DSC 2003 Working Papers
(Draft Versions)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

GIS and the Random Forest Predictor: Integration in R for Tick-Borne Disease Risk Assessment*

C. Furlanello†, M. Neteler†, S. Merler†,S. Menegon†, S. Fontanari†, A. Donini‡, A. Rizzoli‡, C. Chemini‡
†*ITC-irst* and ‡*CEA, Trento, Italy*

Abstract

We discuss how sophisticated machine learning methods may be rapidly integrated within a GIS for the development of new approaches in landscape epidemiology. A multitemporal predictive map is obtained by modeling in `R`, analyzing geodata and digital maps in `GRASS`, and managing biodata samples and weather data in `PostgreSQL`. In particular, we present a 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.

1 Background

GIS, Machine Learning and Landscape Epidemiology

In a landscape epidemiology problem, data are collected on sampling sites and then generalized using a Geographical Information System (GIS): we can think of the procedure in terms of a machine learning problem, in which from few hundreds of examples a classification or regression function (e.g. presence or absence of an infective agent, or density of infection vectors) is estimated for an entire territory of thousands or of millions of cells, from several input variables, each defined through a digital map. The variables may be numerical (e.g. altitude from a digital elevation model) or categorical (e.g. vegetation class coverages). Multitemporal high resolution remote sensing sources are in particular now available that may allow to model time-varying output maps.

Scripts or Dedicated Programs?

In previous studies, we demonstrated that machine learning models as single or aggregated classification trees may be effectively used to develop GIS digital maps

*This study was partially funded by the ECODIS Project of the FUR-PAT

of the probability distribution of infected nymphs of the vector *Ixodes ricinus (L.)* in the Province of Trento at high resolution scale [11, 7, 14]. Our 1996 model was obtained through a script interface between the statistical computing environment `S-Plus` and the GIS map building and visualization tools of the GIS `GRASS` [13]. The procedure was rather laborious: a table of site data descriptions and of tick sampling data was imported in `S-Plus`, a sequence of tree-based classification models was constructed and a model selected according to a particular bootstrap method, the .632+ rule [6, 12]. A customized version of the corresponding `S-Plus tree` object was then exported, parsed by a Perl script, and fed into the `r.mapcalc` interpreter for map algebra in GRASS. Although this approach gave better results than a standard linear discriminant model, automation was limited. More complications were then needed in order to use more accurate models, as with multiple classifier combination (or *ensemble learning*) approaches as bagging [4] or boosting [15], which were thus implemented as stand-alone procedure in `GRASS`.

The Integration of R and GRASS

The progresses of the R-GRASS interface [2, 1, 13] provide an important simplification of the risk modeling phase, or at least of the development of the prototype model. In this paper we show how easily we were able to integrate into GRASS the recent randomForest [5] ensemble prediction method. The randomForest technique was recently ported as an `R` package [10], and it is thus now available also for GIS analysis without requiring a direct implementation in a GIS system, once the needed data are sourced into `R` objects. The randomForest method uses a classification tree as a base model, thus allowing a mix of numerical and categorical input variables, a typical situation with GIS models. Useful subproducts of randomForest computation, as the variable importance plots are also available.

A System for Environmental Risk-mapping

As it will be detailed below, all the critical data in this problem are endowed with geographical coordinates: they are maintained in a GIS location or in the project database. The working system is actually a mixed environment of GIS and database management systems tools and data structures. We may especially take advantage of the date/time structures available in `PostgreSQL`, and of all the data conversion libraries which allow high integration of `GRASS` and `PostgreSQL`. Thus the modeling phase requires to extract examples of associations between predictor and target variables from this mixed environment. Connecting `R` to `PostgreSQL` and to `GRASS` at the same time, as well as connecting `PostgreSQL` and `GRASS` together as needed, makes `R` a very productive working environment. Both interactive (e.g. the ESS Emacs mode) and batch modes are available for variable preprocessing, analysis, model development and selection, and finally for map production. It is important to note that we may use time varying data inputs as meteorological data. The climatic data may be directly obtained as raster maps from remote sensing imagery. Otherwise the maps needs to be computed by spatial interpolation of climatic time series from stations. Above all, in order to extend to all the territory the prediction modeled from the data collected on sampling sites, it makes sense to use for model development only variables which are effectively available as maps, dropping detailed descriptions available only for the sampling sites or for a fraction of the target territory.

