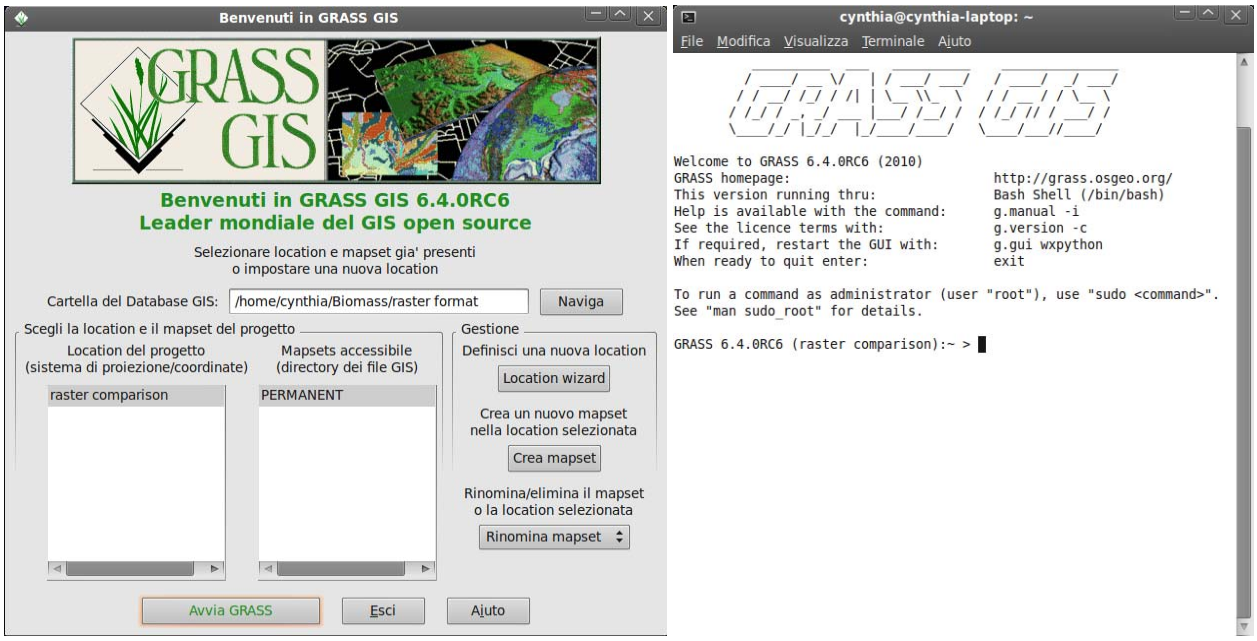

Figure 2. GRASS GIS 6.4.0 RC6 startup screen

As a first step, all the data used for the INTERMEDE experiment were imported and visualized in the GRASS GIS environment As an example, Fig. 3 to 7 show the global CO emission maps for the month of August 2003, for the five emission products considered.

