Data Min Knowl Disc
DOI 10.1007/s10618-011-0213-2

Estimating the risk of fire outbreaks in the natural environment

Daniela Stojanova · Andrej Kobler · Peter Ogrinc · Bernard Ženko · Sašo Džeroski

Received: 21 May 2010 / Accepted: 29 January 2011

Abstract A constant and controlled level of emission of carbon and other gases into the atmosphere is a pre-condition for preventing global warming and an essential issue for a sustainable world. Fires in the natural environment are phenomena that extensively increase the level of greenhouse emissions and disturb the normal functioning of natural ecosystems. Therefore, estimating the risk of fire outbreaks and fire prevention are the first steps in reducing the damage caused by fire. In this study, we build predictive models to estimate the risk of fire outbreaks in Slovenia, using data from a GIS, Remote Sensing imagery and the weather prediction model ALADIN. The study is carried out on three datasets, from three regions: one for the Kras region, one for the coastal region and one for continental Slovenia. On these datasets, we apply both classical statistical approaches and state-of-the-art data mining algorithms, such as ensembles of decision trees, in order to obtain predictive models of fire outbreaks.

Responsible editor: Katharina Morik, Kanishka Bhaduri and Hillol Kargupta.

This paper has its origins in a project report (Kobler et al. 2006) and a short conference paper (Stojanova et al. 2006) that introduced the problem of forest fire prediction in Slovenia, using GIS, RS and meteorological data. However, this paper significantly extends and upgrades the work presented there. In particular: We consider a wider set of data mining techniques, from single classifiers to ensembles; We present a comparison of the predictive performance in terms of several frequently used evaluation measures for classification; We present an example of the results obtained from the modeling task in the form of decision rules, explain and interpret their meaning; We generate geographical maps and compare them with other fire prediction models (e.g., FWI fire risk danger maps) provided by other services.

D. Stojanova (✉) · B. Ženko · S. Džeroski
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: daniela.stojanova@ijs.si

A. Kobler · P. Ogrinc
Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, Slovenia

Published online: 02 March 2011

In addition, we explore the influence of fire fuel information on the performance of the models, measured in terms of accuracy, Kappa statistic, precision and recall. Best results in terms of predictive accuracy are obtained by ensembles of decision trees.

Keywords Fire outbreaks · Fire prediction · Greenhouse emission · Remote sensing · Classification · Rules · Trees · Ensembles

1 Introduction

1.1 Fires: threat, damage and consequences

Fires present a global threat to the natural environment. They violate the functions of natural ecosystems and can cause a serious damage to the natural environment and human assets. Even though fires can also have a beneficial ecological function, e.g., start the rejuvenation of a forest, in most cases they cause significant material damage, both from an economic and an ecological point of view. The damage is especially reflected in ecosystem services, landscape structure and global infrastructure, as well as in species composition, biodiversity of ecosystems and human life. Fires increase the emissions of particles and gases into the atmosphere (especially carbon dioxide). They also alter the water infiltration rates in the soil, making burnt areas more prone to erosion, soil loss and landslides. The extent of damage caused by these natural phenomena can rise to critical levels, especially when combined with droughts, changes in landuse, wind and topographical factors.

Furthermore, extreme and frequent fires degrade habitat quality and destroy ecosystems, including forests, which need time to develop. The forests themselves are a very important segment of the natural environment. They serve as a natural carbon sink by accumulating and storing carbon dioxide, at the same time releasing oxygen and helping in the reduction of greenhouse gas emissions into the atmosphere. Sustainable management practices keep forests growing at a higher rate over a potentially longer period of time, thus providing net sequestration benefits in addition to those of unmanaged forests (Ruddell et al. 2007).

Each year, millions of hectares of forest are destroyed all around the world because of forest fires. During the fires of 2007, an overall 575,531 ha of forest were destroyed in various European nations. From 1980 to 2006, a total of about 1.33 million ha of forest land has been ruined by the fires (European Commission 2008). The fires of 2007 were dramatic both in terms of the size of the affected territory and the number of human deaths caused. The total greenhouse gas emissions resulting from fires in the period from 1994 to 2007 were estimated to 12.5 million tons of carbon dioxide (CO_2).

The deforestation caused by fires further contributes to the increase of CO_2 emissions into the atmosphere. The emissions of particles and gases into the atmosphere can rise to critical levels when combined with extreme weather conditions, especially high temperatures and long drought periods enclosed by storms and winds. With a 2°C increase in average temperature, there is a 30% increased risk of significant deforestation in the northern forests of Eurasia, eastern China, Canada, and the trop-

ical rainforests of central America and the Amazon. This risk would rise to 60% and affect wider areas if temperatures rise by 3°C (Connor 2006).

In short, fires increase the emissions of particles and gases into the atmosphere, especially CO_2, which is believed to be one of the inducers of global warming. They reduce the services and benefits we obtain from natural ecosystems, violate the environmental equilibrium and cause significant material damage. The consequences of fires, combined with the effects of global warming and other hazards, are threatening and can lead to pest outbreaks, changed land usage, as well as new fire outbreaks.

1.2 Fires in Slovenia

Fires are also a threat to the natural environment in Slovenia. Given that a large percent of the country is covered by forests (about 60%), the risk of forest fires is a serious issue. The risk of fires is strongly related to the weather conditions, especially the occurrence of extreme events (e.g., drought). The number of fires in Slovenia between 1994 and 2008 varied as shown in Fig. 1.

The differences in climatic conditions between regions in Slovenia are very large, which implies different fire risks between regions. In our study, we consider three different regions: continental Slovenia, coastal Slovenia and the Karst (Kras) region. Figure 2 presents a map of Slovenia where these regions are shown. Due to the climate differences among the three regions, we will investigate them separately.

The fire threat is not severe in the central part of Slovenia. However, the entire coastal region is threatened due to the sub-Mediterranean climate with high temperatures, long drought periods and powerful winds (Slovenia Forest Service 2005). The drought risk for the Slovenian coast is increasing, primarily due to climate change, and will be amplified by increased water shortage. Fire danger, length of fire season, fire frequency and severity are very likely to increase here and may lead to increased dominance of shrubs over trees.

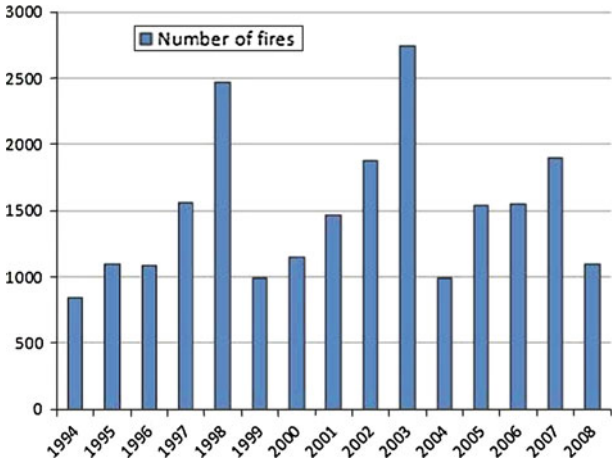

Fig. 1 The number of fires in Slovenia in the period between 1994 and 2008

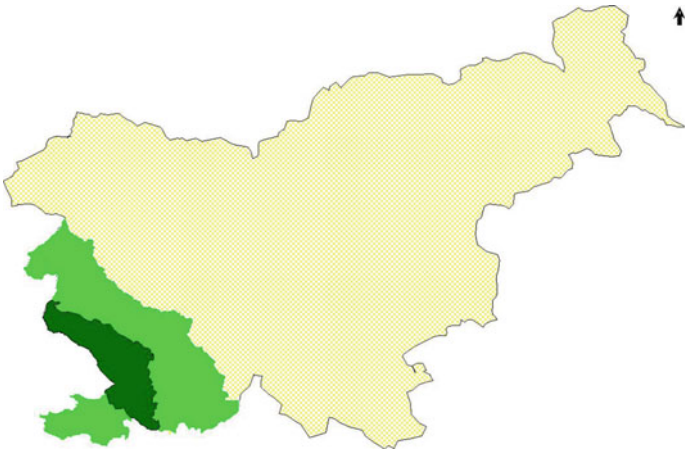

Fig. 2 A map of Slovenia where the study regions are shown. The coastal area is given in *light green (gray) color* whereas the rest of Slovenia is *light yellow (light gray)*. The Kras region that is a subset of the coastal region is presented in *dark green (black) color*

The highest number of fires is recorded in the Kras (Karst) region in the coastal area. Due to the hot and dry sub-Mediterranean climate, brownfield sites, as well as the vegetation adjusted to this type of climate, it has the highest fire risk for the natural environment in Slovenia. More than half of the forests in the Kras region are rated with the highest level of fire risk and at least 50 fires with a total area of over 600 ha occur on average each year (Kobler 2001). The fire threat is further increased by the transport corridors that lead through this area, in particular railways.

1.3 Modeling for reducing fire damage

The measures taken to reduce the damage caused by fire can be roughly divided into fire prevention and fire fighting measures. Measures from the first group aim to prevent fires from occurring in the first place, while measures from the second group aim to reduce the damage of fire spread. Modeling plays an important role for both: Models of fire risk are used in fire prevention, while models for detecting fires as well as models of fire spread and burn severity are used in fire fighting.

The modeling tasks related to fire prevention and fighting, as well as existing approaches to addressing these tasks, are discussed in detail in Sect. 2. Here we briefly touch upon modeling for fire prevention. In particular, we discuss the modeling of fire risk and the probability of fire outbreaks as its special case, as this is the topic we address in our application.

Models of fire risk (such as those predicting fire outbreaks) relate the fire threat (e.g., probability of outbreak) and the influence factors, such as availability of fuel and weather conditions. Because the influence factors are more or less geographically determined, such models are usually developed within Geographical Information Systems (GIS). The models can be built manually, by using domain knowledge, or in

an automated fashion with machine learning (ML), by using historical data on fires (fire outbreaks) and influence factors.

An example of a manually developed model is the Fire Weather Index (FWI) (Turner and Lawson 1978), which was originally developed in Canada. FWI is now in extensive use in the European Union through the European Forest Fires Information Service (EFFIS), where it has been modified to better suit the large differences in day length in the EU, when going from the Mediterranean to the Boreal countries. FWI is composed of six sub-indices, calculated from the weather parameters of temperature, relative humidity, wind speed, and rainfall. EFFIS calculates the FWI daily for all EU countries. The calculations are based on weather forecast data received daily from French and German meteorological services. The FWI is calculated at a spatial resolution of 36–45 km.

Increasingly more often, models for predicting fire occurrence or outbreaks are built by using historical data on fires and influence factors, to which statistical and machine learning approaches are applied. A variety of machine learning approaches have been applied in this context, typically one per study, with neural networks being the most commonly applied (and nearest neighbor, logistic regression, and decision trees being applied occasionally). The spatial resolution also varies across the studies, but is typically coarse (going up to 0.25°).