The map is returned to GRASS or to the WebGIS component, a solution for geodata visualization on Internet based on the MapServer Open Source software, which we have extended for epidemiologic data management and used as a notification system in order to add new data with geolocation [8].

As far as the application is concerned, the R-GRASS interface allowed the rapid development of a mesoscale risk map of tick presence at 100×100 m^2 pixel resolution. Note that a openMosix cluster was used to produce several of the predictor variables involved in the study. The project results are distributed through the ECODIS server http://mpa.itc.it/ecodis/.

We will describe in the rest of the paper several technical details of the system, illustrating a prototype model recently developed also with the introduction of multitemporal climatic data.

2 Methods

Connections among PostgreSQL, GRASS 5.1 and R

The recent developments of GRASS 5.1 provide a much stronger linkage between the GIS and RDBMS such as PostgreSQL. Besides attributes storage in the database system, also geometry data can be read and written to PostgreSQL extended by PostGIS. Along with the integration of R into GRASS as well as the DBI interfaces for R circular connections for quick data retrieval and processing is possible. In interactive sessions, R has to be run from within the GRASS shell environment ([3],[1]). The interface is dynamically loading several compiled GIS library functions into the R executable environment. Several crucial information being transferred to R are the GRASS metadata defining the regional extent and the raster resolution of the area of study, known as the GRASS LOCATION. The interface is currently supporting raster and site data. Since March 2003 the R-GRASS interface is placed in the standard contribution section of CRAN.

Field data characterization

To determine the distribution and relative abundance of *I. ricinus*, questing ticks were collected in 434 sites of the Province of Trento (see Fig. 2) from 10 March to 15 October 1996, by dragging vegetation using a standardized procedure. More biological and technical details on data are available in [14]. Data on distribution and abundance of questing ticks collected in each site are available as a database connected to the GRASS. This connection allows characterizing the sampling sites in terms of the environmental variables correlated to the tick presence.

In particular, the following environmental variables were considered in this study:

- altitude: a general, long period indicator of the climate, e.g. of the average annual temperature;
- geological substratum: related to soil humidity and main vegetation types;
- roe deer density: the main tick host within large mammals usually infested by *I. ricinus*;

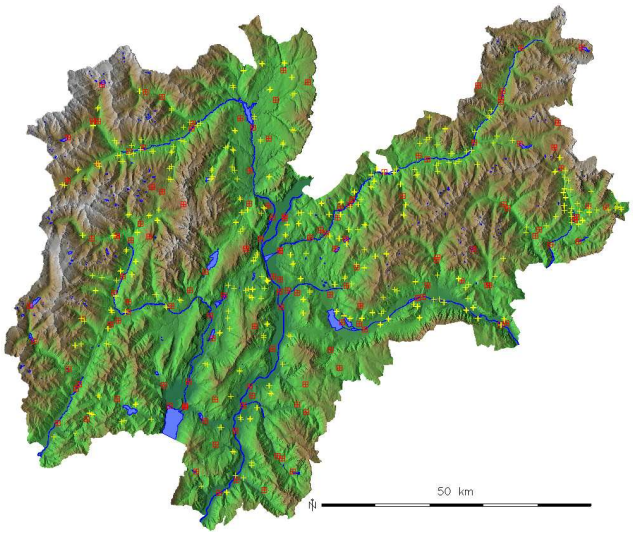

Figure 1: Map showing the elevation, the climatic stations from the Province of Trento (red boxes) and the ticks sampling sites of the 1996 campaign (yellow crosses)

- vegetation: related to the main habitat of other tick hosts, e.g. small mammals;
- winter precipitation: potentially correlated to the abundance of larvae survived with respect to the previous year;
- min temperature: the average of the minimal temperatures of the 30 days before sampling is a short period indicator of the climate;
- convexity: the local profile convexity measures, computed from the digital elevation model are potential descriptors of local microclimatic conditions.