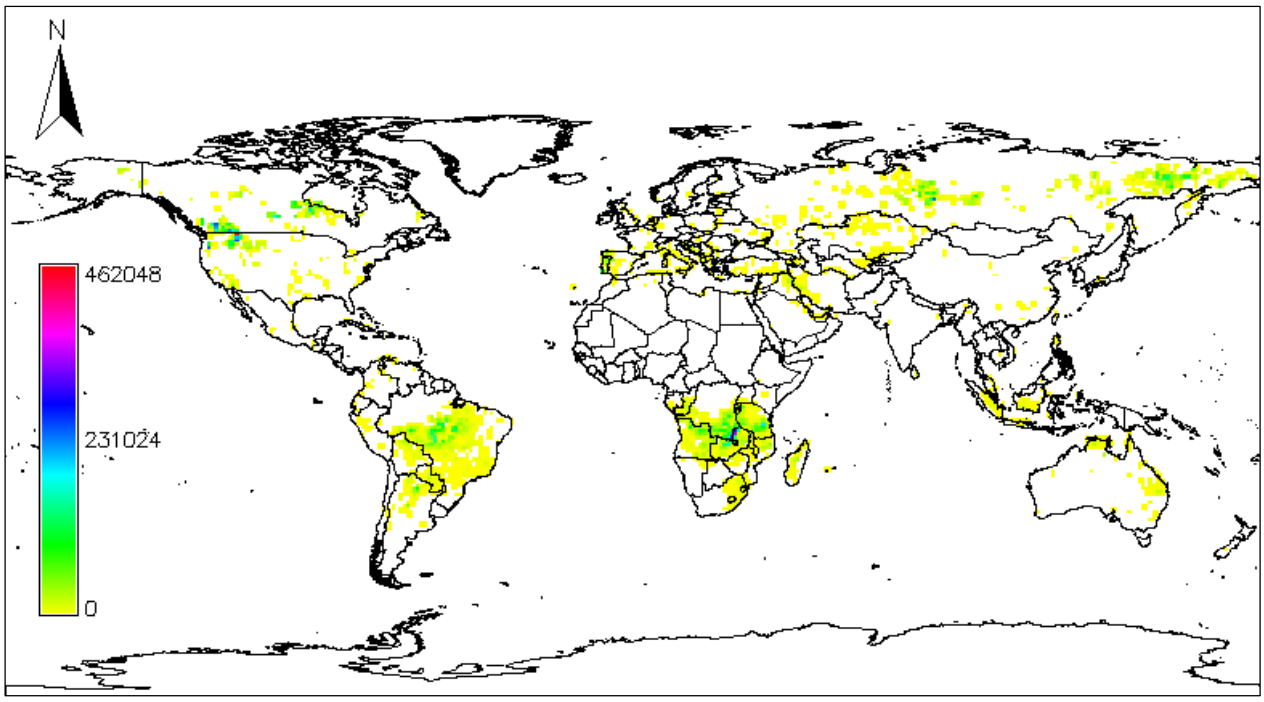

Figure 3. ATSR CO emission map for August 2003; colour scale in tons of CO emissions

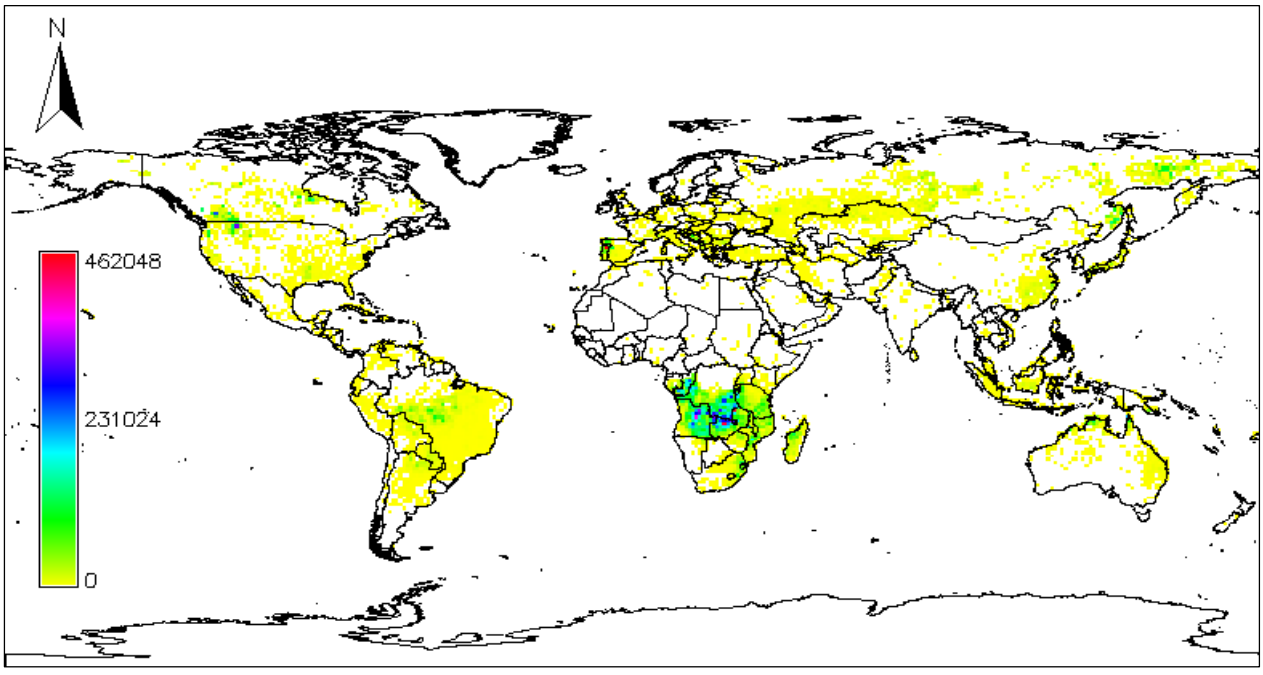

Figure 4. MODIS CO emission map for August 2003; colour scale in tons of CO emissions