1.4 State of the art in Slovenia

As an EU country, Slovenia is covered by EFFIS (mentioned above) through daily calculation of the Fire Weather Index (FWI): However, the spatial resolution of this coverage is very coarse (36–45 km). EFFIS also contains historical fire data for 20 countries, but not for Slovenia. Two operational systems (manually developed expert models) are used locally in Slovenia to assess the potential fire hazard in the natural environment: The first is a regional model and has been developed by the Environment Agency of Slovenia (EARS), while the second only concerns forest fires and has been developed by the Slovenian Forest Service (SFS). The EARS model has very low spatial resolution and fine-grained temporal resolution, while the SFS model has more fine-grained spatial, but coarse temporal resolution.

The Slovenian Forestry Institute (Kobler 2001) developed a regional model of fire risk, combining the fine-grained temporal resolution of the EARS model with the fine-grained spatial resolution of the SFS model. The validity of this model was limited to forests. The model was developed by using statistical approaches on a small quantity of data on previous forest fires: Hence, the achieved accuracy of the model was not adequate for operational use.

Following the above, we have derived two empirical GIS models of fire danger in the natural environment in Slovenia (Kobler et al. 2006). The models have fine-grained spatial and temporal resolution and are based on a much larger dataset of fires in the natural environment. The data, described in detail later in this paper, were used as input for the statistical approach of logistic regression. The resulting model predicts the likelihood of fire outbreaks in the natural environment. This model has

been deployed within a GIS on natural disasters, which is in daily use at the Administration for Civil Protection and Disaster Relief of the Ministry of Defence of Slovenia.

The model for predicting the likelihood of fire outbreaks does not give any further information about the fate of the fire, if the fire in fact occurs. Therefore, it does not anticipate the speed and manner of the fire development, nor does it predict burn severity. We have thus produced a second model that reflects to some extent the danger of fire spreading, in case it actually does break out. In the second model, the output of the first model is weighted by the wind speed (where the weights are based on expert experience): Given the same probability of a fire outbreak for two locations, greater weight will be given to the location where, according to meteorological forecasts, strong winds are expected. This second model has also been deployed as described above.

1.5 The objective of this study

The aim of this study is to build improved models that predict the risk of fire outbreaks in Slovenia by using state-of-the-art data mining techniques, as assessed by predictive accuracy and other relevant performance measures (such as precision and recall). Much like the recent related work (Kobler et al. 2006), we will use data from GIS, Remote sensing (RS) imagery, and weather predictions by the model ALADIN (Aire Limitée Adaptation Dynamique Développement International) (Fischer et al. 2006). Similarly, we will carry out the study on three datasets, from three regions: one for the Kras region, one for the coastal region and one for continental Slovenia. On these datasets, we will apply both classical statistical approaches and state-of-the-art data mining algorithms, such as ensembles of decision trees, showing that the latter perform better.

The remainder of the paper is organized as follows. The next section presents related work along several lines of fire damage prevention research. In Sect. 3, we describe the data, in Sect. 4 the methodology used in this study, and in Sect. 5 the experimental setup of the predictive modeling process. Next, in Sect. 6 we present the obtained models, explain and interpret their meaning, present the fire outbreak risk maps that we generated with our models, and discuss the problem of predicting fire outbreaks in the natural environment. Finally, in Sect. 7 we outline our conclusions and discuss possible directions for further work.

2 Related work

Many measures can be taken in order to reduce the damaging effects of fires. We can roughly divide them into fire prevention and fire fighting measures. The first are supposed to prevent the occurrence of fires in the first place, while the second reduce the damage of fires that have occurred. Fire prevention measures include modeling the risk of fire outbreaks or predicting the probability of fire outbreaks, while fire fighting measures include early (automatic) fire detection and modeling of fire spread

and burn severity.[1] These measures are often supported by computer simulations of weather conditions, models of the fire risk and spread and possible fire damage scenarios. They are very important for successful fire prevention, organization of prevention measures and optimal allocation of fire-fighting resources. In the next paragraphs, we give an overview of existing studies related to these measures.

2.1 Fire outbreak prediction

Fire outbreak prediction can be viewed as the first step in reducing the damage caused by fires. An important tool for the prediction of fire risk is modeling of the relations between the fire threat and the influence factors (e.g., weather conditions, climate data, direction and speed of the wind, etc.). Because these factors are more or less geographically determined, such models are usually developed within a GIS. The models can be constructed manually or built with machine learning or statistical techniques.

Vega-Garcia et al. (1996) applied Neural Networks (NN) to predict human-caused wildfire occurrence in Canada. Within a GIS, they analyzed the historical occurrence of fire data, the Fire Weather Index for a given day, the geographical area of the fire occurrence and the forest districts with high human use. These data were also analyzed with logistic regression models, which served as a "domain expert" to identify the important input variables. The resultant model had four input nodes and two output nodes and correctly predicted 76% of the fire and non-fire observations on the test data.

Alonzo-Betanzos et al. (2003) also used neural networks to predict fire risks classified into four symbolic categories, and obtained an accuracy of 78.9%. Based on daily meteorological data, they built an intelligent rule-based system for forest fire risk prediction and fire fighting management. The application area was the Galicia region in Spain, one of the regions of Europe most affected by fires.

Cheng and Wang (2008) presented an integrated spatio-temporal forecasting framework that uses dynamic recurrent neural networks for forecasting the annual average area of forest fire, based on historical observations. Comparative analysis of this framework with other methods shows its high accuracy in short-term prediction. Its use was illustrated by a case study of forest fire area prediction in Canada.

Felber and Bartelt (2003) used the *k*-Nearest Neighbors algorithm to compare past fire occurrences to current forecast conditions in order to predict forest fire danger in the Swiss Southern Alps. The Nearest Neighbors models are attractive since they are intuitive in their operation and their data requirements are relatively modest. The data used were fuel, weather and forest fire data over a period of several years.

Preisler et al. (2008) developed a statistical method based on logistic regression technique for estimation of the monthly probabilities of large fires on a 1-degree grid cell over the western US. The model used 25 years of historical fire occurrence data, monthly average temperature and monthly mean fire danger indices (FWI, Drought Index, Ignition and Energy Release Component). The statistical models were particularly amenable to model evaluation and production of probability-based fire-danger maps with pre-specified precision. During the 25 years of the study, for the month

[1] Here we discuss in more detail the measures that are related to our study.

of July, fires occurred at 3% of locations where the model forecast was low; 11% of locations where the forecast was moderate; and 76% of locations where the forecast was extreme.

Locatelli et al. (2008) used climatic monthly data for the 1998–2007 period with a 0.25 degree spatial resolution and built decision trees to model the occurrence of forest fires in Central America. The decision trees resulted in 75% accuracy on the 1998–2007 period. Using climate change and socio-economic scenarios, as well as fuzzy indicators of fire risk, they also applied the decision trees to future conditions to create maps of future fire risks, which showed that in some areas fire risk is changing.

2.2 Fire detection

The goal of fire detection is to automatically detect fires that are already active. A large body of research on fire detection exists, where RS of fires is performed using a variety of space-borne systems and sensors. The most widely used sensors for long-term and large-scale fire monitoring are the Advanced Very High Resolution Radiometer (AVHRR), Defense Meteorological Satellite Program (DMSP), Along Track Scanning Radiometer (ATSR), Landsat (Mack 1991) and the Moderate Resolution Imaging Spectroradiometer (MODIS).[2]

Li et al. (2001) present a review of AVHRR-based Active Fire Detection Algorithms in three general categories: single channel threshold algorithms, multichannel threshold algorithms, and spatial contextual algorithms. Five fire detection algorithms (IGBP, MODIS, ESA, CCRS, and an approach by Giglio et al. (1999)) were compared by applying them across the Canadian boreal forest for a 6-month period and comparing cumulative fire pixels with a ground-truth data set. The performance of the algorithms under evaluation differed drastically, which implied that the hot spot detection algorithms are not robust enough for global operational use and no single-sensor algorithm is optimal for global fire detection.

The Enhanced Contextual Fire Detection Algorithm for MODIS by Giglio et al. (2003) runs as a part of the MODIS Rapid Response System, providing information about actively burning fires, including their location and timing, instantaneous radiative power and smoldering ratio, presented at a selection of spatial and temporal scales. Services such as the FIRMS Web Fire Mapper[3] use MODIS data as a data source to map the fire locations and burn areas around the world.

Other detection techniques based on image processing can be used for detection of forest fire spots in satellite images, as well. For example, spatial clustering (FASTCiD) was adopted by Hsu et al. (2002) to detect forest fire spots in satellite images. The clustering problem was presented in terms of image mining as an interdisciplinary endeavor that draws upon expertise in computer vision, image processing,

[2] MODIS Rapid Response System: http://rapidfire.sci.gsfc.nasa.gov/.

[3] FIRMS Web Fire Mapper: http://firefly.geog.umd.edu/firemap/.

image retrieval, data mining, machine learning, databases, and artificial intelligence. Advances in image acquisition and storage technology like object recognition, image retrieval, image indexing, image classification and clustering, association rule mining and neural networks can reveal useful information to the human users and have led to tremendous growth in very large and detailed image databases.

Mazzoni et al. (2005) obtained 75% accuracy at finding smoke at the 1.1-km pixel level by using satellite images from North America forest fires fed into a Support Vector Machine (SVM) pixel classifier. They used spectral, angular and textural features from the NASA's Multi-angle Imaging SpectroRadiometer (MISR) and matched areas containing smoke with fire locations identified by MODIS. For candidate scenes that appear to contain both smoke and fire, they applied machine vision techniques to look for evidence of plume-like shapes. When potential plumes were found, they automatically estimated source location, orientation, and injection height, using histograms of MISR stereo data for the latter. At the end, a human expert examined the results and discarded any false retrievals.

2.3 Modeling fire spread and burn severity

Modeling fire spread and burn severity is very important for preparing fire-fighting strategies in order to minimize the damage caused by fires and to use the limited resources as efficiently as possible. Markuzon and Kolitz (2009) used Random Forests, Bayesian networks and the k-Nearest Neighbor method for estimating fire danger, i.e., modeling the probabilistic risk of a currently burning fire becoming large and dangerous. They used data from MODIS images and weather information. None of the classifiers showed significantly superior performance over the others. However, the study demonstrated a significant predictive power of fire models that are based on remote sensing observations. The combination of data sources with different modalities increased the predictive power of the models to useful levels.

Cortez and Morais (2007) predicted the burned area of forest fires using SVMs and Random Forests. Four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes) were tested on recent real-world data collected from the northeast region of Portugal. The best configuration uses a SVM and four meteorological inputs (i.e., temperature, relative humidity, rain and wind) and is capable of predicting the burned area of small fires, which are more frequent.