Sampling and Climatic data

Climatic data were available for ten years (1990-1999) for 142 meteorological stations of the Province of Trento, Italy (see Fig. 2 for the distribution of sampling sites and climatic stations). The climatic data as well as the ticks sampling data are kept in `PostgreSQL` tables. Three time series of temperatures (tmean, tmax, tmin for 1995–1997) were extracted with SQL queries from station tables and used to create temperature maps for the whole area (about 6250 km^2) for 63 stations valid along the whole period. An approach adapted from [9] was used, based on 3D Regularized Splines with Tension (RST) interpolation. As data were not interpolated during the winter time, thus no temperature inversion was to be expected in the bottoms of the valleys. The interpolated temperature maps were used to pick daily temperature for the sampling sites, and those values were returned into SQL tables for later use. A similar procedure was performed for accumulated precipitation for the two winters before the sampling year (1994 and 1995), for the time intervals [1 Nov – 28 Feb]. Also, direct potential radiation maps (with shadow) at days 21

March, 21 June, 23 September, 21 December, were computed in GRASS, evaluated for the sampling sites, and organized in SQL tables.

The model

The target of this study was the use of the randomForest predictor and developing a spatial model of the probability of tick presence, given environmental biotic and abiotic input variables. The environmental data described above, being defined as maps along most of the area of study, allowed to create tables in the DBMS of all the variables of potential interest. A project data frame was thus developed, with absence/presence of ticks as the target variable (presence: 224 sites, absence: 210 sites), and altitude (alt.dem.10m), plan convexity (plancs11.30m) and profile convexity (profcs11.30m), main geological substratum (genesis), roe deer density (densita.caprioli), vegetation (vegetazione: 11 classes), accumulated precipitation (prec.winter.95: 108 – 351 mm), mean of min temperature in the 30 days before the sampling (min.temp.3).

In the random forests algorithm, prediction is obtained by aggregating classification or regression trees each constructed using a different random sample of the data (as in bagging), and choosing splits of the trees from subsets of the available predictors, randomly chosen at each node [5, 10]. The randomForest model in this study was obtained by aggregating 1000 trees as base classifiers, with 2 variables tried at each split. Given an input pattern, the model outputs the probability of tick presence: by using the connection between R and GRASS, the model was applied to each cell (100×100 m^2) of the entire study area, for a total of 545015 outputs (mostly waters, urban areas and intensive agriculture areas were omitted from computation).

The main result of this procedure is thus a map of the probability of tick presence covering all Province of Trento. As the min temperature variable introduces a time-varying effect, the model allows simulating how the risk map changes with respect to the short horizon temperature.

The main components of the software environment were `R` (1.6.2, with randomForest, lattice, DateTime classes, Rdbi), `GRASS` (5.0.0, 5.1.0), `R/GRASS` interface (0.2-6), `PostgreSQL` (7.2). Due to the multitemporal structure of the data the `R` *DateTime* classes were heavily used, in particular for treating climatic data from stations.

3 Results

In this study we use the "out-of-bag" (OOB) estimate of the error rate of the model. At each bootstrap iteration, a tree is grown on data extracted on the bootstrap sample, and apply for prediction on data not in the bootstrap sample. The error rate calculated by aggregation of the OOB prediction is the OOB estimate of the error rate. The out-of-bag model error estimate for the randomForest model of 1000 trees is 28.6%. More specifically, the error on the tick presence sites is 27.2%, and 32.9% on the absence sites.

The random forest algorithm also produces extra information that allow to evaluate the importance of each explanatory variable: following [5], the implementation of random forests in `R` at the time of preparation of this study provided four "variable importance" measures. There are two main types of importance methods,

