Exploring the relationships of fire occurrence variables by means of CART and MARS models

Giuseppe Amatulli, Andrea Camia [1]

Abstract

Recently, in the framework of long-term fire risk assessment, researcher have implemented spatial and non-spatial non-parametric prediction models to discover complex relationships among wildfire variables. The main scope was to overcome the assumption of spatial stationarity in the relationship among the response variable and the predictors, assumed by the traditional regression techniques. The present article aims to test and compare the potential of the CART and MARS models in predicting fire occurrence at local scale. The test is performed in the Arno River Basin, a fire prone area located in the central part of Italy. Road network, topographic variables and population data were implemented to build up fire prediction model using 1621 ignition points recorded during the period 1997-2003. The models produce two prediction maps slightly similar. In general the CART model over-perform compare to the MARS one. Nonetheless, the MARS model produces a smoothened surface that theoretically better follow the probability of a fire event.

Introduction

Wildfires are the main cause of land degradation in Mediterranean countries. Analysis of the geographic distribution of the fire events has always been an important issue in fire pattern modelling, and now is going to receive more attention for landscape modelling and for fire management and prevention actions. In particular, the prediction of wildfires in terms of fire occurrence or fire frequency for long-term assessment opens an important field of research. Besides, the availability of spatial and flexible regression models combined with the potential of fast computation reached in the last decade provides a practical help to complex data analysis.

The analytical framework of fire risk assessment is generally based on multiple regressions where the fire occurrence response variable is related to a suite of environmental and human predictor variables. Recently, in this context, scientists have used spatial (Koutsias et al., 2005) and non-spatial (McKenzie et al., 2000; Amatulli et al., 2006) non-parametric prediction models to discover complex relationships among wildfire variables. The main aim was to overcome the assumption of spatial stationarity in the relationship among the response variable and the predictors, assumed by the traditional regression techniques. In fact, the regression techniques commonly used identify coefficients implemented as weights for considered predictor variables to estimate the response variable, in this case the fire occurrence. Therefore, scientists are testing models able, not only to consider the several relationships among the variables, but also to reveal non-additive behaviour

[1] European Commission - DG Joint Research Centre,
Institute for Environment and Sustainability
T.P. 261, Via E. Fermi, 21020 Ispra (VA), Italy
Giuseppe.Amatulli@jrc.it , Andrea.Camia@jrc.it

of variables and yielding understandable output readily available to the final users (Amatulli et al., 2006).

In particular, Koutsias, et al. (2005) explore the potential of Geographic Weighted Regression (GWR) (Fotheringham et al., 2002). GWR relies on the theory that the relationships between the response variable and the predictor variables are not constant over the space, as a result the technique allows the estimation of parameters that change over the space producing a continuous surface across the study region (Fotheringham et al., 2002; Koutsias et al., 2005). Nevertheless, GWR does not permit the use of discrete variables, and the use of dummy variables still remains a limitation in practical application. Besides, Amatulli et al. (2006) and McKenzie et al. (2000) highlighted the potential of Classification and Regression Tree (CART) (Breiman et al., 1984) technique to identify and express in a relatively simple form non-linear and non-additive relationships among wildfire occurrence variables. Nonetheless, the authors have highlighted two main aspects that have an impact on the use of the results and the model performance. Firstly, the regression rules created by the CART analysis produce fire occurrence map in the form of zone map that represent the fire probability as constant phenomena for the so called fire management unit (Amatulli et al., 2006). Such aspect can be considered positive for a nation/regional fire management planning where local variation are not considered relevant and management decisions are taken for wide areas. On the contrary, for local planning prediction is required to represent the fire phenomena as continuous and smoothest surface in order to detect local variation of fire probability. Secondly, a potential problem that might arise in the use of the CART analysis is that the decision rules do not take into account the values of the neighbouring cells, in other words, they do not consider their spatial relationships while the spatial pattern and dependency do exist in biogeographical data.

Hence, this work tries to overcome the mentioned limitations testing and comparing the Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) technique against the CART analysis, with the help of ancillary layers. The MARS model is more indicate for regression analysis due to its capability in predicting unique value for each observation. The models will be applied to a study case using common fire occurrence predictors such as road network, population density, land cover and topographic variables, in the contests of Arno River Basin. The final aim is to discover advantage and drawbacks of the two models in producing a reliable long-term fire risk map.