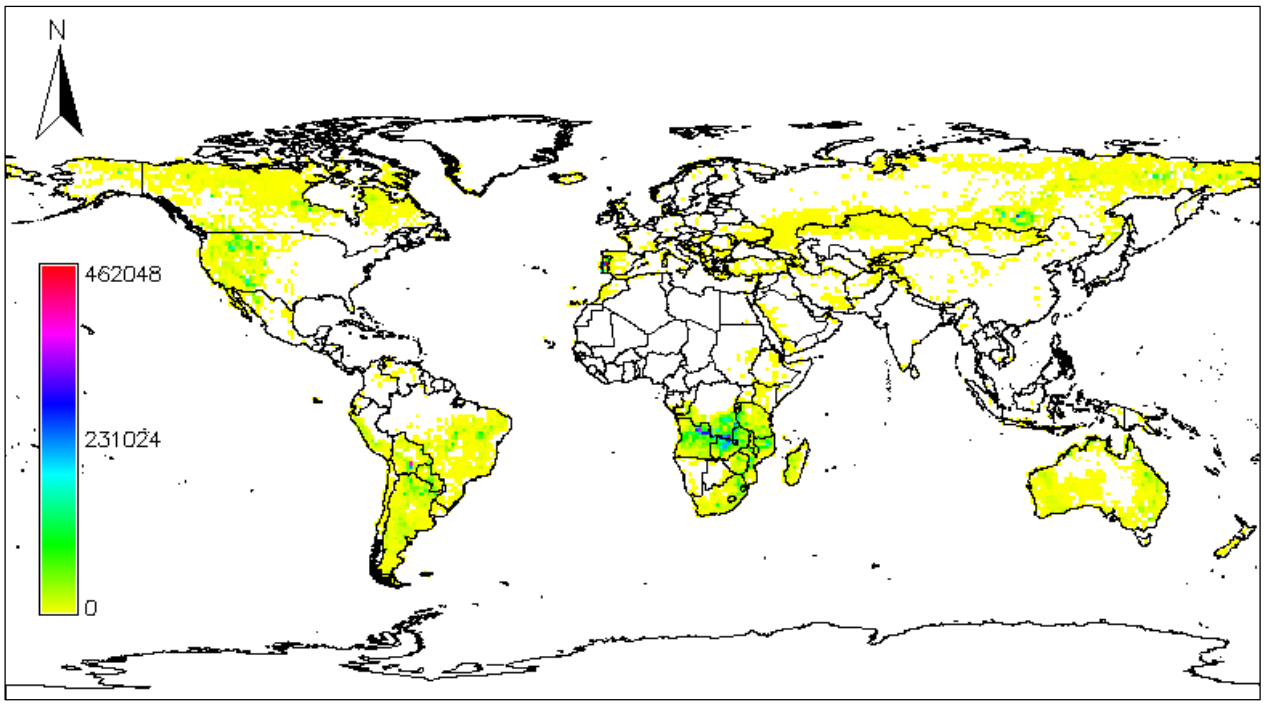

Figure 5. VGTCOR CO emission map for August 2003; colour scale in tons of CO emissions

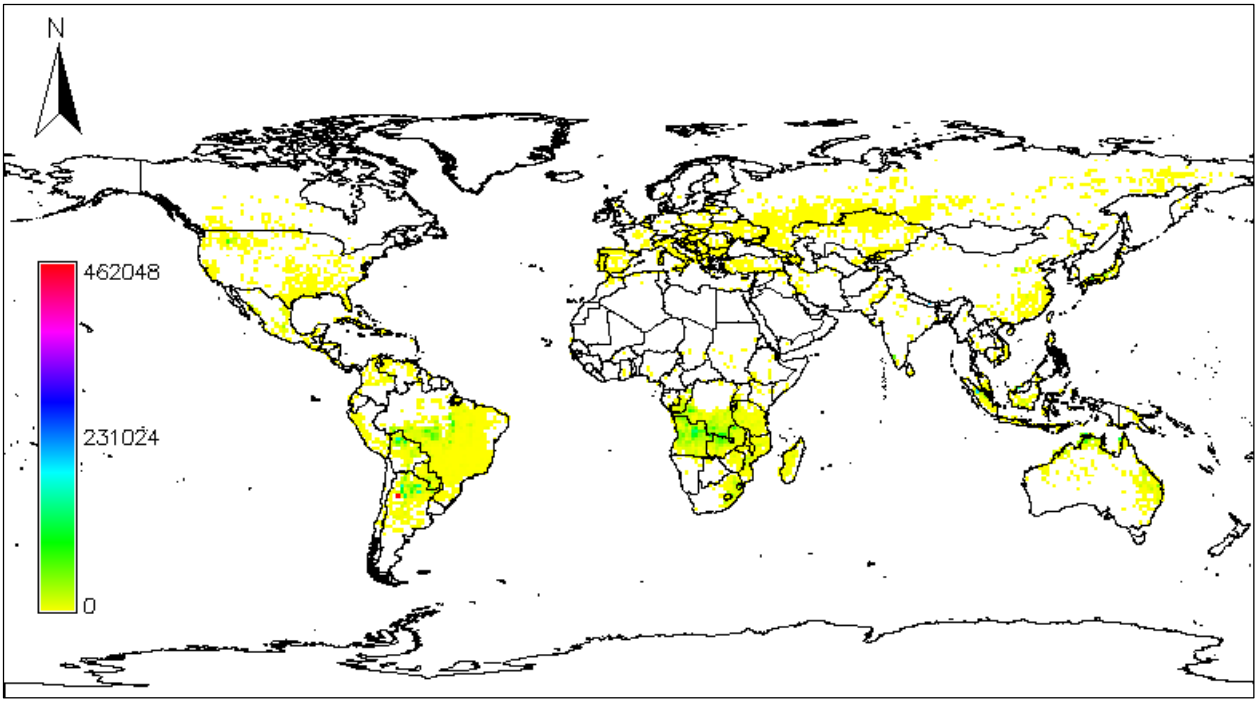

Figure 6. ITO PENNER CO emission map for August 2003; colour scale in tons of CO emissions

