Free GIS Software meets zoonotic diseases: From raw data to ecological indicators

Markus Neteler
Fondazione Mach, Centre for Alpine Ecology, 38100 Viote del Monte Bondone (Trento), Italy,
neteler@cealp.it

Abstract

Emergence and spread of infectious diseases in a changing environment require the development of new methodologies and tools for risk assessment, early warning and policy making. Of special interest are zoonotic diseases that are able to be transmitted from animals to humans, often by a vector. Zoonotic diseases cause major health problems in many countries, they are driven by environmental changes, pathogen changes as well as political and cultural changes.

GIS modelling is routinely used to perform risk assessment for the prevention of these diseases. In combination with available geospatial data, a new quality of ecological indicators can be extracted from new high temporal resolution satellite data time series. Especially the Moderate Resolution Imaging Spectroradiometers (MODIS sensor) which are flown on two satellites, deliver an almost complete Earth coverage four times a day at different resolutions (from 250m to 1km pixels). MODIS data constitute the base for the new indicators.

Free GIS Software (Open Source GIS) is an appropriate choice for scientific computing as the source code is published under a Free Software license with the rights to run the program for any purpose, to study how the program works, to adapt it, to redistribute copies including modifications. Available software packages greatly support the processing of large amounts of geospatial data to generate the ecological indicators oriented toward landscape epidemiology as proposed in this paper.

1. Introduction

Zoonotic diseases are infectious diseases that are able to be transmitted from animals to humans, often by a vector (e.g, ticks, mosquitoes). Both wildlife (e.g, roe and red deer, rodents) and domestic animals are reservoirs for zoonoses. Zoonoses involve all types of agents (bacteria, parasites, viruses and others). Zoonoses constitute a major public health problem in many countries (Kruse et al., 2004). While they have been known since many centuries, several of them are (re-)emerging, i.e., they increase in incidence or they expand in geographical, host or vector range. Environmental changes, pathogen changes as well as political and cultural changes are driving the spread of emerging diseases (Kuhn et al., 2004; Sumilo et al., 2008).

In order to better understand the spatial distribution of zoonoses, in particular rodent-, tick- and mosquito-borne diseases, GIS modelling is routinely used to perform risk assessment for the prevention of these diseases. Related studies aim at the identification of relevant patterns in spatial

ISBN 978-0-620-42117-1

and temporal time series (Hay et al., 2000). Data are integrated from heterogeneous databases which contain common GIS data as well as new remote sensing data. In particular, the integration of remote sensing data from new sensors in combination with spatial data analysis helps to understand the epidemiology of diseases and to improve disease prevention and control. Like other vectors, ticks and tick-borne disease systems are highly susceptible to changes in environmental conditions, including abiotic and biotic factors (Randolph, 2006). This suggests that continuous and high frequency monitoring is needed to observe the emerge of infectious diseases. Especially environmental satellite data can fulfil the data demand at reasonable costs.

Satellite data have been successfully used for more than two decades in health and epidemiological applications. Being crucial for providing detailed descriptions of the current environmental conditions, they permit to improve the understanding of disease patterns (Hay et al., 2000). A list of potential links between remotely sensed indices and infectious diseases is given in Beck et al. (2000). The Moderate Resolution Imaging Spectroradiometer (MODIS sensor) plays an increasing role in the provision of environmental data from space. MODIS can be considered as a much enhanced successor of the AVHRR instrument onboard a series of NOAA satellites. MODIS improves upon the performance of AVHRR by providing both higher spatial resolution and greater spectral resolution.

The Terra and Aqua satellites with their MODIS sensors provide four global coverages per day at pixel resolutions from 250m to 1000m. The high temporal resolution helps to compensate for the relative low spatial data resolution. The more than 40 MODIS products from NASA encompass land surface temperature (LST), vegetation indices (NDVI/EVI), surface reflectance and maximum snow extent. From daily LST maps, derivable indicators include temperature minima and maxima at annual/monthly/decadal periods, the identification of late frost periods, unusual hot summers, growing degree days, spring temperature increase and autumnal temperature decrease (Neteler, 2005; Rizzoli et al., 2007; Carpi et al., 2007). EVI permits to detect seasonal vegetation differences, spring/autumn detection and the length of growing season. Furthermore, the Normalized Difference Water Index (NDWI) can be calculated. While the raw data are of limited interest to landscape epidemiological applications, time series aggregation of the new sensor data leads to a new quality of ecological indicators which have not been available earlier.