Study area

The Arno River Basin is located in the central part of Italy, in the Tuscany Region, and it cover 9269 km^2 (Figure 1). The Arno River Basin present several land cover classes that go from urban/industrial to agriculture/forest classes. The main areas prone to wildfire were selected from the CORINE land cover by merging wildland classes, as result an area of 4225 km^2 was considered for the whole fire risk analysis. The vegetation is a mix of different patches ranging from dense forest of *Quercus spp.* to scrubby Mediterranean maquis, combined with presence Mediterranean pinus. The Arno River Basin is affected by wildfire principally due to anthropic reasons caused by negligence and/or voluntary actions.

Metodology and dataset

For the implementation of the study case, the long-term fire risk assessment is intended as the “predicted fire occurrence” assessed through a regression of

explanatory variables (by means of MARS and CART models) against the “observed fire occurrence”. Often, to have the real perception of the fire phenomena the “fire occurrence” is expressed in terms of number of fire per unit area, in this context this will be called density of fire events. Commonly, the fire risk predictor models enforce the identification of regression rules or regression coefficients based on a response variable against several predictors. Herein, in the following chapter the involved variables and the related methodology to derive them are explained and described.

Dataset

Response variable

For this study case the response variable was based on the fire occurrence recorded in the Italian national fire database during the period 1997–2003. The National Forest Service organizes the activity of the Nucleo Antincendio Boschivi (AIB) units, which are directly involved in wildfire prevention and suppression. Details about each single fire come from the AIB forms, and are collected and processed to compile the national fire database. To verify whether the fire ignition points coordinates correspond to the place where the event really occurred the coordinates are cross validated by comparing them with ancillary data contained in the database (e.g., the municipality, locality, toponym). A preliminary spatial representation of the ignition points was performed to assess their distribution, to allow the spatial query of the points falling in Arno Basin River, and to highlight possible coordinate mistakes. Consequently, a ‘‘reliability index’’ (RI) (Amatulli et al., 2005) was attributed to each ignition point using ArcView 3.2 software (ESRI, 1997). This code was attributed to each fire event classifying it according to three different categories: records with no geographic coordinates and/or no municipality information data; records with geographic coordinates correctly located; and records with geographic coordinates incorrectly located. In the case of records that were not correctly located, the municipality or other information coming from the complete fire records of events happening in the same area, were used to assign coordinates to the fire event (Amatulli et al., 2005). As consequence, 1621 ignition points having a reliability of 99.81% were used to build a fire density map following the technique and the relative calibration procedure described in Amatulli, Perez-Cabello et al. (2007). For this study case the fire density was expressed as number of ignition points for square kilometre (ip/km^2). The fire occurrence was referred to the entire study period (7 years) to avoid long decimal numbers.

Predictor variables

The predictor variables were chosen based on their estimated ability to predict fire occurrence, and they are listed in Table 1. The predictor variables can be associated to two main groups: physical and human variables. Among the former, the choice was also driven by the availability of the geo-datasets in the study area. The road network was used to set up two predictors: road density and road distance, respectively for the primary and secondary road. These variables are often used to provide information on the accessibility to the forest for suppression action but also for ignition purpose (Leone et al., 2003). The aim to use both of variables was to test which one combined with the response variable gives information on fire occurrence pattern. The pixel resolution was of 100 m for the whole geo-dataset. Finally in order to give information concerning the spatial relationship, x and y coordinates were added as predictors. In order to have an idea about the importance of each variable in

discriminate the wildfire phenomena a Pearson correlation coefficient between each predictor and the response variable was calculated.

Table 1— Response and predictor variables used in the fire occurrence model prediction.