Holden et al. (2009) predict the burn severity for the Gila National Forest from historical (over a 20-year period) burn severity data, topographic and biophysical variables. They used the random forest algorithm and a stratified random sample to build an empirical model predicting the probability of occurrence for severe burns across the entire study area. The model classified severity with a classification accuracy of 79.5%. Severe fires occurred more frequently at higher elevations and on north-facing, steep slopes and at locally wet, cool sites, which suggests that moisture limitations on productivity in the southwestern US interact with topography to influence vegetation density and fuel production that in turn influence burn severity.

3 Materials and methods

3.1 The study area

In this study, we investigate three datasets that contain data from different parts of Slovenia: the Kras region, which represents a sub-region of the coastal region of Slovenia (70 km^2, Kras dataset), the coastal region (132 km^2, Coastal dataset), and the continental part of Slovenia (20141 km^2, Slovenia dataset). The area of the Kras dataset is included in the Coastal dataset. Figure 2 presents a map of Slovenia where these regions are shown.

3.2 Data description

The data used for this study comprises fire outbreaks in the natural environment within Slovenia. Each location of a fire outbreak is described with a series of environmental attributes. These can be grouped into geographical, remote sensing and meteorological attributes.

3.2.1 The outbreaks

Each fire outbreak is a positive example for our data mining task. The data on fire occurrences were provided by the Administration for Civil Protection and Disaster Relief of Slovenia and the Slovenian Forestry Institute. They cover a 5 year period (2000–2004). Each fire outbreak is specified by the approximate location and time (date and hour) of the outbreak.

For building predictive models of outbreaks, we also need negative examples (of fire non-occurrence), which are generated from the positive examples by using the following procedure:

1. For each fire outbreak (positive example) repeat steps 2–4;
2. Find all fire outbreaks within the period of ± 3 days from the selected fire outbreak;
3. For all fire outbreaks found in step 2, make a 15 km (11 km for Kras dataset) surrounding region;
4. Outside the areas defined in step 3 randomly select one location in a forested region.

The locations of positive and negative examples of fire occurrence were next spatially and temporally linked to the descriptive environmental data.

3.2.2 Environmental data

The data describing the environment of the outbreaks includes geographical (GIS), remote sensing (RS) and meteorological data. The spatial unit of the analysis was a 1 km $\times$ 1 km quadrant and all the attributes were aggregated to this resolution.

Geographical (GIS) data. The geographical attributes are time independent and describe the following properties for each of the 1 km $\times$ 1 km quadrant: median

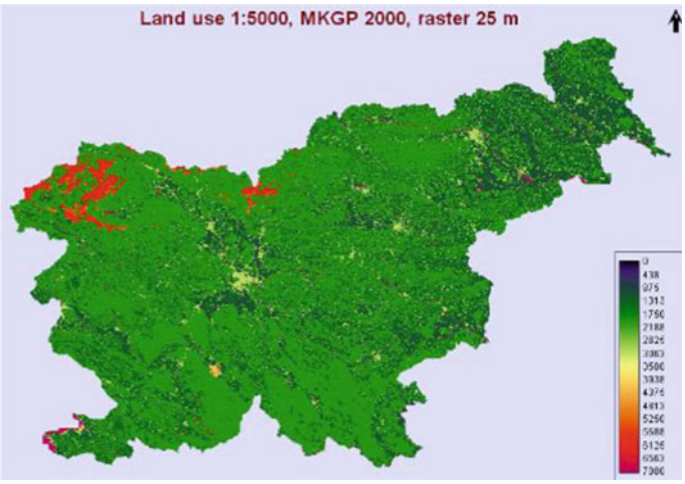

Fig. 3 The state landuse map, provided by the Ministry of Slovenia for Agriculture, Forestry and Food

altitude above sea level, slope and aspect of the relief, mode of exposition of the relief, distance to roads, highways, railways and settlements, and the distribution of land usage. The latter is represented as the percentage of land use of different types (e.g., fields, gardens, forests, buildings and others) and originates from the state landuse map presented in Fig. 3.

Remote Sensing (RS) data. Remote Sensing (RS) involves gathering of spatially organized data and information about an area of interest by detecting and measuring signals composed of radiation, particles and fields emanating from objects located beyond the immediate neighborhood of the sensor devices, offering a potential for more efficient resource assessment (Sabins 1978). RS can collect data on inaccessible and dangerous areas, and replaces expensive and time-consuming data gathering on the ground, ensuring in the process that areas or objects are not disturbed. RS observations can be used to distinguish among forest cover types on the basis of forest structure and species composition, to detect and quantify landscape pattern and structure, to give precise estimates of variables such as leaf area index and biomass for use in ecosystem process models.

Observations acquired at multiple scales and resolutions can be used to continuously estimate forest conditions from plots to stands to ecosystems. Multi-temporal RS observations are essential for various change detection applications. Our RS data includes multi-temporal MODIS and LiDAR data.

Multi-temporal MODIS data. MODIS (Moderate-resolution Imaging Spectroradiometer) (Chu et al. 2002) is an instrument launched into Earth orbit by NASA in 1999 on board the Terra and the Aqua satellites. The instruments capture data in 36 spectral bands ranging in wavelength from 0.4 μm to 14.4 μm and at varying spatial resolutions (2 bands at 250 m, 5 bands at 500 m and 29 bands at 1 km). The data is stored in text (ASCII) files, which cover 7×7 km of the field site.

Each row in the file contains data from one 8-day, 16-day, or annual period (depending on the temporal frequency of the data product represented). From the NASA archives of MODIS satellite images, we collected data on land temperature and net primary production of plants for the period of 5 years (2000–2004) with a spatial resolution of 1 km and time resolution of 8 days. The multi-temporal MODIS satellite data implicitly give information about the response of the vegetation in periods of drought and the types of fire fuels (King et al. 2003). The MODIS attributes describe the following properties: average temperature in Kelvin for a specific quadrant for a day x of the year, where x takes the values of $1, 9, 17, 25, \ldots, 361$ and average net primary production for a specific quadrant for day x of the year. For days other then $1, 9, 17, 25, \ldots, 361$, the closest available day is taken.

LiDAR data. In recent years, LiDAR (Light Detection And Ranging) has become one of the most promising RS techniques for detailed measurement of forest parameters. LiDAR is an optical RS technology that measures properties of scattered light to find range or other information of a distant target (Fujii and Fukuahi 2005). Like most of the passive or active systems, LiDAR can be used for mapping. The main characteristics of LiDAR are the high spatial resolution and 3D detailed measurements, which provide a more detailed picture of the complex forest structure than other passive optical sensors.

For the Kras region, we introduce additional fire fuel information derived from LiDAR data (Stojanova et al. 2010). These data contain attributes that describe the height structure of vegetation (Džeroski et al. 2006), (Kobler et al. 2006). The values of the attributes were derived by applying machine learning models to Landsat (Mack 1991) images of the study area. The models were learned by using both Landsat images and LiDAR data for a small subset of the study area (2 km $\times$ 20 km) and then applied to the entire Kras area. All the data have a resolution of 25 m $\times$ 25 m and are further aggregated to 1 km quadrants.

The LiDAR attributes describe the following forest properties: forest canopy cover, average vegetation height, maximum vegetation height and the vertical vegetation profiles (a histogram of shares of vegetation reaching the different heights). For illustration, the maps of vegetation height and canopy cover generated from LiDAR are presented in Fig. 4. For each of the attributes, we obtain 4 statistic measures: minimum, maximum, average and standard deviation on a 1 km quadrant.

Meteorological data. These data consists of weather forecasts made by the ALADIN/SI model, which is a version of the ALADIN (Fischer et al. 2006) meteorological model for Slovenia. The ALADIN Numerical Weather Prediction Project is a collaboration between the European Centre for Medium-Range Weather Forecasts (ECMWF) and Météo-France in the field of Numerical Weather Prediction (NWP), which provides the basis for the forecasting tools of modern meteorology. The model can run on computing systems with different power, from personal computers to supercomputers. The ALADIN model can simulate the atmosphere on any selected area on Earth using the initial state of the atmosphere in the region and the projected value at the edges of the area.

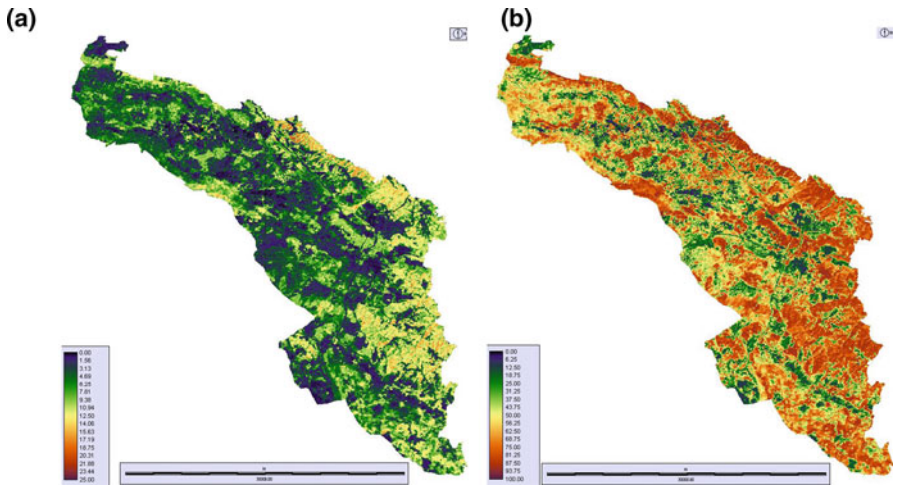

Fig. 4 Maps of (**a**) vegetation height and (**b**) canopy cover for the Kras region. The legend shows the vegetation height in meters/canopy cover in percentages

The ALADIN data used in this study contains meteorological weather predictions made by the ALADIN/SI model. These are issued daily by the Environment Agency of the Republic of Slovenia. The data include weather predictions for every 3 h (00.00–21.00 UTC) for 10 weather attributes: atmospheric precipitation, solar radiation energy, velocity, direction and speed of the wind, evapotranspiration, transpiration, evaporation, relative humidity and temperature. Average values of the meteorological parameters for 1, 2, 4 and 14 days are added in order to help in noise removal. The spatial resolution of the data is 11 km. The relief is taken into account implicitly, because the model for weather predictions is adapted for Slovenia.

3.3 Datasets

Given the purpose of this study, i.e., to predict fire outbreaks in the natural environment, the target attribute is nominal and is related to the fire occurrence ('yes' or 'no'). The list of the attributes with brief descriptions is presented in Appendix A. The different regions of Slovenia discussed above give rise to three classification datasets: continental Slovenia (Slovenia dataset), coastal Slovenia (Coastal dataset) and Kras region (Kras dataset).

