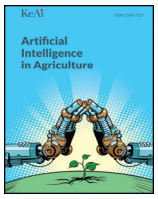
Artificial Intelligence in Agriculture 5 (2021) 24–42

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight

David A. Wood *

DWA Energy Limited, 25 Badgers Oak, Bassingham, Lincoln LN5 9JP, United Kingdom

ARTICLE INFO

Article history:
Received 27 October 2020
Received in revised form 5 January 2021
Accepted 8 January 2021
Available online 16 January 2021

Keywords:
Wildfire feature selection
Non-regression machine learning
Data-matched prediction accuracy
Cumulative absolute error differentials
Predicting highly skewed distributions

ABSTRACT

An optimized data-matching machine learning algorithm is developed to provide high-prediction accuracy of total burned areas for specific wildfire incidents. It is applied to a well-studied forest-fire dataset from Portugal Montesinho Natural Park considering 13 input variables. The total burned area distribution of the 517 burn events in that dataset is highly positively skewed. The model is transparent and avoids regressions and hidden layers. This increases its detailed data mining capabilities. It matches the highest burned-area prediction accuracy achieved for this dataset with a wide range of traditional machine learning algorithms. The two-stage prediction process provides informative feature selection that establishes the relative influences of the input variables on burned-area predictions. Optimizing with mean absolute error (MAE) and root mean square error (RMSE) as separate objective functions provides complementary information with which to data mine each total burned-area incident. Such insight offers potential agricultural, ecological, environmental and forestry benefits by improving the understanding of the key influences associated with each burn event. Data mining the differential trends of cumulative absolute error and squared error also provides detailed insight with which to determine the suitability of each optimized solution to accurately predict burned-areas events of specific types. Such prediction accuracy and insight leads to confidence in how each prediction is derived. It provides knowledge to make appropriate responses and mitigate specific burn incidents, as they occur. Such informed responses should lead to short-term and long-term multi-faceted benefits by helping to prevent certain types of burn incidents being repeated or spread.

1. Introduction

Recent years have been unprecedented in the extent of burned areas of wildfires, particularly in forested regions across the world that increasingly impact substantial population areas. This highlights the urgent need to improve our ability to predict and better understand the factors influencing total burned areas likely to occur in certain conditions (Finney et al., 2013), and real-time, wildfire-spread modelling (De Gennaro et al., 2017). This requires efficient and penetrating data mining of datasets of historical wild-fire incidents. It is essential to identify the key factors influencing the impacts of forest fires in specific locations and climatic zones as global climate conditions seem likely to evolve rapidly (Liu and Wimberly, 2016; Young et al., 2017) and as human land use expands (Montgomery, 2014). Alternative methods continue to emerge to suppress (Houtman et al., 2013), to prevent and to simulate (Rios et al., 2019) wildfires. Although human-caused and lightning-caused ignitions are common features in forest fire initiation (Veraverbeke et al., 2017) a range of meteorological and environmental factors influence their ability to spread and the speed at which they spread. These factors include local climatic conditions, topography and vegetation type. Also, the condition and density of vegetation are fundamental factors that influence the behavior of wildfires and their ultimate wide ranging impacts on surrounding agriculture, ecology, environments and forestry.

The majority of wildfires that spread rapidly occur in windy conditions. Forest wind profiles can be complex, especially in the sub-canopy due to vegetation structure (Moon et al., 2019). Topography, and in particular canyons and junctions between canyons, tend to be wildfire hotspots in hilly or mountainous terrains, where localized high winds fan flames (Raposo et al., 2018; Rodrigues et al., 2019). Intermittent and varying heat flux from wind-driven flames (Tang et al., 2019) and coherent structures within flames impacted by wind conditions (Miller et al., 2017) influence the directions in which wildfires propagate and at what rates. Smouldering and flaming wildfires behave quite differently (Santoso et al., 2019), and the spontaneous transition between the two states is also influenced by complex relationships between the prevailing weather and environmental conditions (Yang et al., 2018).