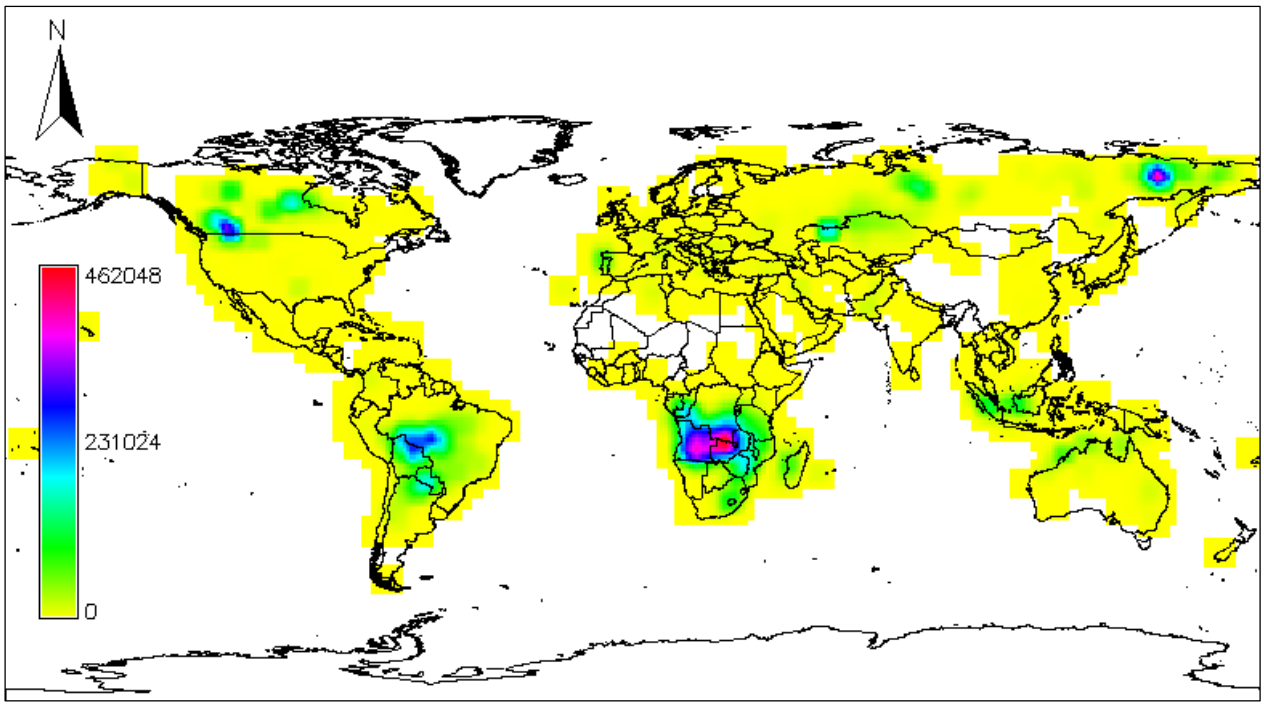

Figure 7. PETRON CO emission map for August 2003; colour scale in tons of CO emissions

For importing data to GRASS, the GDAL (Geospatial Data Abstraction Library) OGR 1.7.2 release was exploited. GDAL is a translator library for raster geospatial data formats released under an X/MIT style Open Source license by the Open Source Geospatial Foundation.

The tools developed in ArcGIS for the analysis and comparison of CO emission data had then to be ported to the GRASS GIS environment. The reference for the creation of such tools was represented by the GRASS 6 Programmer's Manual (see GRASS Development Team, 2010b). The source code of the r.covar tool (developed by Michael Shapiro of the US Army CERL) helped defining the main structure of the routine developed for the computation of the comparison coefficients for each couple of datasets. This tool has been valuable also in order to identify useful functions for the loading and processing of raster maps.

The new GRASS tool r.compare has been written in C language: it enables to compute all the coefficients defined in Paragraph 2 (Ji and Gallo agreement coefficient AC, Mielke's measure of agreement and Robinson's coefficient of agreement). In particular, r.compare provides an easy interface which helps the users to enter input data (see Fig. 8) and to visualize the corresponding documentation in HTML format.

The results are organized in an N × N matrix, where N represents the number of CO emission maps for which the chosen coefficients are computed. The output can be visualized on the screen and saved in an ASCII file. As a test all coefficients have been computed for the global maps of CO emissions for the month of August 2003. In this specific case, the results highlighted very dissimilar behaviours

between all the couples of CO emission products. This is probably due to the large variability shown by the phenomenon of biomass burning at global and even at continental scale.