Variable type	Variable file-name	Explanation notes	Range		Units
			From	To	
Response	Dk[5]	Ignition point density – adaptive kernel density	0	3.74	ip km^2
Predictor	CORINE (Code)	Wild Land cover by CORINE (CLC90)	422532		n. of pixels
	21	Agriculture land with significant areas of natural vegetation	39084		n. of pixels
	22	Agro-forestry areas	32		n. of pixels
	23	Broad-leaved forest	260746		n. of pixels
	24	Coniferous forest	19934		n. of pixels
	25	Mixed forest	762663		n. of pixels
	26	Natural grasslands	1561		n. of pixels
	28	Sclerophyllous vegetation	1799		n. of pixels
	29	Transitional woodland-shrub	22606		n. of pixels
	32	Sparsely vegetated areas	424		n. of pixels
	33	Burnt areas	83		n. of pixels
Predictor	ALT	Altitude	0	1651	m
Predictor	POP	Population distribution by CORINE (CLC90)	0	59.64	Citizen/km^2
Predictor	ASP_SIN	Sine of aspect	-1	+1	-
Predictor	ASP_COS	Cosine of aspect	-1	+1	-
Predictor	SLOPE	Slope	0	46	Degree
Predictor	PARK	Mask with Protected areas limits	0	1	1=park; 0=outside
Predictor	R1DENS	Continuous grid density from secondary roads	0	1073	m/km^2
Predictor	R2DENS	Continuous grid density from primary roads	205	10199	m/km^2
Predictor	R1DIST	Continuous grid distance from secondary roads	0	23526	m
Predictor	R2DIST	Continuous grid distance from primary roads	0	3252	m
Predictor	Y	Latitude	4758747	4885247	m UTM
Predictor	X	Longitude	1602201	1758401	m UTM

Methodology

In order to assess the fire occurrence based on predictors variables, and keeping in mind the constraints described in the introduction, the CART and MARS models were processed account to identify regression rules and/or regression coefficients. The proposed models have several advantages such as being less restrictive in terms of assumptions, they non-parametric retrieving the data distribution from the training dataset, they can handle discrete and continuous variables and the resulting regression functions are easy to interpret. Nevertheless, they can not be considered pure spatial model such as GWR, therefore, the introduction of spatial predictors such as the x and y coordinates are fundamental to provide spatial relationship to other predictors.

CART and MARS theories

The CART model operates by recursively splitting the data until ending points, or terminal nodes, are achieved (Stroppiana et al., 2003). It begins by analyzing all the input variables and determining which binary division of a single predictor variable best reduces deviance in the response variable. The process is repeated for each portion of the data resulting from the first split, continuing until homogeneous

terminal nodes are reached in the hierarchical tree. The technique creates a tree that explains substantially all of the deviance in the original data. The CART model returns response average values for a delimited area, giving the typical aspect of a zonal map. Beside, the MARS model (Friedman, 1991) is an innovative and flexible tool that automates the building of accurate predictive models for continuous and binary dependent variables. MARS excels at finding optimal variable transformations and interactions that are common present in large geo-datasets (Leathwick et al., 2006). The main functions are defined in pairs using a knot, or value of a variable, which identify an inflection point along the range of predictor. When fitting a MARS model the knots are chosen automatically in a forward stepwise manner (Hastie and Tibshirani, 1990). The results of a MARS model is a map predicting smoothed response values.

Results

Fire occurrence

The map shows in Figure 1 depicts the fire ignition density in terms of fire ignition point per km^2 of wildland. Two main hot spot areas can be identified. The largest one, located in the central-North-western part of the study area, consists of several peaks ranging from 2.5 to 3.5 ip/km^2. The smallest one is situated in the central-South-eastern part and has a main peak of 3.74 ip/km^2 and smaller ones of 2 ip/km^2. In the remaining part of the study area the fire occurrence slightly decreases to values close to 0 ip/km^2. The spatial pattern of the fires distribution has the typical behaviour of clustered phenomenon, due to the persistence of the human-caused wildfires. The most prone areas to the wildfire are the wildland in contact with agriculture areas and subjected to high level of fragmentation. The typical spatial pattern of the fire occurrence is quite well know in forest fire literature and is very common in Mediterranean landscape where the human pressure is very high (Leone and Lovreglio, 2003; Maselli et al., 2003; de la Riva et al., 2004; Amatulli et al., 2007).

Variable correlations

Table 2 shows the correlation coefficient among the response variable and the continuous predictor variables. As it can be noted all the coefficients do not reach high values due to the complexity of the fire phenomena. Nonetheless, some observations can be drawn. The road variable produces a positive correlation when they are expressed in terms of road density, on the contrary low and slightly negative correlation is founded in case of road distance. A similar situation is detectable for the road classes. The secondary roads, expressed in terms of density, are the most critical ones due to their presence in the wildland areas. The topographic variables have low correlation with fire occurrence due to the fact that they are more important for fire spreading rather than for fire occurrence. The geographic variables, x and y, are also correlated

Table 2. Correlation Coefficient (r) of each continuous predictors and the response variable.