* Corresponding author.
E-mail address: dw@dwasolutions.com.

https://doi.org/10.1016/j.aiia.2021.01.004

The rate of wildfire spread depends on a wide range of variables combining with wind speeds and directions, displaying complex non-linear interactions. The condition of the fuel bed (McAllister, 2019) and its geometry (McAllister and Finney, 2016) influence the spreading rate of wildfires. Likewise the wind-driven transportation of firebrands (Caton-Kerr et al., 2019; Tohidi and Kaye, 2017; Manzello and Suzuki, 2017) and the nature of the firebrands (Hedayati et al., 2019) impact how wildfires develop and spread in specific regions. The aerated condition of the litter layer beneath the trees and shrubs, particularly its permeability in broadleaf forests (Wang et al., 2019; Fehrmann et al., 2017), is influential because it affects the extent of wind driven convection through the fuel bed (De Gennaro et al., 2017).

This study is organized into the following sections:

- Literature review (Section 2)
- Materials and methods (Section 3)
- Results and discussion (Section 4)

2. Literature review

Forest fire prediction systems, especially the Forest Fire Weather Index (FWI) System developed for applications in Canada using mainly meteorological data (Stocks et al., 1989; Taylor and Alexander, 2006), and the Forest Fire Behavior Prediction (FBP) System developed for applications in China incorporating satellite sensors (Zhang et al., 2011) are widely used. They are suitable and exploited for worldwide applications (Di Giuseppe et al., 2018) as they combine a wide range of relevant input variables. Such systems and their component variables make it possible to compile datasets with historic information recording a wide range of prevailing conditions and geospatial data (Forsell et al., 2009; Alkhatib, 2014) linked to the total burned area resulting from each incident. Such datasets are conducive to machine-learning applications attempting to predict total burned area based on specific conditions, and for data mining to extract useful comparative information relating specific events. Common challenges using such datasets are consolidating data from diverse sources (De Souza et al., 2015) and the positively skewed nature of burned area distributions (Cortez and Morais, 2007). This requires careful feature selection to generate manageable models. In some locations specific variables are known to be highly influential, e.g. vapor pressure deficit and fraction of spruce cover in the case of Alaska's boreal forests (Coffield et al., 2019).

There have been many attempts over recent decades to apply machine learning and simulation methods to assist in the prediction of forest fires (Finney, 1998; Yongzhong et al., 2004) and to evaluate the feasibility of wildfire prediction (Malarz et al., 2002). Additionally, various models have evolved to assess multiple aspects of environmental and ecological change influencing wildfire development (Recknagel, 2001). The machine-learning techniques evaluated include multivariate regression, artificial neural networks (applying both multilayer perceptrons and radial basis functions), support vector machines, decision trees, in some cases supported by genetic programming, random forest, and Markov chain decision process (Garzón et al., 2006; Cortez and Morais, 2007; Li et al., 2011; Sitanggang and Ismail, 2011; Castelli et al., 2015; Mcgregor et al., 2016). Airborne and satellite imagery to provide detailed geospatial, and other information (e.g. surface relative humidity, Peng et al., 2006)), are the focus of some wildfire machine learning studies (Sehgal et al., 2006; Saranya and Hemalatha, 2012; Subramanian and Crowley, 2017, 2018) with some applying fuzzy logic techniques to generate total burned-area predictions (Angayarkkani and Radhakrishnan, 2009).

In recent years, a range of optimizers and machine learning models have been employed to assist in wildfire prediction. These include genetic algorithms coupled with wind field and environmental models (Artes et al., 2016; De Gennaro et al., 2017). Some models distinguishing wind velocity and direction are exploited using domain decomposition of linear relationships within datasets (Sanjuan et al., 2016) and probabilistic analysis (Quill et al., 2019). Zhang et al. (2019) developed a spatial prediction model for forest fire susceptibility employing a convolutional neural network verified with data from historic fire locations in Yunnan Province, China. An adaptive neuro-fuzzy inference system (ANFIS) was applied to classify forest fire hotspot data in central Kalimantan, Indonesia (Wijayanto et al., 2017). Jaafari et al. (2019) evaluated hybrid models combining neuro-fuzzy logic and metaheuristic optimization to predict the probability of wildfires spatially in the Hyrcanian region of Iran.