The creation of ecological indicators focuses on new satellite sensors which are yet rarely used for landscape epidemiological studies (Beck et al., 2000; Tatem et al., 2004; Herbreteau et al., 2005). A special challenge is the complex terrain as it dominates the study area in Northern Italy. It requires special attention to data processing and outlier detection.

The sheer amount of data requires adequate software and computational resources such as clusters. Free Software GIS (Open Source GIS) tools permit to execute parallel batch jobs to process large amounts of satellite data on high performance computing (HPC) facilities, even without modifications to the source code. Free Software GIS is an appropriate choice for scientific computing as the source code is published under a Free Software license with the rights to run the program for any purpose, to study how the program works, to adapt it, to redistribute copies including modifications. Data preparation often requires a substantial amount of work regarding

data reformatting, geocoding, reprojection and quality assessment. Free GIS software tools (PROJ4, GDAL/OGR, GRASS GIS etc.), are well suited for the entire work flow and provide all needed methods and algorithms (Neteler & Mitasova, 2008). Based on this tool set, raw GIS and remote sensing data can be transformed into ecological indicators as used in epidemiological and risk assessment models.

Further integration with machine learning techniques (e.g., *randomForest* in the R Language and Environment for Statistical Computing) permits to perform variable selection and ranking from a larger catalogue of indicators (Furlanello et al., 2003). Based on significant variables and further GIS data, static or dynamic absence/presence maps and risk models are developed (Benito Garzòn et al., 2006).

2. Material and methods

2.1 Study area

The set of indicators was developed in the provinces of Trento and Belluno, a mountainous region covering an area of 9850 km^2 located in north-eastern Italy. The climate is temperate-oceanic and 70% of the territory lies above 1,000 m a.s.l. and 55% is covered by coniferous and deciduous forests. The rugged terrain leads to shading effects, presence of haze in the valleys and locally elevated cloud contamination which requires special attention when processing satellite data.

2.2 Data collection and preprocessing

Besides common cartographic GIS layers, we obtained a time series of Moderate Resolution Imaging Spectroradiometer (MODIS, production level V004) data through the NASA EOS Data Gateway (http://edcimswww.cr.usgs.gov/pub/imswelcome/). We received Land Surface Temperature (LST) maps and surface reflectance maps with daily temporal resolution, 8 days snow extent maps and Normalized Difference/Enhanced Vegetation Index (NDVI/EVI) maps aggregated to 16 days.

From March 2000 to May 2002, only the Terra satellite was operative, delivering daily two overpasses (at 10:30, and 22:30, local time). With the subsequent launch of the Aqua satellite two more overpasses (at 13:30, and 01:30, local time) became available. We used daily LST and surface reflectance to have the possibility of data aggregations at different temporal scale. NDVI/EVI and snow maps with slower temporal dynamics were taken as 16 days composition maps to minimize cloud contamination. All data are delivered in Sinusoidal projection, the LST maps (in Kelvin) at 1x1 km^2 pixel resolution, surface reflectance and snow maps at 500x500 m^2 pixel resolution. All maps were reprojected to the Italian Gauss-Boaga cartographic system using the MODIS Reprojection Tool (MRT V 3.3 from USGS). The LST maps were converted from Kelvin to degree Celsius.

All MODIS products used in this study consist of the data layer and a related quality assurance (QA) map. These QA maps are a bit-pattern encoded quality indications which are used to filter out poor quality pixels. For the LST maps, an additional outlier/cloud detector was developed and applied to eliminate remaining cloud-contaminated pixels not discovered by the NASA QA

algorithms that were used during production of the LST maps. There is indication that the recent reprocessing of all MODIS data by NASA to production level V005 reduces these problems (http://landweb.nascom.nasa.gov/cgi-bin/QA_WWW/newPage.cgi?fileName=MODLAND_C005_changes). The outlier detector is based on a simple histogram analysis to identify and remove pixels which show unusually low LST values, typically caused by cloud contamination (Neteler, 2005; Rizzoli et al., 2007). Individual MODIS images which contain areas devoid of data due to clouds or snow and ice were closed by means of geostatistics, taking into account the complexity of the terrain of the study sites and the temperature gradient in the area. This is a particular computational intensive operation which was performed on a HPC cluster.