Predictor Variables	Correlation Coefficient (r) against the response variable.
X	-0.265
R2DIST	-0.159
R1DIST	-0.121
ALT	-0.076
ASP_COS	-0.003
ASP_SIN	0.002
R1DENS	0.002
POP	0.071
SLOPE	0.116
Y	0.284
R2DENS	0.402

with the fire occurrence. Lastly population density does not give information on the spatial pattern of fire density.

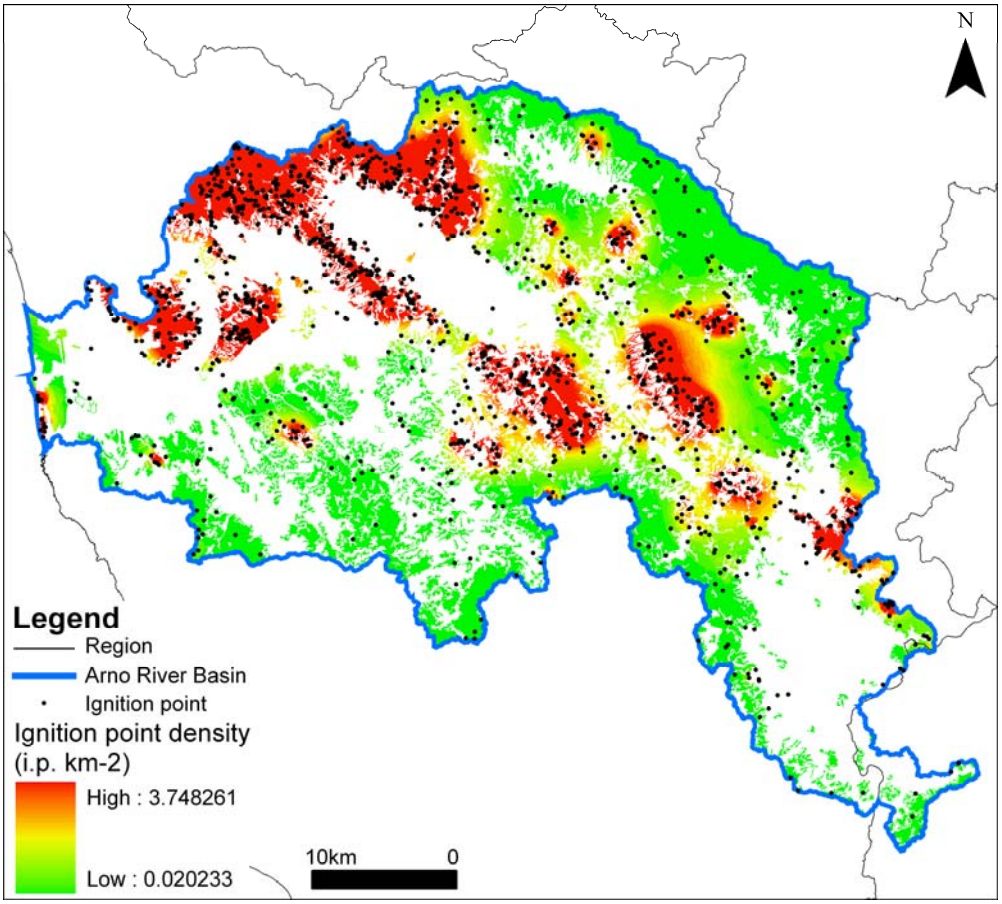

Figure 1- Study area, ignition point location and density (in relation to the wildland land cover area) of the Arno River Basin.

The CART model

The predicted fire density map estimated using the CART model is depicted in Figure 2. The number of predictor variables, the high heterogeneity of the study area, in the sense of topography and wildland patches, and the strong variability of the fire occurrence map, created a very complex tree with 67 nodes. The rules identified several thresholds unique for each variable and specific for each unit, useful to predict 67 average densities, ranging from 0 to 2.65 ip/km^2, smoothing the maximum values (3.74 ip/km^2) of the fire occurrence map. The unit size ranges from quite large (47,303 grid cells) to small areas (51 grid cells). Analyzing the spatial shape of the resulting units, a squared pattern was detected in some units. This can be seen in the enlarged circled area in Fig. 2. This squared shape of the occurrence unit is due to the use of coordinates as predictors.