based respectively on label permutation and on impurity decrease. In permutation methods, in order to estimate the importance of the m-th variable, prediction on the test cases is computed after that all values of the variable are permuted: the amount of variation of the error rate and of the classifier margin (the proportion of votes for the true class minus the maximum of the proportion of votes for the other classes) are then considered. Alternatively, we can consider the accumulated reduction at nodes according to the criteria used at the splits, an idea from the original CART formulation. In randomForest, the splitting criterion is the Gini index, thus the sum of all decreases in the forest due to a given variable, normalized by the number of trees, is used to define the Gini variable importance measure. It is clear that this measure may reveal variables which can cause many small decreases summing up to a large contribute to model deviance reduction. For our model with 1000 trees, in Fig. 2 we report the 4 measures of variable importance (1: error increase; 2: average margin increase; 3: differential of margin increases; 4: Gini decrease). Measure 2 and 4 were found the more stable. With respect to measure 4, the most important variable is the accumulated winter precipitation before the year of sampling, followed by altitude, profile convexity, min temperature and geological substratum. For both measures, it seems that the climatic variables, in the short and long period, are the key factors to predict the tick presence. This is a novel result for Trentino. With respect to previous published models, the geological substratum is confirmed to play a key role for identifying the favorable tick habitats. Figures 3, 4 and 5 are partial dependence plots of probability of tick presence, as predicted by the model, as functions of the variables altitude, winter precipitation and min temperature respectively. As found by single tree and bagging models, the probability of tick presence drastically decreases as the altitude becomes > 1000 m. This is not a surprising results, since it is well recognized that the temperature plays a key role in the tick life cycle. Interestingly the short period climatic indicator, the min temperature, is highly effective. Moreover, the winter precipitation is highly correlated with the probability of tick presence, possibly by influencing the quantity of larvae surviving with respect to the previous year.

The three maps in Fig. 6 are our first attempt to define a multitemporal (actually: temperature-dependent) tick risk map. The risk maps are obtained by assuming a min temperature of 0° C (top), 6° C (middle) and 13° C (bottom) for each cell, respectively. We can thus imagine to effectively use the model to estimate the risk of a being bitten by a questing tick given the (known, as computed in GIS from the meteorological station data, which are available on-line) mean of the min temperature in the previous month.

4 Discussion

This model must be regarded as preliminary, as we need to provide, on the basis of biological hypothesis being collected from experts, a complete set of potentially relevant time-varying factors. In addition, a full methodological set-up needs to be organized in an effective model selection protocol. However, this is the first time the effect of accumulated winter precipitation and temperature has been demonstrated: this randomForest prototype model was constructed in a remarkable short time, and the connections between `R`, `PostgreSQL` and GRASS have been instrumental to obtain this result. Linux systems were used for the computations described in this paper, but at the present time the whole environment described in this paper

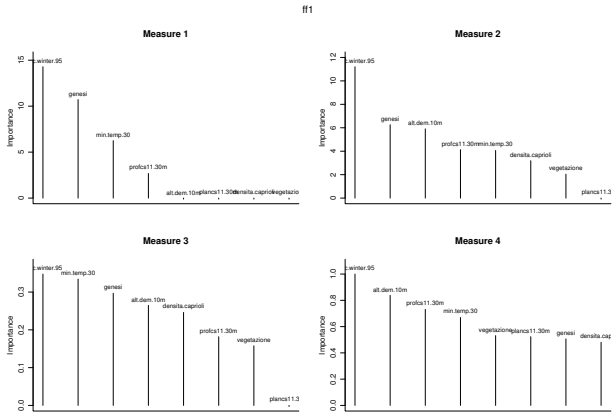

Figure 2: Four measures of variable importance for tick presence/absence model as measured by the randomForest predictor

is available also under MS-Windows with the support of the Cygwin tools.

Acknowledgments

We are grateful to Roberto Flor and Alessandro Soraruf for their assistance in developing the MPA Linux Cluster. The authors thank Stéphane Dray for valuable comments on this paper.

References

[1] R. Bivand and M. Neteler. Open source geocomputation: using the R data analysis language integrated with GRASS GIS and PostgreSQL data base systems. In *Proc. 5th conference on GeoComputation (CDROM), 23-25 August 2000, University of Greenwich, U.K.*, `http://reclus.nhh.no/gc00/gc009.htm`, 2000.

[2] R. S. Bivand. Integrating GRASS 5.0 and R: GIS and modern statistics for data analysis. In *Proc. 7th Scandinavian Research Conference on Geographical Information Science, Aalborg, Denmark*, pages 111–127, 1999.

[3] R. S. Bivand. Using the R statistical data analysis language on GRASS 5.0 GIS data base files. *Computers & Geosciences*, 26(9/10):1043–1052, 2000.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: The .632 + bootstrap method. *JASA*, 1997.