2.3 Data analysis: creation of ecological indicators

Based on the available data, a series of ecological indicators was created. The overall processing chain was implemented in Unix (Linux) shell scripts in order to download the MODIS data from NASA, to reproject them from Sinusoidal to the Italian national coordinate system, to import the maps into GRASS GIS (V6.3.0, http://grass.osgeo.org/) and to apply quality control and outlier detection. GRASS itself uses the PROJ4 library (V4.5.0, http://trac.osgeo.org/proj/) for projection support as well as the GDAL/OGR libraries (V1.4.2, http://www.gdal.org) for interoperability.

2.3.1 Land Surface Temperature (LST) derived indices

The aim of the indices proposed here is the creation of short period indicators in order to capture temporal dynamics unlike other studies where multi-annual aggregation is performed (Scharlemann et al., 2008). A number of indicators could be generated in GRASS GIS using time series analysis (aggregating daily LST maps with *r.series* module, thresholding and filtering with *r.mapcalc*):

- annual, monthly and decadal temperature minima/maxima: for example ticks are active when the temperature is >= 10°C;
- late frost periods: relevant for masting of trees and seed production which influences rodent reproduction;
- growing degree days (GDD) for phenological status of vegetation;
- hot/cold summers identification through comparison of inter-annual mean temperature differences;
- autumnal temperature decrease (“autumnal cooling”, Randolph et al., 2000), and the opposite spring temperature increase (“spring warming”) which influence the life cycle of ticks; the indicator is the de-/increase as obtained by linear regression relative to the annual maximum of the monthly mean LST in midsummer (northern hemisphere);
- aggregated daily minimum/maximum temperature and diurnal temperature range (DTR);
- mean winter temperatures which are relevant for the ticks and mosquito egg survival.

2.3.2 Enhanced Vegetation Index (EVI) derived indices

We considered only EVI in our study as it tends to perform better than NDVI. It is less prone to saturation as well as less sensitive to haze due to the inclusion of the blue channel (Huete et al., 2002). The latter is of interest in mountainous regions where valleys are often relatively hazy. A series of indicators could be derived from the 16 day composition maps:

- calculation of seasonal differences across years by pixel-wise map subtraction;
- length of vegetation growing period through EVI thresholding;
- spring and autumn “detection" by finding relevant raise or decline of EVI: over short distances the seasons can be slightly shifted due to effect of valley orientation and exposition.

2.3.3 Maximum snow extent map

Snow extent and duration has a variety of ecological implications. It determines behaviour and winter survival of rodents, as well as survival of ticks and eggs of mosquitoes. The detection of snow for the maximum snow extent map product is based on various algorithms including the Normalized Difference Snow Index (NDSI) with a seasonally varying NDSI threshold for snow presence (Hall et al., 2002). Comparison to the distributed hydrological model Geotop (http://www.geotop.org/) showed that the 8 days maximum snow extent maps are a valid substitute to the more data intensive and complex Geotop snow modelling (Endrizzi et al., 2006). As ecological indicators, besides snow cover duration and inter-annual differences also the last day of snow for each pixel in spring time as well as the first day with snow in autumn was calculated. This required simple map algebra with r.mapcalc and r.series in GRASS.

2.3.4 Humidity and moisture indices from surface reflectance

From daily MODIS surface reflectance maps two humidity and moisture indices can be generated: the Normalized Difference Water Index (NDWI) and the Land Surface Water Index (LSWI). These indices have, however, limited significance as they are based on surface observations which include bare soil, rocks and vegetated area, often even as mixed pixels. Again, they are generated with simple map algebra.

2.4 Free GIS software integration with machine learning methods

The further use of the generated ecological indicators in epidemiological applications requires an extension of the (geo)statistical capabilities of GRASS. A working option is the GRASS interface to the R Language and Environment for Statistical Computing (Bivand & Neteler, 2000). The R session is launched within the GRASS session which makes all maps in the GRASS database accessible to R. Furthermore, it can be connected to RDBMS like PostgreSQL for the non-spatial host, vector and meteorological data.

R was used to prepare multi-temporal predictive maps by analysing geodata from GRASS, and biological data samples (host and vector observations) and meteorological data from PostgreSQL. A first risk mapping system for tick-borne diseases, applied to model the risk of exposure to Lyme borreliosis and tick-borne encephalitis (TBE) in Trentino, Italian Alps was presented in Furlanello et al. (2003).