The Slovenia dataset has 127 attributes and 8476 (4264 positive and 4212 negative) examples. The Coastal dataset has 106 attributes and 2442 (1229 positive and 1213 negative) examples. The Coastal and Slovenia dataset do not include additional LiDAR fire fuel attributes. The Kras dataset contains all the attributes discussed above and additionally the LiDAR fire fuel attributes derived by using predictive models generated by machine learning. The total number of attributes for the Kras dataset is 159. It has 1439 (959 negative and 480 positive) examples. In addition, we consider

the Kras dataset without the LiDAR fire fuel data (KrasWithoutLidar dataset), which has 126 attributes and 1439 (959 negative and 480 positive) examples.

4 Classification algorithms

The task we address is to estimate the risk of fires in the natural environment. The corresponding data mining task is to learn a model (or a set of models) for predicting fire occurrence in Slovenia. The target attribute is the occurrence of a fire and the descriptive attributes are extracted from GIS, the ALADIN weather prediction model and RS imagery as described above.

To learn a predictive model, we use several different classification algorithms that are implemented in the WEKA (Witten and Frank 2005) data mining suite. Each of the models gives as a prediction a probability estimate of fire occurrence (rather than just a binary answer of whether a fire will occur or not). We take these probabilities to be an estimate of the risk of a fire outbreak at a specific location at a specific time.

We decided to use a diverse set of classifiers in order to assess their suitability for our task. We included algorithms that induce interpretable models (e.g., decision trees and rules) as well as algorithms that tend to learn more accurate, but less interpretable models (such as ensemble methods). The single classifier methods include *k*-Nearest Neighbors, Naive Bayes, J48 decision trees, jRIP classification rules, Logistic regression, Support Vector Machines (SVM), Bayesian Networks, while the ensemble methods include Boosting, Bagging and Random Forests of decision trees.

4.1 Algorithms for learning single classifiers

k-Nearest Neighbors classifier. The *k*-Nearest Neighbors algorithm (KNN) (Aha et al. 1991) is an instance-based learning method, classifying examples based on the closest training examples in the feature space. An example is classified by a majority vote of its neighbors, with the example being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). The neighbors are taken from a set of examples for which the correct classification is known. This can be thought of as the training set for the algorithm, even though no explicit training set takes palce.

Naive Bayes. A naive Bayes classifier (NB) (John and Langley 1995) is a simple probabilistic classifier based on applying Bayes´ theorem (from Bayesian statistics) with strong (naive) independence assumptions. It assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independence of the variables is assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. It can

often work much better than one might expect in many complex real-world situations, even when the independence assumption does not hold.

J48. The J48 algorithm is an implementation of the C4.5 decision tree learner (Quinlan 1993). It uses the top down induction of decision trees approach, a greedy search technique. A decision tree (Quinlan 1986) is a hierarchical structure, where the internal nodes contain tests on the descriptive attributes. Each branch of an internal test corresponds to an outcome of the test and the prediction for the value of the target attribute is stored in a leaf. To obtain a prediction for a new example, the example is sorted down the tree, starting from the root (the top-most node of the tree). For each internal node encountered on the path, the test stored in the node is applied. Depending on the outcome, the path continues along the corresponding branch. The resulting prediction of the tree is taken from the leaf at the end of the path.

jRIP. The jRIP algorithm implements the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) propositional rule learner (Cohen 1995). JRip is a bottom up method that learns rules, which in the end cover all the examples. For each class from the less prevalent one to the more frequent one, it grows a rule, by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e., 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain. It balances overfitting with generalization and can efficiently handle large noisy datasets.

Logistic Regression. Logistic Regression (LogR) falls within the category of statistical models called generalized linear models (Agresti 1996). Logistic regression allows prediction of a discrete outcome, such as group membership, from a set of attributes that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response attribute is dichotomous, such as 'yes' or 'no'. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an odds ratio. Logistic regression also provides knowledge of the relationships among the variables and their strengths (Hosmer and Stanley 1989).

Support Vector Machines. Support Vector Machines (SVM) belong to the class of supervised learning algorithms in which the learning machine is given a set of attributes (or inputs) with the associated labels (or output values) (Saravanan et al. 2008). Each of these attributes can be looked upon as a dimension of a hyper-plane that separates the data into two classes (this can be extended to multi-class problems). The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin (Kandola et al. 2003). The most serious problem with SVMs, from a practical point of view, is the high algorithmic complexity, choice of the kernel and the high memory and time requirements.

Bayesian Networks. A Bayesian Network is a combination of a directed acyclic graph of nodes and links and a set of conditional probability tables. Nodes can

represent features or classes, while links between nodes represent the relationship between them. Conditional probability tables determine the strength of the links. There is one probability table for each node (feature) that defines the probability distribution for the node given its parent nodes. If a node has no parents the probability distribution is unconditional. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents.

Learning of a Bayesian network is a two-stage process. First the network structure is formed (structure learning) and then probability tables are estimated (probability distribution estimation). There are numerous combinations of structure learning and search techniques that can be used to create Bayesian Networks.

We use a Bayesian score metric to form the structure, while node quality is determined by using the Tree Augmented Naive Bayes (BNet) local search algorithm, where the tree is formed by calculating the maximum weight spanning tree and the Bayesian Metric (Bouckaert 2005). An estimation algorithm is used to create the conditional probability tables for the Bayesian Network. We use the Simple Estimator, which estimates probabilities directly from the dataset. The simple estimator calculates class membership probabilities for each instance, as well as the conditional probability of each node given its parent node in the Bayes network structure.

4.2 Algorithms for learning ensemble classifiers

An ensemble method constructs a set (called ensemble) of predictive models (called base models). The ensemble makes a prediction for a new example by combining the predictions of the individual models for that example. Ensembles perform better than individual classifiers, if the classifiers in the ensemble are accurate and diverse. The diversity in an ensemble can be introduced in different ways: by manipulating the training set (e.g., bootstrap sampling, change of weights of the data instances) or by manipulating the learning algorithm used to obtain the base models (e.g., introducing random elements in the algorithm). While ensemble methods can have much better performance than individual classifiers, the ensembles learned are large and difficult to interpret.

Boosting. Adaptive Boosting (AdaBoost) (Freund and Schapire 1996) is an algorithm for constructing classifiers by using a set of many weak or base classifiers in order to improve the overall performance. AdaBoost calls a given weak base learning algorithm repeatedly in a series of rounds. The algorithm maintains a distribution or a set of weights over the training set. Initially, all weights are set equally, but in each round, the weights of incorrectly classified examples are increased so that the base learner is forced to focus on the hard examples in the training set. At the end, the predictions of all weak classifiers are combined into a single prediction with weighted voting. AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. It is capable of reducing both bias and variance of the basic classifiers and has good generalization properties. The algorithm is sensitive to noisy data and outliers. However, it is less susceptible to the overfitting problem than most other learning algorithms.

Bagging. Bagging (Bag) (Breiman 1996) is an ensemble method that constructs different base models by making bootstrap replicates of the training set and using them to build the individual models. Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set. The bootstrap sample and the training set have an equal number of instances. Bagging can give substantial gains in predictive performance, when applied to an unstable learner (i.e., a learner for which small changes in the training set result in large changes in the predictions), such as classification and regression tree learners.

Random Forests. A random forest (RF) (Breiman 2001) is an ensemble of trees, where the diversity among the individual trees is obtained from two sources: (1) by using bootstrap sampling and (2) randomization of the attribute selection step of the tree generation algorithm. At each node in the decision tree, a random subset of the input attributes is taken and the best split is selected from this subset. The size of the random subset is a function of the number of descriptive attributes. Prediction is made by aggregation (majority vote for classification or averaging for regression) of the predictions of the individual models in the ensemble.

Random Forests produce highly accurate classifiers for many learning problems. The results are competitive with boosting and bagging. They are fast to build and work very efficiently for large datasets.

5 Experimental setup

5.1 Parameter settings for the classification algorithms

As described in Sect. 4, we use the following algorithms: *k*-Nearest Neighbors classifier (KNN), Naive Bayes (NB), J48 decision trees (J48), jRIP classification rules (jRIP), Logistic regression (LogR), Support Vector Machines (SVM), Bayesian Networks (BNet), as well as AdaBoost (AdaBoost), Bagging (BagJ48) and Random Forests (RF). We use these algorithms as implemented in the WEKA (Witten and Frank 2005) data mining suite.

The algorithm parameters in our experiments are set to the default values except: 5 nearest neighbors for the KNN classifier and 4 examples as the minimal number of examples that form a leaf for the trees and rules. For the ensemble methods, we use J48 as a base classifier and set the minimal number of examples that form a leaf to 4 examples, the number of iterations to 10 and the size of each bag to 100% of the training set size. The above mentioned settings were selected in a set of preliminary experiments, which investigated the influence of the different parameter settings on the accuracy of the results.

5.2 Performance measures

For a classification problem with two possible classes, most measures of performance are based on the four values of the contingency table, obtained by applying the classifier to the test set. These are the true positives TP, false positives FP, true negatives TN,

and false negatives FN. Using these values we define the following standard evaluation measures:

Accuracy: the proportion of correct predictions (both true positives and true negatives) in the population: A = (TP+TN)/(TP+TN+FP+FN);
Precision: the proportion of the true positives against all the positive predictions (both true positives and false positives): P = TP/(TP+FP);
Recall: the proportion of the true positives against all positives (the true positives and false negatives): R = TP/(TP+FN).

To evaluate our models we use accuracy, precision and recall. We also use the Kappa statistic and the Area Under the ROC Curve (AUC). The Kappa statistic is a measure of the degree of agreement between the predicted and observed classification of a dataset, which takes into account the agreement that occurs by chance.

It has been suggested that the Area Under an ROC Curve (AUC) can be used as a measure of performance in many applications (Swets 1988). A ROC graph (curve) is a plot with the false positive rate on the x axis and the true positive rate on the y axis. The AUC corresponds to the area under this curve. In classification, the AUC is interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

5.3 Statistical comparison

We estimate the above mentioned performance measures for our models on unseen examples by using 10 fold cross-validation. All the algorithms were evaluated on the same folds in order to allow statistical significance testing and comparison. To assess whether the differences in performance between the different algorithms are statistically significant, we use the corrected Friedman test (Friedman 1940) and the post hoc Nemenyi test (Nemenyi 1963) as recommended by (Demšar 2006).

The Friedman test is a non-parametric test for multiple hypotheses testing. It ranks the algorithms according to their performance for each dataset separately, thus the best performing algorithm gets the rank of 1, the second best the rank of 2 and so on. In the case of ties, it assigns average ranks. If the Friedman test shows statistically significant difference in performance, we can proceed with a post hoc test.

The Nemenyi test is used to compare all the classifiers to each other. In this procedure, the performance of two classifiers is significantly different if their average ranks differ by more than some critical distance. The critical distance depends on the number of algorithms, the number of datasets and the critical value (for a given significance level) that is based on the Studentized range statistic and can be found in statistical textbooks.