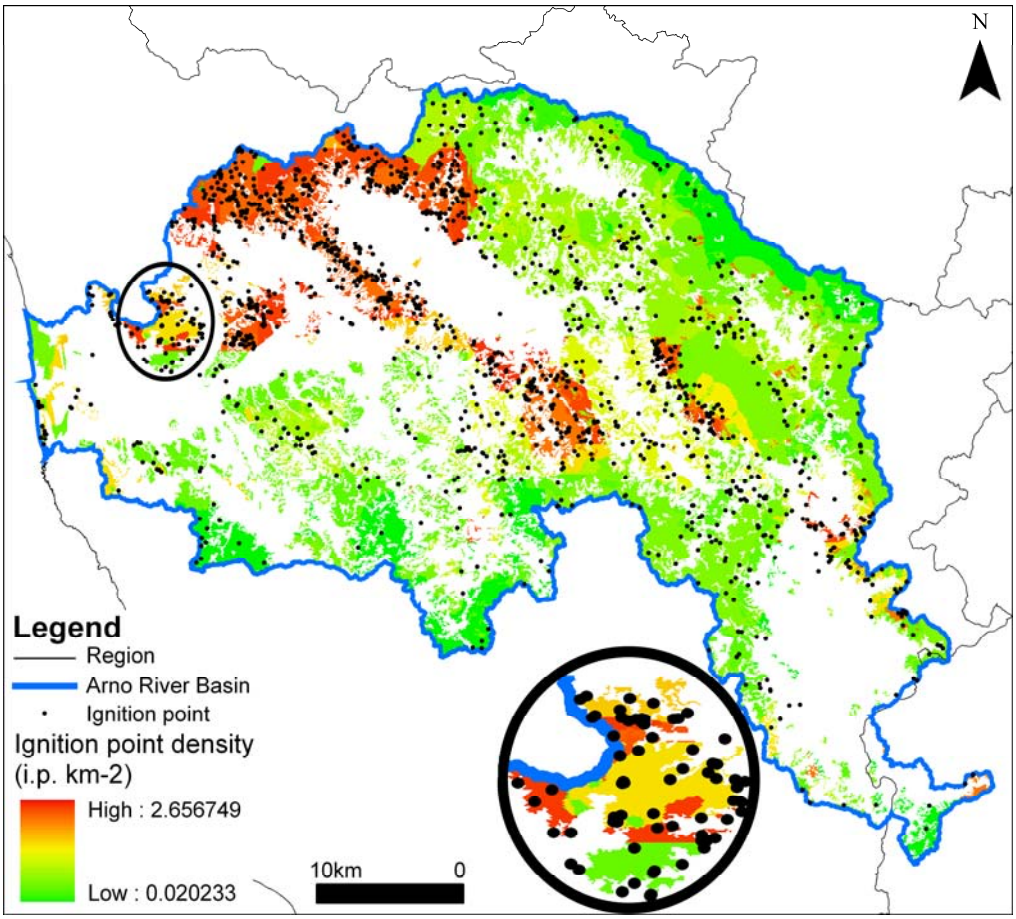

Figure 2 - Fire risk map obtained implementing the decision rules of the CART model.

The MARS model

The estimated fire density map derived from the implementation of the regression coefficients obtained with the MARS model is depicted in Figure 3. As it can be seen from visual interpretation, the predicted fire occurrence has a smoother surface compared to the CART model. Also the selected area (circle in Fig. 3) shows smoother transition compared to the CART one. The estimates reach very low

maximum value (0.70 ip/km^2) and a negative minimum value (-0.66 ip/km^2), clearly under estimating the target variable due to the difficulty of modelling the fire phenomenon by means of unique coefficients. In fact, in the MARS setting the number of bending points (variable discontinuity) was set to a minimum value, producing little fluctuations in the predictions.