However, there is much room for progress and improvement in the prediction accuracy of total burned areas associated with specific wildfires. To date, limited research has been conducted to explore the potential of new machine-learning techniques that avoid regressions, hidden layers and focus on transparency, as well as accuracy of individual predictions.

A common characteristic of the machine learning techniques traditionally applied to predict total burned areas of forest fires from data-driven models is that they involve regression, correlation and/or statistical distribution analysis in their calculation processes, often with hidden layers and difficult to penetrate calculations contributing to the prediction of each forest fire event. Such models tend to be insufficiently transparent in their calculations to be conducive to rapid and detailed data mining of the datasets they are modelling. In this study, a recently developed optimized data matching model, the transparent open box (TOB) network (Wood, 2018) is configured and developed specifically to predict and data mine a well-studied forest fire dataset (Cortez and Morais, 2007). This technique, for the first time applied here to assess a forest fire dataset, avoids using correlations, regressions or statistical relationships between its variables. By doing so, it readily reveals the details of each prediction it makes and the closely matched historical fires to each event. This makes the model much more conducive to detailed data mining applications, which are essential to better understand historical forest fire datasets.

3. Materials and methods

3.1. Optimized-data-matching algorithm configured for forest fire data mining

The TOB machine-learning method is now well described and has been implemented with a range of dataset from different sectors (Wood, 2019a). Fig. 1 provides a summary of the TOB workflow configured to predict total burned area associated with forest fires based on forestry, weather and environmental variables. A novel aspect of the algorithm is that it is computed in two stages (1 and 2) and provides two estimates, the second optimized. Comparing the two-stage estimates with training and testing data subsets helps to identify and reject optimized solutions that overfit the underlying data variables.

TOB establishes the best ($Q \leq 10$) data-record matches in a large training data subset for each specific data record in relatively small tuning subsets of data records (between about 100 and 150 data records based on sensitivity analysis). It does so by comparing the sums of the variable squared differences (*VSD*) for all the input variables between specific data records in the tuning subset with all the data records in the larger pool of training subset data records. Eq. (1) describes how weights are applied in the calculation to of the weighted sum of each VSD.

$$\sum VSD_{jk} = \sum_{n=1}^{n=N} VSD(\mathrm{X}n)_{jk} * (Wn) \tag{1}$$

where

N = number of input variables.
Xn = normalized value of variable n.
$j = j^{th}$ tuning-subset data record (from small pool of data records).