6 Results and discussion

As described in the previous section, we learned predictive models on four datasets (Slovenia—S, Coastal—C, Kras—K and Kras without Lidar—L) with a series of different learning methods. We now present these models and their predictive

performance. We give some interpretation for the human readable ones as they provided us with insight into the factors that influence fire outbreaks. At the end, we present the fire outbreak risk maps that we generated with our models.

6.1 Predictive models

We present the predictive performance of the obtained models in terms of their accuracy, Kappa statistic, precision, recall, and AUC. The results are presented in Table 1, grouped in panels a) to e), by performance measures. Each column contains the results of one algorithm. Each performance figure in each panel is the average of the 10 folds with the standard deviations.

The results of the significance tests are presented in the form of average rank diagrams in Fig. 5, for each evaluation measure separately. The ranks are depicted on the x axis, in such a manner that the best ranking algorithms are at the right-most side of the diagram. The critical difference (CD) interval, for the significance level of 0.05, is computed by the Nemenyi test and is plotted in the upper left corner; algorithms whose average rank difference is larger than this critical difference can be considered significantly different with 95% probability. The algorithms that do not differ significantly are connected with a horizontal line.

Overall, the best results in terms of predictive accuracy are obtained with bagging of decision trees. The Nemenyi test shows (Fig. 5a) that these results are statistically significantly better than Naive Bayes and kNN. If we take a look at the ensemble methods (Boosting, Bagging and Random Forests), we can see they are better than any of the single classifier methods, but the differences between them are not statistically significant. In general, most of the differences between the models are not significant due to the small number of datasets and the relatively high number of learning methods. Still, we believe the average ranks diagrams can be used to show the general trends in performance.

The accuracy as a measure of model quality is mostly used when the focus is on predicting the target attribute. In our case, the task is somewhat different, i.e., to estimate the risk of fire outbreaks or to estimate the conditional probability of the occurrence of a fire given the values of the other attributes. Therefore, some other quality measures might be more suitable for our task and this is the reason we decided to investigate several of them (besides accuracy, also precision, recall, Kappa, and AUC). However, the conclusions drawn using these measures are quite similar (see Fig. 5b–d).

Two measures that are especially important in fire outbreak prediction are precision and recall. High precision means that there are a small number of false positives (FP); in our case this means that we have a small number of false alarms, i.e., predicted fire outbreaks that do not actually happen. Recall (or sensitivity for binary classification), on the other hand, can be seen as a probability that a positive example (fire outbreak) is indeed predicted as positive, and is especially important in fire prediction since non-predicted outbreaks can be very costly. Therefore, if we are interested in a small number of false alarms, Bagging is again the most appropriate method. However, if we are more interested in the sensitivity of our predictions, Random Forests seem to be better (Table 1; Fig. 5).

Table 1 Comparison of the performance measures (Accuracy, Kappa Statistic, Precision, Recall and AUC) of the predictive models built in this study

	iBk	NB	J48	JRip	LogR	SVM	AdaBoost	BagJ48	RF	BNet
(a) Accuracy										
S	80.5 ± 1.1	81.0 ± 1.0	78.6 ± 1.2	81.5 ± 1.2	83.0 ± 0.8	83.0 ± 0.7	83.3 ± 1.2	**84.9 ± 1.9**	82.5 ± 1.2	81.7 ± 0.9
P	81.6 ± 1.5	74.9 ± 1.7	80.5 ± 1.8	81.4 ± 2.8	82.3 ± 2.1	83.3 ± 2.1	83.2 ± 1.7	**86.0 ± 1.6**	84.5 ± 1.3	83.2 ± 1.2
K	75.5 ± 1.7	63.8 ± 2.9	76.7 ± 2.8	77.5 ± 1.5	77.2 ± 2.1	78.1 ± 2.8	80.1 ± 1.6	**82.7 ± 1.1**	82.0 ± 1.7	77.3 ± 0.8
L	74.0 ± 1.4	62.8 ± 3.8	78.2 ± 1.7	79.4 ± 1.2	76.2 ± 2.1	77.4 ± 2.3	80.5 ± 1.7	**81.8 ± 2.6**	79.4 ± 1.4	77.0 ± 1.1
(b) Kappa statistic										
S	60.1 ± 2.3	63.0 ± 1.0	57.3 ± 0.5	63.1 ± 2.1	66.0 ± 0.9	66.0 ± 1.3	66.6 ± 1.3	**69.8 ± 1.9**	65.1 ± 1.5	63.4 ± 0.7
P	63.3 ± 1.7	50.0 ± 1.1	61.1 ± 1.1	62.7 ± 2.0	64.7 ± 1.9	66.6 ± 2.4	69.7 ± 1.4	**72.1 ± 2.0**	69.0 ± 1.3	66.3 ± 0.8
K	42.6 ± 0.9	28.8 ± 1.4	47.7 ± 2.1	47.5 ± 1.7	46.1 ± 1.0	47.6 ± 2.2	54.5 ± 1.2	**59.4 ± 2.9**	56.5 ± 1.7	49.3 ± 0.6
L	38.3 ± 2.0	26.4 ± 2.8	50.6 ± 2.0	51.2 ± 1.5	42.9 ± 0.4	45.1 ± 1.3	53.7 ± 1.3	**57.2 ± 2.9**	52.6 ± 1.3	48.8 ± 0.8
(c) Precision										
S	79.1 ± 1.3	79.5 ± 1.9	78.2 ± 1.2	79.5 ± 0.6	82.4 ± 1.2	82.4 ± 1.7	83.2 ± 2.0	**84.6 ± 0.9**	80.1 ± 2.4	81.2 ± 2.2
P	77.9 ± 1.1	79.3 ± 2.0	80.7 ± 1.3	79.5 ± 1.9	81.8 ± 1.2	82.0 ± 2.3	84.1 ± 1.5	**85.0 ± 0.8**	81.8 ± 2.2	83.3 ± 2.7
K	65.9 ± 1.6	47.3 ± 1.2	65.1 ± 1.6	69.4 ± 0.9	69.6 ± 1.1	72.7 ± 1.3	71.3 ± 1.4	**78.2 ± 0.9**	73.5 ± 2.3	65.5 ± 2.1
L	64.3 ± 1.4	46.3 ± 1.8	67.8 ± 1.9	73.6 ± 1.6	68.9 ± 1.9	73.1 ± 1.4	72.8 ± 1.4	**76.3 ± 1.1**	71.1 ± 2.7	64.8 ± 2.8
(d) Recall										
S	83.2 ± 1.2	81.0 ± 1.6	79.7 ± 0.4	84.7 ± 2.0	84.2 ± 2.0	84.2 ± 2.2	83.6 ± 1.5	84.7 ± 1.3	**86.8 ± 1.2**	82.9 ± 3.0
P	88.8 ± 1.1	67.8 ± 1.7	80.7 ± 1.2	84.9 ± 2.2	84.3 ± 1.9	85.6 ± 2.4	86.2 ± 3.0	86.9 ± 2.6	**89.0 ± 1.3**	83.2 ± 0.7
K	55.2 ± 1.6	74.4 ± 1.5	65.2 ± 1.0	58.5 ± 1.1	56.3 ± 0.4	55.0 ± 1.1	67.3 ± 0.9	66.5 ± 2.6	**67.5 ± 1.7**	67.5 ± 2.1
L	50.2 ± 1.4	71.7 ± 1.6	66.3 ± 1.3	59.4 ± 1.6	52.1 ± 1.2	51.0 ± 1.3	66.5 ± 1.5	65.2 ± 3.9	64.6 ± 1.4	**67.9 ± 0.8**

Table 1 continued

	iBk	NB	J48	JRip	LogR	SVM	AdaBoost	BagJ48	RF	BNet
(e) AUC										
S	87.6 ± 1.1	88.4 ± 0.7	81.1 ± 1.0	84.0 ± 0.3	90.8 ± 0.5	83.0 ± 2.2	91.1 ± 0.8	**91.9 ± 0.9**	90.7 ± 0.2	89.0 ± 0.8
P	88.6 ± 1.2	84.8 ± 2.1	83.2 ± 1.3	84.6 ± 1.2	90.5 ± 0.5	83.3 ± 1.9	92.0 ± 2.3	**92.6 ± 1.3**	92.2 ± 0.7	90.4 ± 0.7
K	79.1 ± 1.5	72.4 ± 1.9	75.6 ± 0.5	72.9 ± 1.4	81.2 ± 0.2	72.3 ± 1.8	84.8 ± 2.5	**88.8 ± 1.4**	84.3 ± 0.4	82.2 ± 0.9
L	75.6 ± 1.3	71.6 ± 1.7	78.1 ± 0.8	74.4 ± 1.6	81.3 ± 1.3	70.8 ± 1.8	85.8 ± 2.2	**87.8 ± 2.0**	83.5 ± 0.3	83.5 ± 0.8

The values of the performance measures were estimated by 10-fold cross-validation and are given in the format average ± standard deviation. The datasets labels are as follows: *S* Slovenia dataset, *C* Coastal dataset, *K* Kras dataset and *L* Kras without LiDAR dataset. The algorithm labels are as follows: *kNN* *k*-Nearest Neighbors classifier, *NB* Naive Bayes, *J48* J48 decision trees, *JRip* jRIP classification rules, *LogR* Logistic regression, *SVM* SVM classifier, *AdaBoost* Boosting, *BagJ48* Bagging, *RF* Random Forests, *BNet* Bayesian Network. The best results for each performance metric are given in bold face