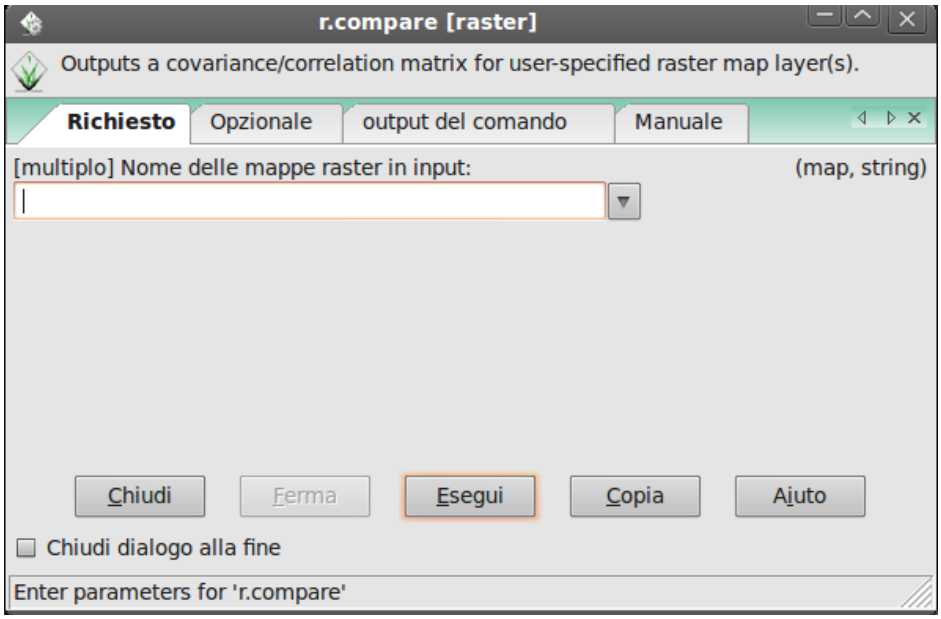

Figure 8. The r.compare module GUI (for the time being the dialog boxes have been written in Italian)

For this reason, it was decided that it could be useful to implement the computation of the statistical indices also for more limited geographical windows, which any user may decide to adapt to his/her specific purposes. For example, in Fig. 9 the map of active cells for the five emission products (month: August 2003) is shown: here the active cells are plotted in greyscale, meaning that the darker is the cell, the higher is the number of products showing emission of CO. Looking at such a map, it is easy to identify at least two relatively limited areas (window 1 in South America and window 2 in Africa) where biomass burning appears to be substantial. For these areas (set as regions in GRASS) it is particularly important to understand if the CO emission products are in good agreement and this can be verified by computing both the correlation coefficient and other comparison coefficients.

Examples of results (again, for the month of August 2003) are reported in Table 1 to Table 4.

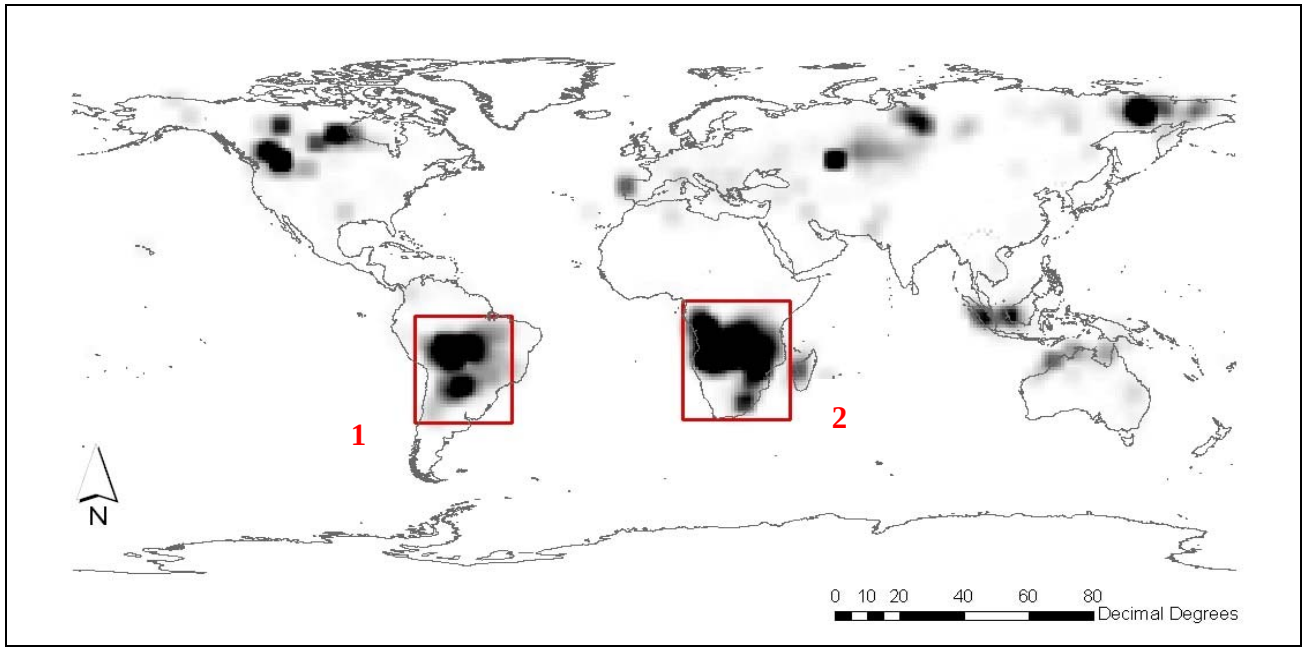

Figure 9. Example of selection of small “regions” in GRASS for the CO emission data comparison

In particular, Table 1 reports the values of the correlation coefficient computed by means of the r.covar tool included in the GRASS GIS software package, while for the computation of the other indices (Table 2 to Table 4) the r.compare tool was used. At the moment the latter tool can only be launched from a command line.