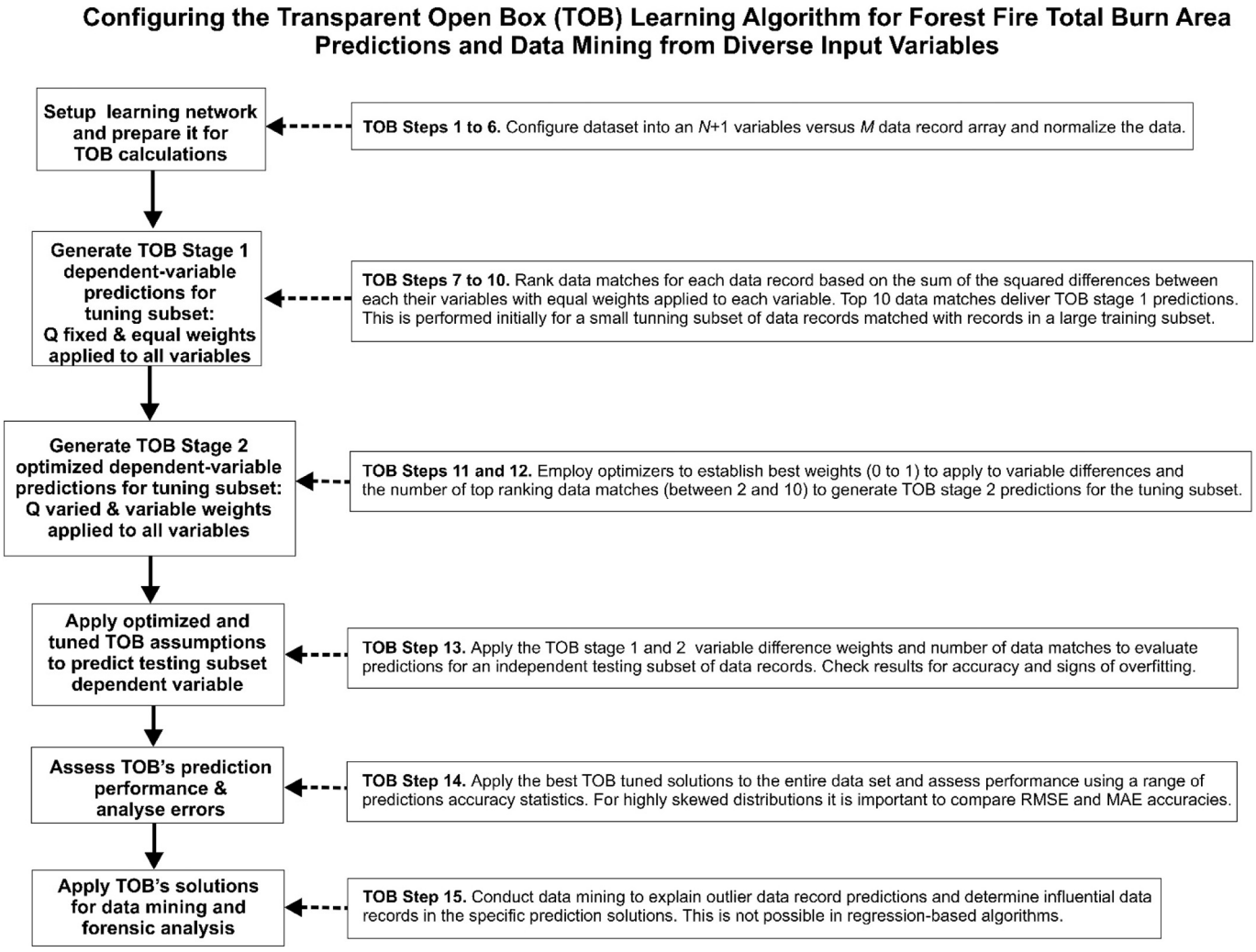


Fig. 1. TOB optimized data matching methodology configured for predicting and data mining highly skewed total burned areas in *forest-fire datasets.*

$k = k^{th}$ training-subset data record (from large pool of data records).

Wn = weight applied to variable n in all data-record comparisons. In TOB Stage 1 Wn is fixed at a constant value (e.g. 0.5) and Q (number of best matching records) is set to 10. In TOB Stage 2 variable weights are allowed to vary such that $0 < Wn \leq 1$ and Q is also allowed to vary ($Q \leq 10$) enabling the optimizer to minimize $\sum_{n=1}^{N} VSD_{jk}$.

Eq. (2) calculates the relative magnitude of $\sum VSE$ for the best matching data records

$$f_{jq} = \sum VSD_{jq} / \left[\sum_{r=1}^{r=Q} \sum VSD_{jr} \right] \tag{2}$$

where

$q = q^{th}$ highest-ranked training-subset data record for the j^{th} tuning-subset data record.

$r = r^{th}$ highest-ranking training-subset data record ($Q \leq 10$) for the j^{th} tuning-subset data record.

f_{jq} = fractional contribution to VSD of q^{th} highest-ranked training-subset data record for the j^{th} tuning-subset data record, constrained such that $\sum_{q=1}^{q=Q} f_{jq} = 1$.

The matching data record with the lowest $\sum VSD_{jk}$ has the lowest calculated f_{jq}. Eq. (3) ensures that the training-subset data record with the lowest f_{jq} value contributes the most to the total burned area prediction for the j^{th} tuning-subset data record.

$$(X_{N+1})_j^{predicted} = \sum_{q=1}^{q=Q} \left[(X_{N+1})_q * \left(1 - f_q\right) \right] \tag{3}$$

where

$N + 1$ = dependent variable (total burned area).

$(X_{N+1})_q$ = total burned area of q^{th} of Q highest-matching training-subset data record.

$(X_{N+1})_j^{predicted}$ = predicted total burned area value for j^{th} tuning-subset data record.