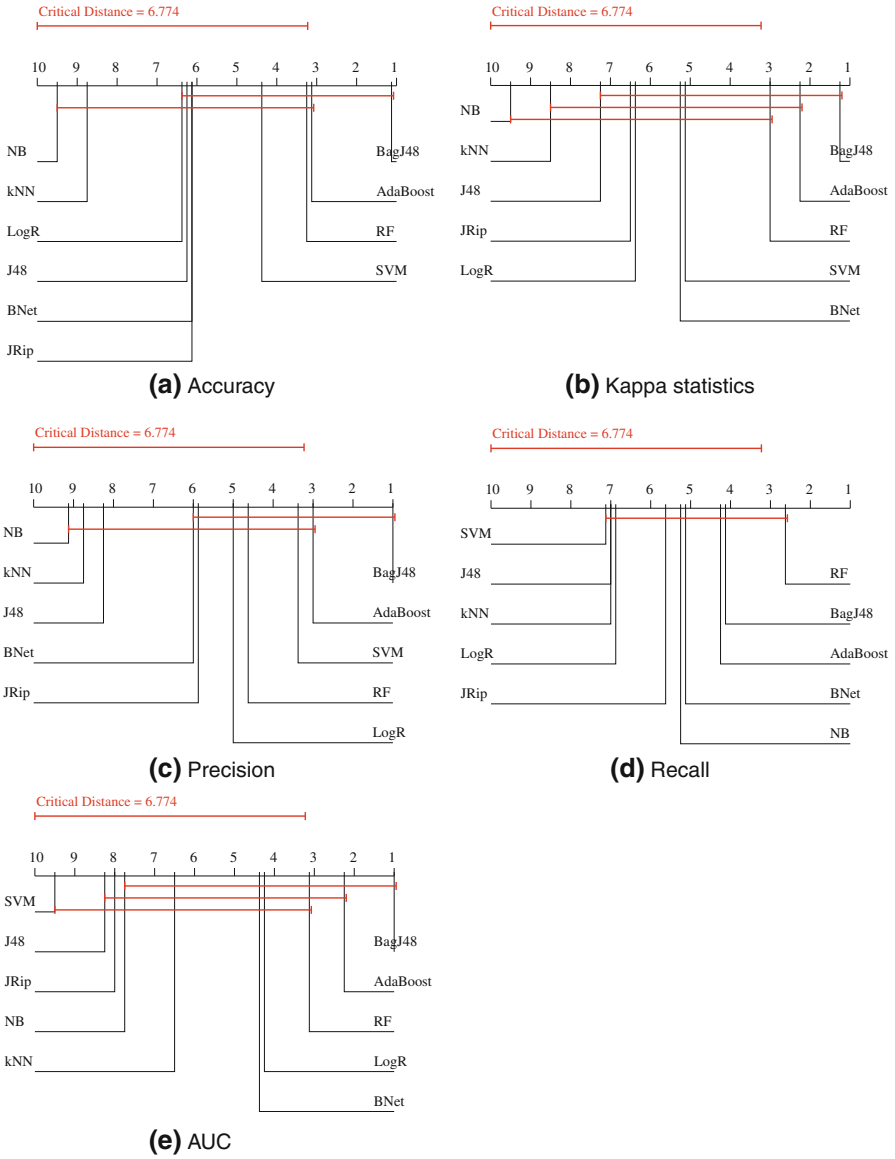

Fig. 5 Average ranks diagrams: (**a**) Accuracy, (**b**) Kappa Statistic, (**c**) Precision, (**d**) Recall and (**e**) AUC. Algorithms whose average rank difference is larger than the critical difference can be considered significantly different with 95% probability. The algorithms that do not differ significantly are connected with a line. Algorithm labels are as follows: *kNN* *k*-Nearest Neighbors classifier, *NB* Naive Bayes classifier, *J48* J48 decision trees, *JRip* jRIP classification rules, *LogR* Logistic regression, *SVM* SVM classifier, *AdaBoost* Boosting, *BagJ48* Bagging, *RF* Random Forests; *BNet* Bayesian Network classifier

The predictions given by the learned models are the probabilities of fire outbreaks: In this context, it is interesting to consider how well the different learning methods model the conditional probabilities, i.e., their calibration. An empirical study comparing predictions by different learning methods and true posterior probabilities (Niculescu-Mizil and Caruana 2005) suggests that while some methods (e.g., Boosting and Naive Bayes) produce quite distorted probability predictions, other methods (e.g., Bagging and Neural networks) predict well calibrated probabilities. If we combine these findings with our results (showing that the performance of Bagging was very good according to all investigated quality measures, except recall), we can conclude that this method is most suitable for predicting the risk of fire outbreaks. In case we want to reduce the number of non-predicted fire outbreaks, Random Forests might be preferable over Bagging. The only drawback of Bagging as well as Random Forests is the fact that the ensemble models are large and very hard to interpret.

Comparing the results for different datasets (Slovenia—S, Coastal—C, Kras—K and Kras without Lidar—L) we can see that the best performing models are learned for the Costal dataset, followed by Slovenia and Kras datasets. Considering the Kras dataset, we can see that some single classifier methods (kNN, Naive Bayes, SVM, Logistic Regression, Baysian Networks) perform better when the LiDAR fire fuel information is not included in the learning data. The introduction of the additional attributes does not affect the decision trees, rules and the ensemble methods.

For the Kras dataset, the choice of the data mining algorithm has a much higher influence on the performance as compared to the use or non-use of the LiDAR attributes. This questions the value of the LiDAR data. Because the LiDAR data used in this study was obtained by using a model learned from training data from only a small area, adding LiDAR data for the entire Kras region or introducing vegetation indices from RS imagery would probably improve the models.

Finally, we compare the results of Bagging, as the best performing method, to those of logistic regression. The later are used in the model currently deployed at the Administration for Civil Protection and Disaster Relief of the, Ministry of Defence of Slovenia. According to the Wilcoxon test, Bagging performs better on all performance metrics and all datasets at the 99% ($p < 0.01$) significance level.

6.2 An example: the Kras model

As already mentioned, the models learned with Bagging of decision trees are, although the most accurate, very hard to interpret. In order to get further insight into the problem domain, we investigate the decision rules learned with JRip, as these are very concise. As an example of what kind of knowledge can be extracted from decision rules, we present the models for the Kras dataset without the LiDAR fire fuel attributes. The rules are presented in Table 2.

From the presented rules, we can conclude that in the Kras region the forest fires mostly occur at the edge of settlements or small villages located at elevations higher than 378 m that are close to railways (relative distance less than 3 km). Often the railway lines do not follow security regulations for railway tidiness, therefore sparks from the railway lines can cause fires in the forests next to them.

Table 2 JRip rules learned from the Kras dataset without LiDAR attributes

if ((distRailways ≤ 2970) **and** (elevation ≥ 378) **and** (percBuiltUp ≥ 0.875)) **then**
  fireOutbreak = YES
else if ((percBuiltUp ≥ 0.875) **and** (distRailways ≤ 1487) **and** (percOver ≤ 2.875) **and**
    (percRiparian ≥ 0.625)) **then**
  fireOutbreak = YES
else if ((percSwMead ≥ 26) **and** (percArable ≥ 0.1875) **and**
    (evapoTranspiration_48 ≥ -0.9)) **then**
  fireOutbreak = YES
else if ((distRailways ≤ 2970) **and** (elevation ≥ 350) **and** (percRiparian ≥ 0.125) **and**
    (distRailways ≥ 1897)) **then**
  fireOutbreak = YES
else
  fireOutbreak = NO
end if

Moreover, forest fires are likely to occur at riparian overgrowth areas and forest hedges of the numerous small villages in the Kras region located nearby railways (relative distance less than 1.5 km). Human activities in the numerous small villages in the Kras region are a known issue and a problem to be solved. The farmers are usually not aware of the potential threat of forest fires that can be easily caused by the intensive agricultural use that involves burning of agricultural residuals on riparian overgrowth areas and forest hedges. The agricultural residuals are represented by large amounts of dry meadow biomass that is easily inflammable. The inflamed large meadow biomass can often cause uncontrollable fires that expand and spread into the nearby forests.

Over the past 30 years, the settlements in the Kras region spread too much into the forest (at a distance of 30–100 m from the forest) increasing the potential threat of forest fires. Some of the settlements in the Kras region, for example Sežana, are already fully surrounded by fire-endangered forests and their further enlargement could take place only in the direction of the forest. The awareness of the potential fire hazard, caused by the more recent increase in the population density near the forest, and the possible consequences from it is still at a very low level. On the other hand, the forests are increasingly used for recreational purposes and high human activity is noted in this region.

Given the expectations that the problem of fire risk in the coming decades will be strengthened, due to climate change, there is a need to consider the potential threat of forest fires in the land use planning and ensure a safe distance from residential areas, which would prevent the spread of forest fires to objects and vice versa. Otherwise, the number of conflict areas between forest and settlements will continue to grow because the future expansion opportunities of the settlements will adjust to the growing demand for residential areas. Replacing the belt of conifers (mostly pine) near the railway lines with deciduous trees could be considered as an alternative that can reduce or temper the fire risk and the potential for large fires, in this part of the Kras region.

Additionally, fires are likely to occur in gardens, fields and swampy meadows when the evapotranspiration in the last 48 h is very high. The fields and gardens located in the western part of the Kras region, which is closer to the sea, have high evapotranspiration

rate, especially in the summer, due to the hot and dry sub-Mediterranean climate, brownfield sites, as well as the vegetation adjusted to this type of climate.

Forest fires in the Kras region are also likely to occur on riparian overgrowth areas and forest hedges at altitudes of 350 m and higher that are very close to railways (relative distance between 1.5 and 3 km). As some of the agricultural areas traditionally used for grazing, mowing, small businesses and farms where abandoned at some small villages, they have become overgrowth and can be also seen as a potential threat.

The rapidly overgrowing of Kras woods started in the middle of the last century, due to the abandonment of grazing and logging in this region. As a result, a mosaic interlace of meadows, pastures and forest slopes was formed under the influence of grazing on an extremely modest layer of soil among bare rocks. This typical Kras ecosystem is distinguished by the dry extensive meadows which represent a precious habitat for exceptionally rich flora and specific fauna. Many of the plant species that compose this flora are typical karst species and some of them are endemic.

The knowledge that can be extracted from the predictive models can be used for better understanding of the causes of forest fires. It can be also used to improve short and long term forecast models specific for Slovenia and especially for the Kras region, which is the region with the highest level of fire risk in Slovenia. Finally, it can be used for identifying the landuse types that are most endangered by forest fires, which can lead to improvement in the (planning of) future landuse.

6.3 Maps of the probability of fire outbreaks and fire danger in Slovenia

An important contribution of our work is the generation of maps of the probability of fire outbreaks and the fire danger in Slovenia using the predictive models obtained in our study. In order to generate geographical maps, we used the models derived from J48 decision trees and fed it with environmental data (static GIS data and dynamic RS and meteorological data for a particular date and time).

We obtained different models for each of the investigated datasets. Then we combined the predictions of the models built using data from the Coastal and Slovenia dataset. These datasets contain the same attributes, but due to the climate differences between these regions we investigated them separately in this study (the regions are presented in Fig. 2). The predictions of the model built using data from the Kras dataset are used separately, since they contain a different set of attributes. Next, we translated the predictive model into Python[4] functions that were later used in the GIS system to visualize the predictions in the form of a map.

For the mapping task, we consider the probabilities of fire occurrence, rather than just a binary answer of whether a fire will occur or not. We consider these probabilities to be an estimate of the risk of a fire outbreak at a specific location at a specific time. In addition, we also estimate the fire danger by using an empirical model (Kobler et al. 2006) obtained by weighting the probability of fire outbreaks with the wind speed.

Finally, for demonstration purposes, we present two maps: a map of the probability of fire outbreaks (Fig. 6) and a map of the fire danger (Fig. 7) for 14 June 2006 at

[4] http://www.python.org/.