Finally, Table 5 reports a synopsis of the results obtained for the different coefficients.

Of course this analysis can be repeated for each month of the year or any other time period and for any geographical window suitably chosen.

The interpretation of the results of the comparison among the different products is not the purpose of this paper, anyway it is possible to see that the correlation/agreement between the CO emission maps is not high. It is possible to state that the development of tools allowing the researchers to explore where and when the CO emissions datasets provide similar or dissimilar information can be very useful.

Table 1. Correlation coefficient for all couples of CO emission products (window 2, Aug. 2003)

Correlation	**ATSR**	**MODIS**	**VGTCOR**	**ITO-PENNER**	**PETRON**
ATSR	1.000	0.667	-0.005	0.191	0.305
MODIS		1.000	0.033	0.181	0.483
VGTCOR			1.000	0.158	0.050
ITO-PENNER				1.000	0.146
PETRON					1.000

Table 2. Agreement coefficient AC for all couples of CO emission products (window 2, Aug. 2003)

AC	ATSR	MODIS	VGTCOR	ITO-PENNER	PETRON
ATSR	1.000	-0.444	-3.791	-2.866	-2.331
MODIS		1.000	-2.011	-2.351	0.287
VGTCOR			1.000	-4.664	-5.983
ITO-PENNER				1.000	-6.680
PETRON					1.000

Table 3. Robinson's coefficient for all couples of CO emission products (window 2, Aug. 2003)

Robinson	ATSR	MODIS	VGTCOR	ITO-PENNER	PETRON
ATSR	1.000	0.591	0.496	0.593	0.599
MODIS		1.000	0.483	0.513	0.544
VGTCOR			1.000	0.577	0.517
ITO-PENNER				1.000	0.543
PETRON					1.000

Table 4. Mielke's coefficient for all couples of CO emission products (window 2, Aug. 2003)

Mielke	ATSR	MODIS	VGTCOR	ITO-PENNER	PETRON
ATSR	1.000	0.242	-0.005	0.188	0.201
MODIS		1.000	0.013	0.062	0.261
VGTCOR			1.000	0.155	0.034
ITO-PENNER				1.000	0.085
PETRON					1.000

Table 5. Values of the comparison coefficients computed by r.covar and r.compare

Dataset X	Dataset Y	Correlation	AC	Robinson	Mielke
ATSR	ITO PENNER	0.191	-2.866	0.593	0.188
ATSR	MODIS	0.667	-0.444	0.591	0.242
ATSR	PETRON	0.305	-2.331	0.599	0.201
ATSR	VGTCOR	-0.005	-3.791	0.496	-0.005
MODIS	ITO PENNER	0.181	-2.351	0.513	0.062
MODIS	PETRON	0.483	0.287	0.544	0.261
MODIS	VGTCOR	0.033	-2.011	0.483	0.013
VGTCOR	ITO PENNER	0.158	-4.664	0.577	0.155
VGTCOR	PETRON	0.050	-5.983	0.517	0.034
ITO PENNER	PETRON	0.146	-6.680	0.543	0.085

5. Discussion and future developments

In this paper, an example of porting a GIS application from the proprietary to the free/open source environment has been presented. GIS tools for the analysis of maps in raster format representing CO emission caused by biomass burning, which had previously been developed in ArcGIS 9.3 integrated with Python 2.5, have now been party implemented in GRASS 6.4.

Although the migration of the whole project to GRASS has not yet been completed, this preliminary experience has allowed drawing some considerations and can be taken as a test bed to evaluate whether it could be worthwhile to completely migrate to a free/open source platform.

CO emission data consultation and processing was automated as much as possible, trying to develop a “user friendly” tool for users that are not particularly acquainted with GIS software, which means that intuitive interfaces had to be prepared to facilitate the data input and the output of results, however this was not possible in all cases.

A first consideration (quite obvious indeed) is that going from Python to C code was at the beginning not trivial, actually just from the point of view of the software philosophy, Python being an interpreted and C a compiled programming language.

Nevertheless, the GRASS software libraries were found to be very helpful and well documented, which is an issue not to be underestimated. Regarding the possible sources of help for newcomers in the GRASS environment, also the role of the on-line community through several mailing lists can be cited (such as gfoss and grass-dev). Besides, the starting step, requiring to import CO emission raster data from the ArcGIS grid format to GRASS, was quite facilitated by exploiting the GDAL library.