An alternative toolbox is Machine Learning Py (mlpy, http://mlpy.fbk.eu) which is a new high-performance Python/NumPy based package for machine learning. It yet lacks integration with GRASS GIS which, however, could be obtained through the GRASS-SWIG interface (http://www.swig.org).

3. Results

In complex terrain such as the study area, meteorological stations and ground surveys are usually sparsely distributed. Applied to areas with complex terrain, traditional geospatial interpolation methods are often error prone and difficult to optimize since the meteorological stations are typically positioned only in valleys or in the accessible parts of the mountain ranges. The integration of satellite data into epidemiological research enhances the spatio-temporal resolution significantly (Hess et al., 2002).

Results of this analysis are related to three areas: data availability, quality of ecological indicators and quality of the software environment. The increased availability of daily satellite data is the key to new ecological indicators. Indicators based on high temporal resolution are crucial for the analysis and the understanding of disease patterns which can now be calculated at global scale (with exception for cloud dominated areas). This will overcome the lack of environmental data in areas with limited availability of sensors.

The quality of ecological indicators, especially in complex terrain is determined by thorough pre- and postprocessing of the data to minimise artefacts in the maps. While vegetation indices and snow cover extent maps are hard to compare with ground truth data, this can be done with land surface temperature. LST data are not directly comparable to air temperatures measured at 2m above ground but it is assumed that the general temperature profile/pattern be very similar. A comparison of LST derived indices to meteorological data in the Italian Alps with both data sets been aggregated to 10 days periods showed good agreement. A Wilcoxon rank sum test performed on the two curves confirmed that they statistically do not differ (W=679, p-value=0.9572, data not shown here). Similar results were gained from comparing maximum and minimum temperatures (see also Neteler, 2005).

Essential for the data processing is the quality of software environment. Besides the ease of use, expandability to special needs as required in this study and performance, also the availability of peer reviewed, public algorithms is an asset. The latter is ensured by the Open Source development model. The work flow to generate ecological indicators was completely based on Free Software. Standard software versions of PROJ4, GDAL/OGR, GRASS GIS, PostgreSQL and R were used and combined with Unix shell scripting to obtain an automated processing chain. The environment worked stable over many days of data processing. It was also possible without major modifications, to process LST maps in parallel for void filling with the GRASS GIS software on a High Performance Computing cluster which reduced the calculations from estimated 8 month on a single CPU to one month on up to 60 CPUs.

4. Conclusions

The results demonstrate that a new set of ecological indicators is becoming available now. They are of interest mainly for landscape epidemiological risk modelling but also for other related research areas. Data availability and their high time/space resolution dramatically improved in recent years, in particular for data originating from polar orbiting satellites. In combination with a robust Free Software environment and enhancements in data processing and machine learning algorithms, epidemiological risk modelling will gain significant improvements in the near future.

The research activities presented in this work will be continued in the EDEN (*Emerging Diseases in a changing European eNvironment*) EU/FP6 project framework (http://www.eden-fp6project.net).

Acknowledgements

NASA LP DAAC is acknowledged for making the MODIS data available. P. Larsson, FOI, Umeå (Sweden) and the HPC2N team are acknowledged for arranging access to the Sarek Opteron cluster at the HPC2N of Umeå University. This research was partially funded by EU grant GOCE-2003-010284 EDEN. I thank D. Gianelle who provided helpful comments on an earlier version of the manuscript.

References

Beck, LR, Lobitz, BM, & Wood, BL, 2000, 'Remote sensing and human health: New sensors and new opportunities', *Emerging Infectious Diseases*, vol. 6, no. 3, pp. 217-227.

Benito Garzòn, M, Blazek, R, Neteler, M, Sánchez de Dios, R, Sainz Ollero, H, & Furlanello, C, 2006, 'Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula', *Ecological Modelling*, vol. 197, no. 3-4, pp. 383-393.

Bivand, R & Neteler, M, 2000, 'Open Source geocomputation: using the R data analysis language integrated with GRASS GIS and PostgreSQL data base systems', *Proceedings 5th conference on GeoComputation*, 23-25 August 2000, University of Greenwich, U.K.

Carpi, G, Cagnacci, F, Neteler, M, & Rizzoli, A, 2007, 'Tick infestation on roe deer in relation to geographic and remotely sensed climatic variables in a tick-borne encephalitis endemic area', *Epidemiology and Infection*, pp. 1-9.