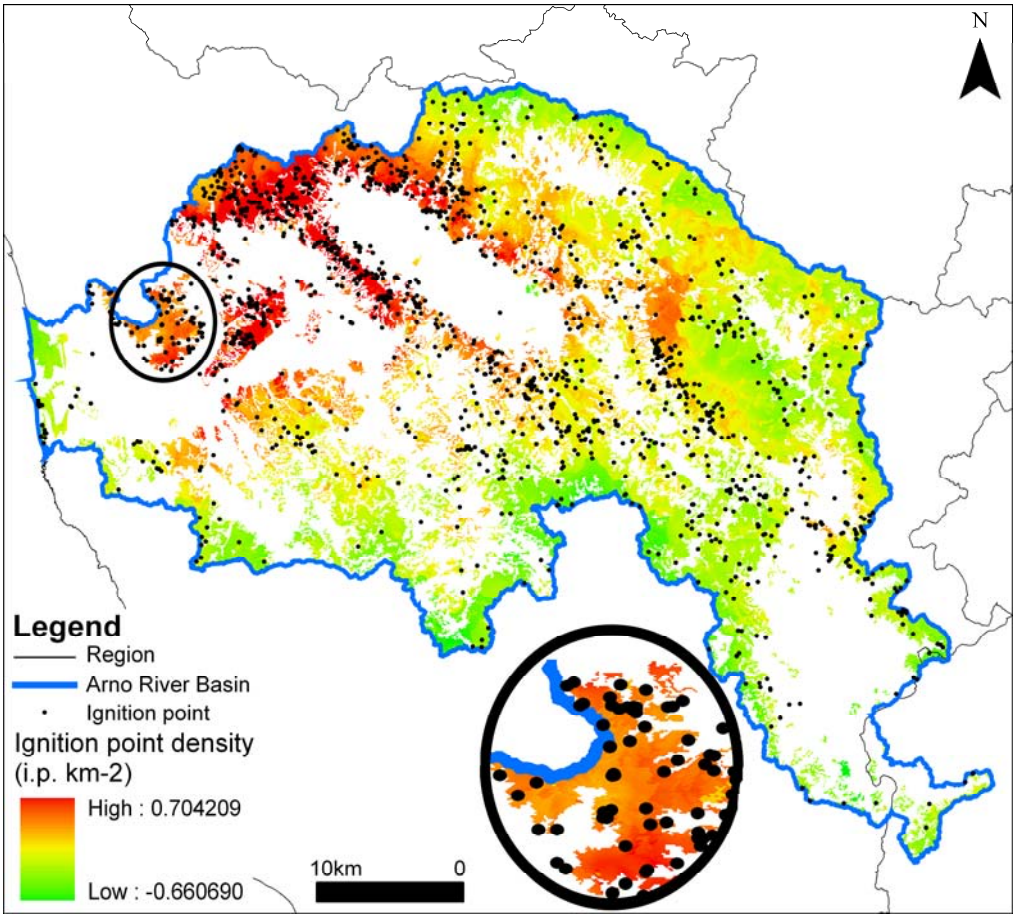

Figure 3 - Fire risk map obtained implementing the regression coefficients of the MARS model.

Scatter plots of observed versus predicted fire density values are given in Figure 4 for the CART and MARS models. The correlation coefficient (r) was 0.83 in the case of CART and 0.56 for the MARS model. For the later it can be noted that the extreme values are not well estimated.

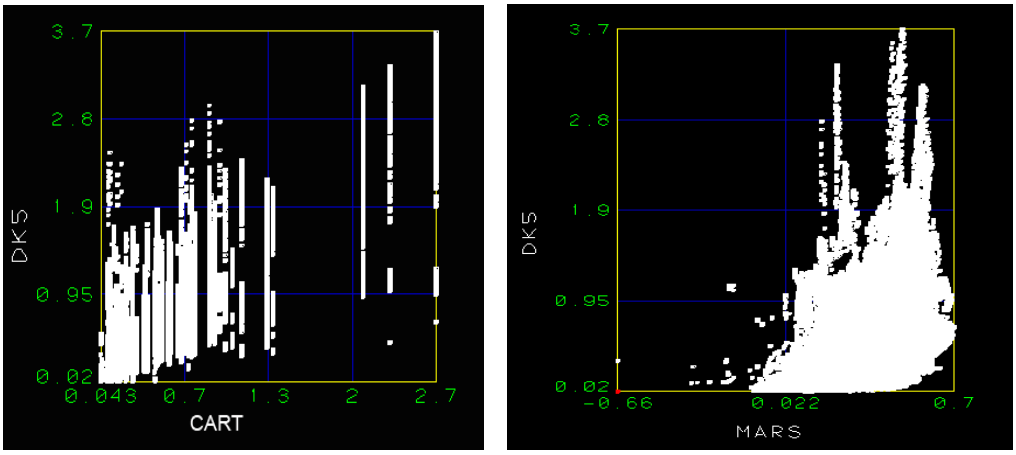

Figure 4 – Scatter plots of observed versus predicted fire density obtained with the CART (left) and MARS (right) models. In the left hand plot the straight line pattern is due to the typical output of the regression tree models: average values for unit areas. On the contrary the plot to the right depicts the correlation between two continuous variables. In this case the estimated values are smoothed resulting in higher homogeneity in the predicted values.