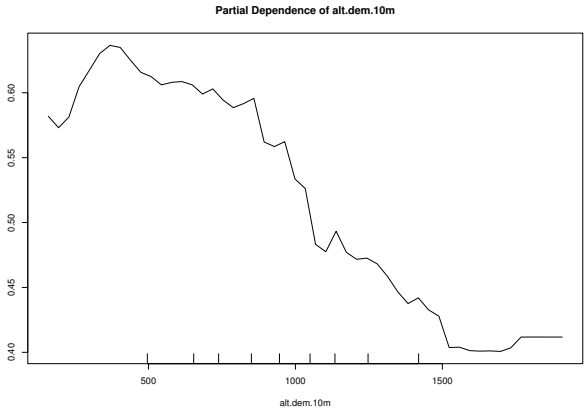

Figure 3: Partial dependence plot showing the effect of the DTM variable on the classification

[7] C. Furlanello and S. Merler. Boosting of tree-based classifiers for predictive risk modeling in GIS. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems, Lecture Notes in Computer Science 1857*, pages 220–229. Springer, 2000.

[8] C. Furlanello, S. Merler, S. Menegon, S. Mancuso, and G. Bertiato. New WEBGIS technologies for geolocation of epidemiological data: an application for the surveillance of the risk of Lyme borrelliosis disease. *GIAC*, 5(1):241–245, 2002.

[9] J. Hofierka, J. Parajka, M. Mitasova, and L. Mitas. Multivariate interpolation of precipitation using regularized spline with tension. *Transactions in GIS*, 6:135–150, 2002.

[10] A. Liaw and M. Wiener. Classification and regression by randomForest. *Rnews*, 2/3:18–22, December 2002.

[11] S. Merler, C. Furlanello C. Chemini, and G. Nicolini. Classification tree methods for analysis of mesoscale distribution of ixodex ricinus (Acari: Ixodidae) in Trentino, Italian Alps. *Journal of Medical Entomology*, 33(6):888–893, 1996.

[12] S. Merler and C. Furlanello. Selection of tree-based classifiers with the bootstrap 632+ rule. *Biometrical Journal*, 39(2):1–14, 1997.

[13] M. Neteler and H. Mitasova. *Open Source GIS: A GRASS GIS Approach*. The Kluwer international series in Engineering and Computer Science (SECS): Volume 689. Kluwer Academic Publishers, Boston, Dordrecht, London, 2002.

[14] A. Rizzoli, S. Merler, C. Furlanello, and C. Genchi. Geographical Information System and Bootstrap Aggregation (Bagging) of tree-based classifiers for Lyme disease risk assessment in Trentino, Italian alps. *Journal of Medical Entomology*, 39(3):485–492, 2002.

[15] R. E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

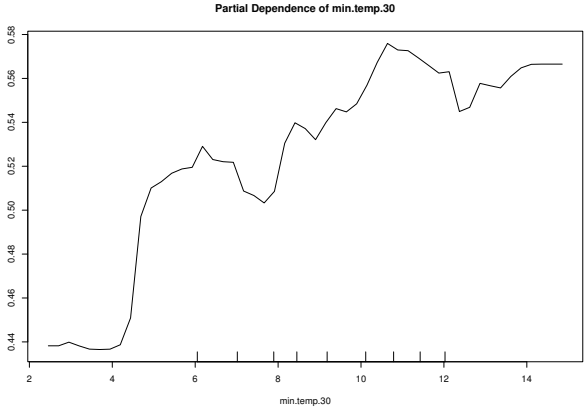

Figure 4: Partial dependence plot showing the effect of the minimum temperature variable on the classification

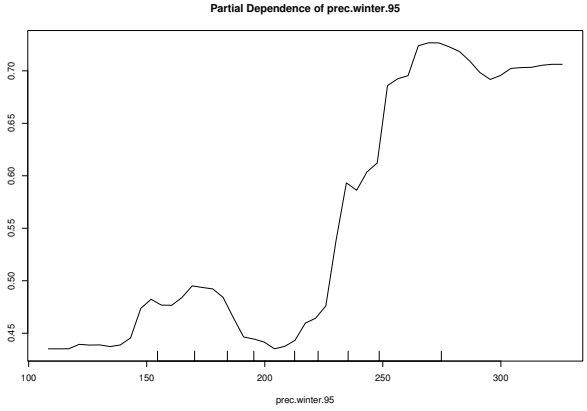

Figure 5: Partial dependence plot showing the effect of the precipitation variable on the classification