Endrizzi, S, Bertoldi, G, Neteler M & Rigon R, 2006, 'Snow cover patterns and evolution at basin scale: GEOtop model simulations and remote sensing observations', *Proceedings of the 63rd Eastern Snow Conference*, Newark, Delaware (USA), pp. 195-209.

Furlanello, C, Neteler, M, Merler, S, Menegon, S, Fontanari, S, Donini, A, Rizzoli, A & Chemini, C, 2003, 'GIS and the randomForest Predictor: integration in R for tick-borne disease risk assessment', *Proceedings Distributed Statistical Computing*, 20-22 March 2003, Vienna, Austria.

Hall, DK, Riggs, GA, Salomonson, VV, Digirolamo, NE, & Bayr, KJ, 2002, 'MODIS snow-cover products', *Remote Sensing of* Environment, vol. 83, no. 1, pp. 181-194.

Hay, SI, Randolph, SE, & Rogers, DJ, 2000, 'Remote Sensing and Geographical Information Systems in Epidemiology', *Advances in Parasitology 47*. Academic Press.

Hess, G, Randolph, SE, Arneberg, P, Chemini, C, Furlanello, C, Harwood, J, Roberts, M, & Swinton, J, 2002, 'Spatial aspects of disease dynamics', in: PJ Hudson, A Rizzoli, BT Grenfell, H Heesterbeek & AP Dobson, *The Ecology of Wildlife Diseases*, Oxford University Press, pp. 102–118.

Herbreteau, V, Salem, G, Souris, M, Hugot, JP, & Gonzalez, JP, 2005, 'Sizing up human health through remote sensing: uses and misuses', *Parassitologia*, vol. 47, no. 1, pp. 63-79.

Huete, A, Didan, K, Miura, T, Rodriguez, EP, Gao, X, & Ferreira, LG, 2002, 'Overview of the radiometric and biophysical performance of the MODIS vegetation indices', *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 195-213.

Kruse, H, Kirkemo, AM, & Handeland, K, 2004, 'Wildlife as source of zoonotic infections', *Emerging Infectious Diseases*, vol. 10, no. 12, pp. 2067-2072.

Kuhn, KG, Campbell-Lendrum, DH, and Davies, CR, 2004, 'Tropical diseases in Europe? How we can learn from the past to predict the future', *EpiNorth Journal*, 1.

Neteler, M, 2005, 'Time series processing of MODIS satellite data for landscape epidemiological applications', *International Journal of Geoinformatics*, vol. 1, no. 1, pp. 133-138.

Neteler, M & Mitasova, H, 2008, *Open Source GIS: A GRASS GIS Approach*. 3rd edition, Springer, New York, http://www.grassbook.org/

Randolph, SE, 2000, 'Ticks and tick-borne disease systems in space and from space', in: SI Hay, SE Randolph & DJ Rogers, 2000, 'Remote Sensing and Geographical Information Systems in Epidemiology', *Advances in Parasitology 47*. Academic Press, pp. 217-243.

Randolph, SE, 2006, 'EDEN - Emerging diseases in a changing European environment: tick-borne diseases', *International Journal of Medical Microbiology*, vol. 296, suppl. 1, pp. 84-86.

Rizzoli, A, Neteler, M, Rosà, R, Versini, W, Cristofolini, A, Bregoli, M, Buckley, A, and Gould, EA, 2007, 'Early detection of TBEv spatial distribution and activity in the province of Trento assessed using serological and remotely-sensed climatic data', *Geospatial Health*, vol. 1, no. 2, pp. 169-176.

Scharlemann, JP, Benz, D, Hay, SI, Purse, BV, Tatem, AJ, Wint, GR, & Rogers, DJ, 2008, 'Global Data for Ecology and Epidemiology: A Novel Algorithm for Temporal Fourier Processing MODIS Data', *PLoS ONE*, vol. 3.

Sumilo, D, Bormane, A, Asokliene, L, Vasilenko, V, Golovljova, I, Avsic-Zupanc, T, Hubalek, Z & Randolph, SE, 2008, 'Socio-economic factors in the differential upsurge of tick-borne encephalitis in central and eastern Europe', *Reviews in Medical Virology*, vol. 18, no. 2, pp. 81-95.

Tatem, AJ, Goetz, SJ, and Hay, SI, 2004, 'Terra and Aqua: new data for epidemiology and public health'. *International Journal of Applied Earth Observation and Geoinformation*, vol. 6, no. 1, pp. 33-46.