Conclusions and recommendations

From the application of the two models for the prediction of fire occurrence in the framework of long-term fire risk assessment it can be concluded that the CART model shows a better performance in terms of prediction power. In addition it gives an output that identifies homogenous fire risk management units, that can be useful to support wildfire planning actions. On the other hand the MARS model is able to produce a smoothed prediction surface. The two models were set up using default parameters, therefore better performances could be found testing other setting parameters. The results were also useful to analyze the behaviour of each independent variable in the regression process. The positive relationship of the road variable when expressed in terms of road density is confirmed as in previous works (Chuvieco et al., 1999; Amatulli et al., 2006). The use of x and y variables can supply spatial information to the models, nonetheless they have to be used carefully since unforeseen results can be produced.

Acknowledgments

This work was funded by the European Community under the Sixth EU Framework Programme for Research and Technological Development. Research was carried out as part of the collaborative project “Applied multi Risk Mapping of Natural Hazards for Impact Assessment” (ARMONIA).

References

Amatulli, G., Perez-Cabello, F. and de la Riva, J., 2007. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. Ecological Modelling, 200:321-333.

Amatulli, G., Rodrigues, M.J. and Lovreglio, R., 2005. Mapping forest fire occurrence at national level - Assessing fire density by means of the Adaptive Kernel Density Technique. In: J. De la Riva, F. Peréz-Cabello and E. Chuvieco (Editor), Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to Forest Fire Management: Fire Effects Assessment, Zaragoza, Spain, pp. 51-55.

Amatulli, G., Rodrigues, M.J., Trombetti, M. and Lovreglio, R., 2006. Assessing long-term fire risk at local scale by means of decision tree technique. Journal of Geophysical Research - Biogeosciences, 111.

Breiman, L., Friedman, J.h., Olshen, R.A. and Stone, C.J., 1984. Classsification and regression tree. Wadsworth International Group, Belmont, CA.

Chuvieco, E., Salas, F.J., Carvacho, L. and Rodríguez-Silva, F., 1999. Integrated fire risk mapping. In: E. Chuvieco (Editor), Remote Sensing of Large Wildfires in the European Mediterranean Basin. Springer-Verlag, Berlin, pp. 61-84.

de la Riva, J., Pérez-Cabello, F., Lana-Renault, N. and Koutsias, N., 2004. Mapping wildfire occurrence at regional scale. Remote Sensing of Environment, 92:363-369.

ESRI, 1997. Environmental Systems Research Institute.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.

Friedman, J.H., 1991. Multivariate adaptive regression splines. The Annals of Statistics, 19:1-141.

Hastie, T.J. and Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London, UK.

Koutsias, N., Martínez, J., Chuvieco, E. and Allgöwer, B., 2005. Modeling Wildland Fire Occurrence in Southern Europe by a Geographically Weighted Regression Approach. In: J. De la Riva, F. Peréz-Cabello and E. Chuvieco (Editor), Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to Forest Fire Management: Fire Effects Assessment, Zaragoza, Spain.

Leathwick, J.R., Elith, J. and Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling, 199:188-196.

Leone, V., Koutsias, N., Martínez, J., Vega-García, C., Allgöwer, B. and Lovreglio, R., 2003. The human factor in fire danger assessment. In: E. Chuvieco (Editor), Wildland Fire Danger Estimation and Mapping - The Role of Remote Sensing Data. Singapore, Singapore, pp. 143-196.

Leone, V. and Lovreglio, R., 2003. Human fire causes: a challenge for modelling. In: E. Chuvieco, P. Martín and C. Justice (Editor), Proceedings of the 4th Intern Workshop on Remote Sensing and GIS Applications to Forest Fire Management. Innovative Concepts and Methods in Fire Danger Estimation, Ghent, Belgium.

Maselli, F., Romanelli, S., Bottai, L. and Zipoli, G., 2003. Use of NOAA-AVHRR NDVI images for the estimation of dynamic fire risk in Mediterranean areas. Remote Sensing of Environment, 86:187-197.

McKenzie, D., Peterson, D.L. and Agee, J.K., 2000. Fire frequency in the interior Columbia River basin: building regional models from fire history data. Ecological applications, 10:1497-1516.

Stroppiana, D., Gregoire, J.-M. and Pereira, J.M.C., 2003. The use of SPOT VEGETATION data in a classification tree approach for burnt area mapping in Australian savanna. International journal of remote sensing, 24:2131-2151.