The ArcGIS software is well documented, so that the user is able to understand the use of each tool. Often references to algorithms are provided, but obviously it is not possible to see how the algorithms have been implemented. The GRASS software is widely documented, but the documentation that can be found online is so much that sometimes it is difficult to choose to which one refer to, in particular for the specific version in use. This remark is in our opinion true especially for the installation guide, while the tools documentation is sometimes a bit too concise and it can be necessary to access the source code to understand which algorithm has been implemented.

Although GRASS is available in many operating-systems, one drawback is represented by the fact that some GRASS functionalities are only available when working on UNIX or Linux systems (in the present case, it was Ubuntu 10.04), thus reinforcing the belief that free/open source GIS software is most effective and displays the maximum of its capabilities when running in a UNIX or Linux environment.

From the point of view of the user of the tools implemented in this case study, the interface of the tools, both in ArcGIS and in GRASS, is very similar. Besides, also the display of the results through layouts can be considered as equally user friendly, in particular thanks to the GRASS Wx Python GUI.

Finally, while new GRASS tools can be created and directly integrated in the software, this possibility is not available to the user of the default ArcGIS environment without e.g. the ESRI Developer Network (EDN) add-on package which gives access to ArcEngine that allows to implement new tools.

At the moment, the work has not yet been completed, and more tools are to be implemented in order to enable to graphically visualize some of the numerical results in the form of charts and to produce maps from other types of data. Indeed, since GRASS can be interfaced with R, it will be studied how to take advantage of this possibility in the frame of the CO emission data inter-comparison research. The interface between GRASS and R can be created by means of the spgrass6 package, which provides facilities for using all GRASS commands from the R command line.

Finally, besides increasing the number of functionalities available in the new GRASS tool, it will be investigated how to improve it by porting the functions implemented by the ArcGIS ModelBuilder to the GRASS environment, exploiting, when possible, the capabilities of the Wx GUI-modeler3 (see Fig. 10). In fact, this application is a Wx GUI extension which allows the user to create, edit, and manage models. It has to be noted that at the moment this tool is still under development and distributed as an experimental prototype.