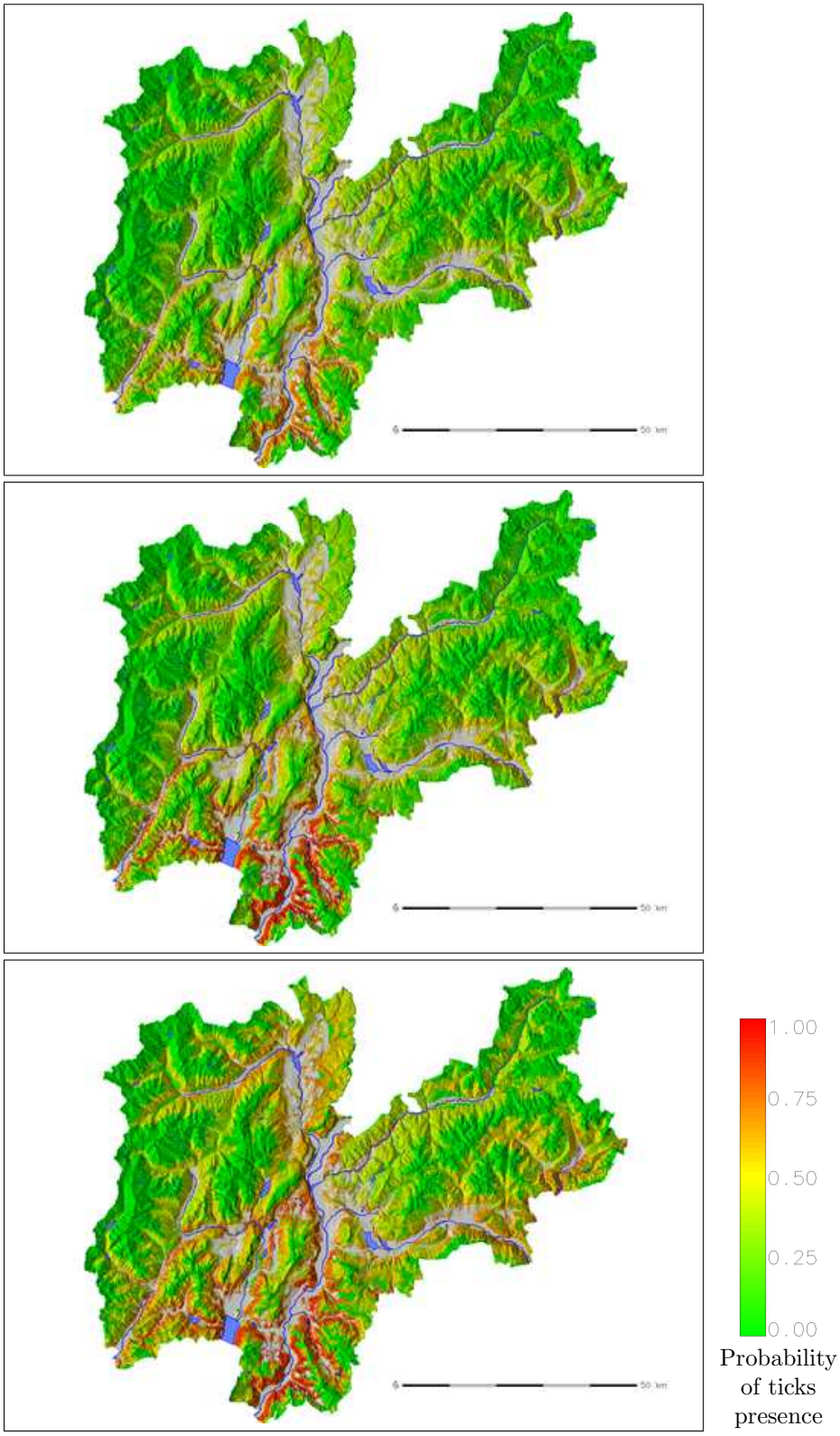

Figure 6: Risk maps for exposure to Lyme borreliosis and TBE in Trentino, Italian Alps. Ticks presence probability assuming a min temperature of 0° C (top), 6° C (middle), 13° C (bottom), in the previous month