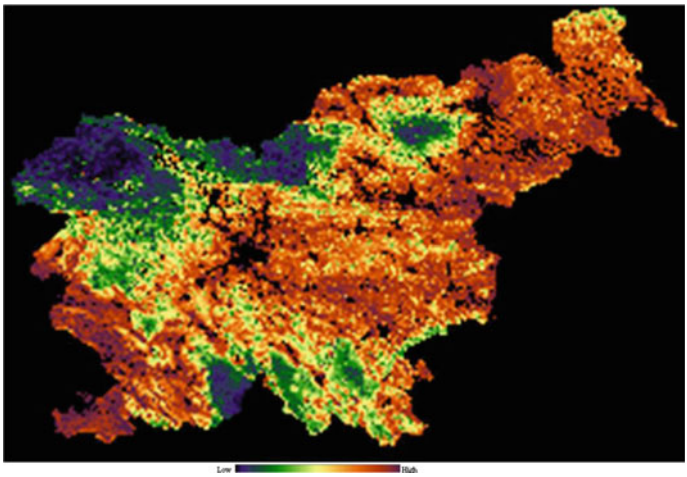

Fig. 6 A map of the probability of fire outbreaks in Slovenia generated by using decision trees, for 14 June 2006 at 12.30 UTC

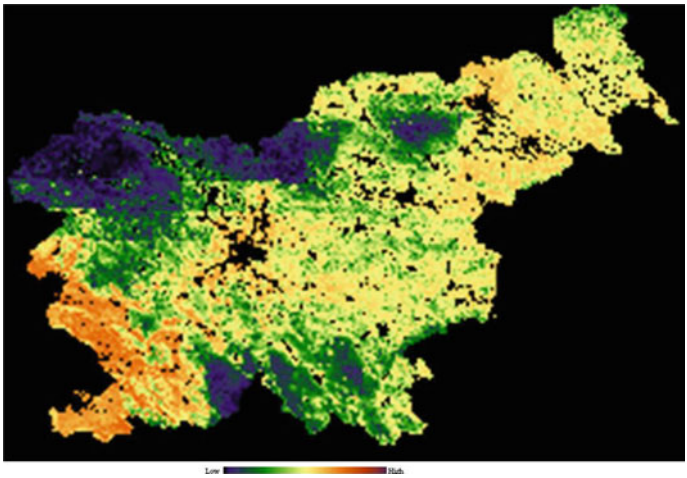

Fig. 7 The fire danger in Slovenia obtained by weighting of the probability of fire outbreaks with the wind speed, for 14 June 2006 at 12.30 UTC

12.30 UTC in Slovenia. The maps were generated by extrapolating the predictions of the model built on the representative sample of fire data and using a set of available environmental data for the whole country.

The maps illustrate the difference in the fire threat in the different regions in Slovenia, stressing the high threat for the Slovenian coast and especially the Kras region. The time interval was selected because of the highest danger of fire in these hours: the sun is highest at noon, and therefore most powerful; the daily maximum temperatures are most often recorded early in the afternoon at the stable weather conditions, and local winds are strongest during this time; therefore, the drying of

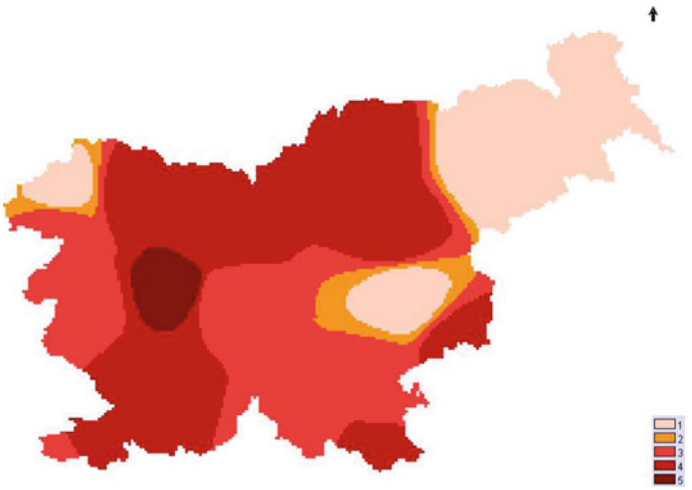

Fig. 8 The model of FWI fire risk danger classes in Slovenia, for 14 June 2006

combustible material is the most intense at that time. In general, the predictions can be made for any desired date and time.

In order to compare the generated maps with maps provided by other services, we present the map of the FWI fire risk danger for the same day (Fig. 8) generated by interpolation of the calculated FWI index using meteorological data from the measuring stations in Slovenia. We generated the FWI map by ourselves instead of using the FWI maps created by the EFFIS system (see Sect. 1.3), because of the coarse spatial resolution and the use of meteorological forecast data from French and German meteorological services in the latter.

Although these maps have different scale from the ones that we generate in this study, we can immediately see the difference in the quality of the maps. The presented FWI map (Fig. 8), besides being of a much coarser resolution, also completely fails to recognize the coastal area as an area with a high fire danger.

The fire danger map built using our models (Fig. 7) predicts moderate to high fire danger in the Coastal part of Slovenia, high fire danger in the Kras region (a subregion of the Coastal region) and low to moderate fire danger in the rest of Slovenia. Since the resolution of the map comes from the resolution of the input data, it is much better compared to the resolution of the map presented in Fig. 8, as well as the prediction maps for Slovenia provided by the EFFIS services. In addition, this kind of maps can be generated for a particular date and time, which makes them suitable for fire outbreak forecasting and prediction of the fire risk. Moreover, the presented methodology can be used within a fire protection system as well as for informing and planing purposes.

7 Conclusions

In this work, we have applied data mining to the task of estimating the risk of fire outbreaks in the natural environment in the country of Slovenia, situated in Central

Europe at the confluence of the Alps and the Mediterranean. Data mining methods were applied to historical data on fires, as well as data on land use from geographical information systems (GIS), remote sensing data and weather forecast data. Predictive models were built that perform well along a number of metrics important for the task at hand: The best models perform significantly better than those built by statistical methods that are currently used in a deployed GIS for supporting activities in response to natural disasters.

As compared to other studies applying statistical and data mining methods to similar problems, our study is unique in several aspects. We use a variety of data at a very fine-grained spatial and temporal resolution, while most studies consider a limited range of data, typically at a much coarser spatio-temporal resolution. We apply a large set of data mining techniques, ranging from techniques that produce simple and understandable models to methods that produce very accurate models that are difficult to interpret: Most studies only consider one or a few learning techniques. Finally, we compare the performance of the data mining methods on a number of related datasets along several relevant performance metrics: Most related studies only consider one dataset and one performance metric, typically accuracy.

While we use a variety of data at a very fine-grained spatial and temporal resolution, this kind of data is readily available for many areas of the world. Data on land use from geographical information systems (GIS) is becoming increasingly available and can also be derived from RS data, the MODIS remote sensing data is also publicly available, and the weather forecast model ALADIN is in widespread use across the world. This means that our methodology is applicable to other areas of the world to produce models suitable for daily predictions and monitoring purposes across large scale areas.

The large set of data mining techniques applied in our study includes a number of data mining techniques that produce a single model and several of these produce understandable models. As an example, we examined in detail a set of classification rules learned from data about the Karst region, the most fire-endangered part of Slovenia. The rules give us insight into the major causes of fires in the natural environment in this region, but also how the severity of the different threats may evolve with climate change. The best predictive performance is achieved by ensemble models, such as Bagging and Random Forests of decision trees.

The performance of the models learned by different data mining methods were evaluated on several metrics. Besides accuracy, these included precision (directly related to the rate of false alarms in predicting fire outbreaks) and recall (directly related to the rate of failed-to-predict fires). The ensemble methods perform best on all metrics, Bagging having better precision and Random Forests having better recall. Both of these approaches are known to produce well-calibrated probability estimates, an important aspect of our task of predicting fire outbreaks.

Several directions for further work remain. We would like to extend our work to include more detailed data on the road network: Besides major roads as considered now, we need to take into account smaller public and forest roads, especially in the areas with high and very high fire risk. We would also like to add further data, such as more precise weather forecasts and day of the week (workday, weekend, etc.), the latter being related to human activity. We would also like to include more historical

data on fire outbreaks over a longer period of time: This could also be achieved by improving the collection of data on fire outbreaks by automated detection of fires from remote sensing data.

Acknowledgements Sašo Džeroski is supported by the Slovenian Research Agency (through the grants P2-0103, J2-0734, and J2-2285), the European Commission (through the grant HEALTH-F4-2008-223451), the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins (operation no.\ OP13.1.1.2.02.005) and the Jozef Stefan International Postgraduate School. Daniela Stojanova is supported by two joint projects between the ARVALIS-Institut du vegetal, Pau, France and the Jožef Stefan Institute, as well as the above mentioned grant J2-2285. Preliminary work on this application, including the data acquisition and pre-processing, was performed within the project "Forecasting GIS model of fire danger in the natural environment" (M1–0032) financed by the Ministry of Education, Science and Sports and the Ministry of Defence of Slovenia.

Appendix A

Table A Description of the attributes and their sources

Attribute	Description	Data Source
fireOutbreak	Yes/No	ACPDR, SFI
*t*0–*t*9	Temperature at *i**8 days ago, where *i*=0–9	MODIS
*npp*0–*npp*9	Net Primary Production (NPP) at *i**8 days ago, where *i*=0–9	MODIS
*nppT*0–*nppT*9	NPP today /t_i, where t_i in *t*0–*t*9	MODIS
*nppSumT*0–*nppSumT*9	NPP from 1.1. until today t_i, where *i* in 0–9	MODIS
nppSum	Sum of NPP from 1.1. until today	MODIS
*nppTavg*0–*nppTavg*9	NPP today /t_i, where *i* in 0–9 and Tavg_i is the average temperature over the last *i**8 days	MODIS
*nppSumTav*0–*nppSumTav*9	Sum NPP from 1.1. until today / Tavg_i, where *i* in 0–9	MODIS
elevation	Median altitude above sea level	SFI
reliefAspect	Median aspect of relief from Digital Relief Model (DMR)	SFI
reliefSlope	Mode of slope of relief from DMR: 0=flat, 1=N, 2=NE,. . ., 8=SW	SFI
distRoads	Average distance to roads (m)	SFI
distSettlements	Average distance to settlements (m)	SFI
distRailways	Average distance to railways (m)	SFI
percArable	Percentage of arable land in a quadrant	MAFF
percTempMead	Percentage of temporary meadows in a quadrant	MAFF
percOthCrop	Percentage of other permanent crops on arable land in a quadrant	MAFF
percVine	Percentage of vineyards in a quadrant	MAFF
percIntOrch	Percentage of intensive orchards in a quadrant	MAFF
percExtOrch	Percentage of extensive orchards in a quadrant	MAFF
percOliveGroves	Percentage of olive groves in a quadrant	MAFF
percOthPermCrop	Percentage of other permanent crops in a quadrant	MAFF
percMeadPast	Percentage of meadows and pastures in a quadrant	MAFF
percSwMead	Percentage of swampy meadow in a quadrant	MAFF