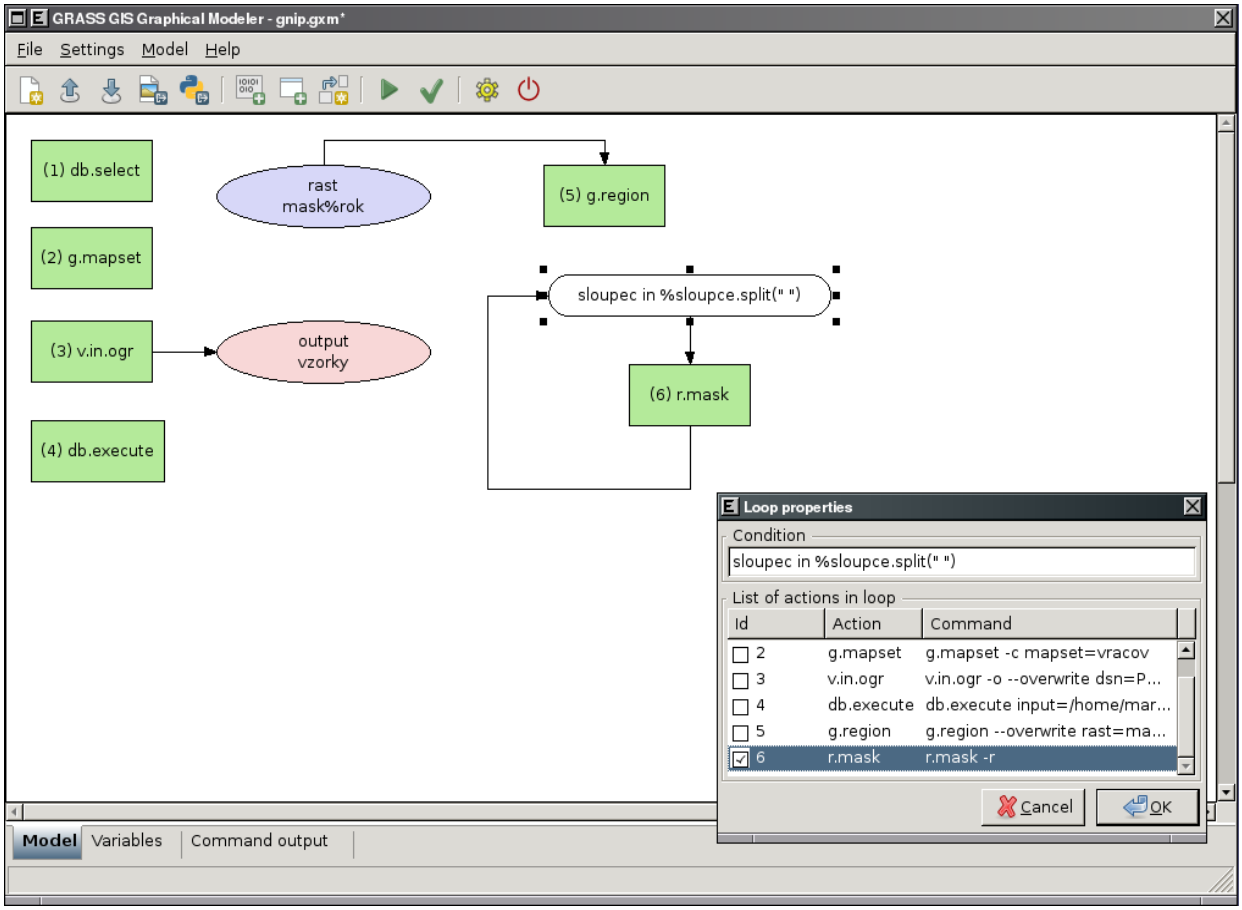

Figure 10. The graphical modeler Wx GUI-modeler3: an example of a model to define loops

References

Full-text journal article

Bartholomé E & Belward A 2005, 'GLC2000: a new approach to global land cover mapping from Earth observation data', *International Journal of Remote Sensing*, vol. 26 (9), pp. 1959-1977.

Boschetti L, Brivio PA, Eva HD & Grégoire JM 2004, "Lessons to be learned from the comparison of three satellite-derived biomass burning products", *Geophysical Research Letters* 31-L21501.

Giglio L, Van der Werf GR, Randerson JT, Collatz GJ & Kasibhata P 2006, "Global estimation of burned area using MODIS active fire observations", *Atmospheric Chemistry and Physics*, vol. 6, pp. 957-974.

Ito A & Penner JE 2004, "Global estimates of biomass burning emissions based on satellite imagery for the year 2000", *Journal of Geophysical Research.*, 109, D14S05, doi:10.1029/2003JD004423.

Jain AK 2007, "Global estimation of CO emissions using three sets of satellite data for burned area", *Atmospheric Environment* 41- 6931-6940.

Ji L & Gallo K 2006, “An agreement coefficient for image comparison”, *Photogrammetric Engineering and Remote Sensing*, Vol. 72, N. 7, pp.823 – 833.

Kasischke ES, Hewson JH, Stocks B, Van der Werf G & Randerson J 2003, “The use of ATSR active fire counts for estimating relative patterns of biomass burning - a study from the boreal forest region”, *Geophysical Research Letters*, 30, doi:10.1029/2003GL017859.

Michel C, Liosse C, Grégoire JM, Tansey K, Carmichael GR & Woo JH 2005, “Biomass burning emission inventory from burnt area data given by the SPOT-VEGETATION system in the frame of TRACE-P and ACE-Asia campaigns**”**, *Journal* of *Geophysical Research* 110-D09304.

Migliaccio F & Pinto L 2009, “Experiences in the automatic validation and cross-validation of spatial datasets and raster maps”, Proceedings of the International Workshop on validation of geo-information products for crisis management - Valgeo 2009, pp. 121 – 126. (JRC, Ispra, Italy, Nov. 23 – 25, 2009), ISBN 978-92-79-14069-3.

Mood AM, Graybill FA & Boes DC 1974, Introduction to the theory of statistics, McGraw Hill.

Pétron G, Granier C, Khattotov B, Yudin V, Lamarque JF, Emmons L, Gille J & Edwards DP 2004, “Monthly CO surface sources inventory based on the 2000-2001 MOPITT satellite data”, *Geophysical Research Letters*, 31, L21107, doi:10.1029/2004GL020560.

Seiler W & Crutzen PJ 1980, “Estimates of gross and net fluxes of carbon between the biosphere and the atmosphere from biomass burning”, *Climatic Change*, 2, 207-247.

Tansey K, Grégoire JM, Defourny P, Leigh R, Pekel JF, Van Bogaert E & Bartholomè E 2008, “A new, global, multi-annual (2000–2007) burnt area product at 1 km resolution”, *Geophysical Research Letters*, 35-L01401.

Web documents

Roger Bivand, spgrass6 - R package version 0.6-16, *Interface between GRASS 6 and R*, 2010, <http://CRAN.R-project.org/package=spgrass6>

Web sites

Geospatial Data Abstraction Library
<http://www.gdal.org/>

Gfoss mailing list
<http://lists.faunalia.it/cgi-bin/mailman/listinfo/gfoss>

GRASS Development Team, 2010a. Geographic Resources Analysis Support System (GRASS) Software, Version 6.4.0. Open Source Geospatial Foundation.
<http://grass.osgeo.org>

GRASS Development Team, 2010b. Geographic Resources Analysis Support System (GRASS) Programmer's Manual. Open Source Geospatial Foundation.
<http://download.osgeo.org/grass/grass6_progman/ >

GRASS-dev mailing list
<http://lists.osgeo.org/mailman/listinfo/grass-dev>

Matplotlib library
<http://matplotlib.sourceforge.net/index.html>

WxGUI_Modeler
<http://grass.osgeo.org/wiki/WxGUI_Modeler>