Table A Continued

Attribute	Description	Data Source
percAlpMead	Percentage of alpine meadows in a quadrant	MAFF
percOvergrowth	Percentage of overgrowth areas in a quadrant	MAFF
percForestPlant	Percentage of forest plantations in a quadrant	MAFF
percRiparian	Percentage of riparian overgrowth and forest hedges in a quadrant	MAFF
percUncultAgr	Percentage of uncultivated agriculture land in a quadrant	MAFF
percForestAgr	Percentage of forest trees on agricultural land in a quadrant	MAFF
percForest	Percentage of forest in a quadrant	MAFF
percBuiltUp	Percentage of built-up areas and related surfaces in a quadrant	MAFF
percSwamp	Percentage of swamps in a quadrant	MAFF
percReed	Percentage of reeds in a quadrant	MAFF
percOtherMarsh	Percentage of other marshy areas in a quadrant	MAFF
percDry	Percentage of dried open areas with special vegetation in a quadrant	MAFF
percNoVeg	Percentage of open areas with little or no vegetation in a quadrant	MAFF
percWater	Percentage of water in a quadrant	MAFF
t2m	Temperature at 2 *m*	ALADIN
rh2m	Relative humidity at 2 *m*	ALADIN
windDirection	Wind direction	ALADIN
windSpeed	Wind speed	ALADIN
WindGusts	Wind gusts	ALADIN
precipitationX	Sum of precipitation in the last *X* hours, where *X* in {0, 24, 48, 96, 336}	ALADIN
aseX	Sum of accumulated solar energy in the last *X* hours, where *X* in {0, 24, 48, 96, 336}	ALADIN
evapoTranspirationX	Sum of evapotranspiration in the last *X* hours, where *X* in {0, 24, 48, 96, 336}	ALADIN
transpirationX	Sum of transpiration in the last *X* hours, where *X* in {0, 24, 48, 96, 336}	ALADIN
evaporationX	Sum of evaporation in the last *X* hours, where *X* in {0, 24, 48, 96, 336}	ALADIN
*vegHeightX**	Vegetation height, where *X* in {minimum, maximum, average, standard deviation}	LiDAR
*percVegX**	Surface percentage of vegetation, where *X* in {minimum, maximum, average, standard deviation}	LiDAR
*canopyCoverX**	Canopy Cover - minimum, where *X* in {minimum, maximum, average, standard deviation}	LiDAR
*vegHeightBX**	Vegetation height until B (50, 75, 95, 99)% biomass, where *X* in {minimum, maximum, average, standard deviation}	LiDAR
*vegHeightX**	Maximum vegetation height, where *X* in {minimum, maximum, average, standard deviation}	LiDAR

MAFF Ministry of Slovenia for Agriculture, Forestry and Food, ACPDR Administration for Civil Protection and Disater Relief, SFI Slovenian Forestry Institute; *These attributes only appear in the Kras dataset

References

Agresti A (1996) An introduction to categorical data analysis. John Wiley & Sons, New York

Aha D, Kibler D, Albert MK (1991) Instance-based learning algorithms, Mach Learn 6(1):37–66

Alonzo-Betanzos A, Fontenla-Romeroa O, Guijarro-Berdiñasa B, Hernández-Pereiraa E, Andradeb M, Jiménezc E, Tarsy C (2003) An intelligent system for forest fire risk prediction and fire fighting management in Galicia. Expert Syst Appl 11(25):545–554

Bouckaert R (2005) Bayesian network classifiers in WEKA, Technical report. Department of Computer Science, Waikato University, Hamilton, NZ

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cheng T, Wang J (2008) Integrated spatio-temporal data mining for forest fire prediction. Trans GIS 12(5):591–611

Chu DA, Kaufman YJ, Ichoku C, Remer LA, Tanré D, Holben BN (2002) Validation of MODIS aerosol optical depth retrieval over land. Geophys Res Lett 29(12):1–4

Cohen W (1995) Fast effective rule induction. In: Machine learning: Proceedings of the 12th international conference, Morgan kaufmann, San Francisco, pp 115–123

Connor S (2006) Global warming' will cause more forest fires, droughts and floods'. The Independent, UK, 15 Aug 2006

Cortez P, Morais A (2007) A data mining approach to predict forest fires using meteorological data. In: New trends in artificial intelligence, proceedings of the 13th Portuguese conference on artifical intelligence, Springer, Berlin, pp 512–523

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Džeroski S, Kobler A, Gjorgjioski V, Panov P (2006) Using decision trees to predict forest stand height and canopy cover from LANDSAT and LIDAR data. In: Managing environmental knowledge, EnviroInfo 2006, proceedings of the 20th international conference on informatics for environmental protection, Shaker Verlag, Aachen, pp 125–133

European Commission (2008) Forest fires in Europe 2007, Report No 8. Technical Report, European Commission, Joint Research Centre, Institute for Environment and Sustainability

Felber A, Bartelt P (2003) The use of nearest neighbor method to predict forest fires. In: Proceedings of the 4th international workshop on remote sensing and GIS applications to forest fire management: innovative concepts and methods in fire danger estimation, pp 100–103

Fischer C, Montmerle T, Berre L, Auger L, Stefanescu S (2006) An overview of the variational assimilation in the ALADIN/FRANCE NWP system. Q J R Meteorol Soc 613(1):3477–3492

Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. In: Machine learning: proceedings of the 13th international conference, Morgan Kaufmann, San Francisco, pp 148–156

Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11(1):86–92

Fujii T, Fukuahi T (2005) Laser Remote Sensing. Taylor and Francis Group, Boca Raton

Giglio L, Kendall J, Justice C (1999) Evaluation of global fire detection using simulated AVHRR infrared data. Int J Rem Sens 20:1947–1985

Giglio L, Descloitres J, Justice C, Kaufman Y (2003) An enhanced contextual fire detection algorithm for MODIS. Rem Sens Environ 87:273–282

Holden Z, Morgan P, Evans J (2009) A predictive model of burn severity based on 20-year satellite-inferred burn severity data in a large southwestern US wilderness area. For Ecol Manag 258(11):2399–2406

Hosmer DW, Lemeshow S (1989) Applied Logistic Regression. John Wiley & Sons, New York

Hsu W, Lee M, Zhang J (2002) Image mining: trends and developments. J Intell Inf Syst 19(1):7–23

John GH, Langley P (1995) Estimating continuous distributions in bayesian classifiers. In: Proceedings of the 11th conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Francisco, pp 338–345

Kandola J, Shawe-Taylor J, Cristianini N (2003) Learning semantic similarity. In: Advances in neural information processing systems, vol 15. Bradford Books. Cambridge, pp 657–664

King M, Closs J, Spangler S, Greenstone R (2003) EOS data products handbook. National Aeronautics and Space Administration, Washington, DC

Kobler A (2001) The final report on the results of the research project: A spatial model of forest fire risk (Končno poročilo o rezultatih podprojekta: prostorski model požarne ogroženosti gozdov). Technical report, Slovenian Forestry Institute, Biotechnical Faculty, Department of Forestry, Ljubljana, Slovenia

Kobler A, Ogrinc P, Skok I, Fajfar D, Džeroski S (2006) The final report on the results of the research project: A predictive GIS model of fire risk in the natural environment (Končno porovcilo o rezultatih raziskovalnega projekta; Napovedovalni GIS model požarne ogroženosti naravneğa okolja). Technical report, Slovenian Forestry Institute, Jožef Stefan Institute, Ljubljana, Slovenia

Li Z, Kaufman Y, Ithoku C, Fraser R, Trishchenko A, Giglio L, Jin J, Yu X (2001) A review of AVHRR-based active fire detection algorithm: principles, limitation, and recommendation. In: Ahern F, Goldammer JG, Justice C (eds) Global and regional vegetation fire Monitoring from space: planning and coordinating international effort. SPB Academic Publishing, The Hague

Locatelli B, Imbach P, Molina L, Palacios E (2008) Adaptation of forests and forest management to changing climate with emphasis on forest health: a review of science, policies and practices. In: Proceedings of the 13th Portuguese conference on artificial intelligence, Guimarães, Portugal, p 15

Mack P (1991) The landsat case. (Book reviews: viewing the earth, the social construction of the landsat satellite system). Science 254:314

Markuzon N, Kolitz S (2009) Data driven approach to estimating fire danger from satellite images and weather information. In: 38th IEEE applied imagery pattern recognition workshop, Washington

Mazzoni D, Tong L, Diner D, Li Q, Logan J (2005) Using MISR and MODIS data for detection and analysis of smoke plume injection heights over north america during summer 2004. AGU fall meeting Abstracts, p B853+

Nemenyi P (1963) Distribution-free multiple comparisons. PhD thesis, Princeton University, Princeton, NY, USA

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Machine learning: Proceedings of the 22nd international conference, Morgan Kaufmann, San Francisco, pp 625–632

Preisler K, Chen S, Fujioka F, Benoit J, Westerling A (2008) Wildland fire probabilities estimated from weather model-deduced monthly mean fire danger indices. Int J Wildland Fire 17:305–316

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Fransisco

Ruddell S, Sampson R, Smith M (2007) The role for sustainably managed forests in climate change mitigation. J For 105:314–319

Sabins F (1978) Remote Sensing: Principles and Interpretation. Freeman, San Francisco

Saravanan N, Kumar Siddabattuni VNS, Ramachandran K (2008) A comparative study on classification of features by SVM and PSVM extracted using Morlet wavelet for fault diagnosis of spur bevel gear box. Expert Syst Appl 35:351–1366

Slovenia Forest Service (2005) Information on forest fires in Slovenia in the period 2000–2004. Technical report, Slovenia Forest Service, Ljubljana, Slovenia

Stojanova D, Panov P, Kobler A, Džeroski S, Taškova K (2006) Learning to predict forest fires with different datamining techniques. In: Proceedings of the 9th international multiconference information Society IS 2006, Ljubljana, Slovenia, pp 255–258

Stojanova D, Panov P, Gjorgjioski V, Kobler A, Džeroski S (2010) Estimating vegetation height and canopy cover from remotely sensed data with machine learning. Ecol Inform 5(4):256–266

Swets J (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

Turner JA, Lawson BD (1978) Weather in the Canadian forest fire danger rating system: a user guide to national standards and practices. Technical report. Inf. Rep. BC-X-177. Canadian Forest Service, Pacific Forestry Centre, Victoria, BC, Canada

Vega-Garcia C, Lee B, Woodard P, Titus S (1996) Applying neural network technology to human-caused wildfire occurence prediction. AI Appl 10(3):9–18

Witten I, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann, San Francisco