Eqs. (1) to (3) are evaluated separately to calculate TOB Stage 1 (no optimizer; $Q = 10$, Wn = constant) and TOB Stage 2 (optimizer applied; $2 \leq Q \leq 10$; $0 < Wn \leq 1$). Optimum Q and Wn values are located by the optimizers to minimize the objective function. That objective function is usually root-mean-squared error (RMSE) or mean absolute error (MAE) of the dependent variable prediction; in this study total burned area.

3.2. Optimizers suitable for TOB applications

Various functional optimization algorithms could be used to perform the TOB Stage 2 optimization routine. Wood (2019b) has demonstrated that standard Microsoft Excel Solver optimizers (Frontline Solvers, 2020) are adequate, but to extract metaheuristic profiling information (Wood, 2016) and to fully code the optimization process a customized

firefly optimizer and to fully code the optimizer customized firefly optimizer works well.

Small tuning subsets of data records selected to sample the full range of total burned area values provide statistically repeatable total burned area predictions. That repeatability is confirmed using independent testing-subsets of data records involving about one hundred data records not involved in the algorithm-tuning routines. To monitor repeatability and expose overfitting five different cases are run each associated with different selections of tuning and testing subset data records. For each case three different optimizer algorithms were applied and each was run three times. This generated forty-five TOB-Stage-2 optimized tuned solutions to test. As well as applying those solutions to the independent testing subsets it is also informative to evaluate each optimized solution using the entire dataset (all data records). If the TOB stage 2 solutions generate lower prediction errors when applied to all data records than the TOB Stage 1 solutions that verifies that the optimized solutions, found using small tuning subsets are not overfitting the dataset as a whole.

3.3. Measures of burned area prediction accuracy for forest fire analysis

The assessment of burned area prediction benefits from comparing a number of well-established metrics that compute prediction accuracy from different statistical perspectives. In this study the prediction accuracy metrics defined in Eqs. (4) to (13) are calculated and assessed.

In these equations, X_i refers to the measured value and Y_i the TOB predicted value of a data record i in the subset being considered.

Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((X_i)-(Y_i))^2 \tag{4}$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\text{MSE}} \tag{5}$$

RMSE calculated with Eqs. (8) and (9). RMSE is used as the objective function of the TOB Stage 2 algorithm.

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i-Y_i| \tag{6}$$

Percent Deviation between measured and predicted values for data set record i (PDi):

$$PD_i = \frac{X_i-Y_i}{X_i}\text{x } 100 \tag{7}$$

Average Percent Deviation (APD):

$$APD = \frac{\sum_{i=1}^{n}PD_i}{n} \tag{8}$$

APD combines both positive and negative percent deviations (Eq. (10)) and is expressed in percentage terms.

Absolute Average Percent Deviation (AAPD):

$$AAPD = \frac{\sum_{i=1}^{n}|PD_i|}{n} \tag{9}$$

AAPD combines the absolute values of the percent deviations (Eq. (10)) and is also expressed in percentage terms.

Standard Deviation (SD):

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(D_i-Dmean)^2}{n-1}} \tag{10}$$

where

Di is $(X_i - Y_i)$ for each (i^{th}) data record of a dataset; and,

Dimean is the mean of the *Di* values of all the data records in a dataset

$$Dmean = \frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i) \tag{11}$$

Correlation Coefficient (R) between variables X_i and Y_i (on a scale between −1 and + 1).

$$R = \frac{\sum_{i=1}^{n}(Xi-Xmean)\,(Yi-Ymean)}{\sqrt{\sum_{i=1}^{n}(Xi-Xmean)^2\ \sum_{i=1}^{n}(Yi-Ymean)^2}} \tag{12}$$

R is expressed on a scale of −1 to +1.

$$Coefficient\ of\ Determination = R^2 \text{ (on scale between 0 and 1)} \tag{13}$$

3.4. Characterization of a well-studied forest fire dataset from portugal

Table 1 provides details and summary statistics for the 517 forest fire incidents recorded in the Montesinho Natural Park (Portugal) between January 2000 to December 2003 that resulted in 39 human deaths and damaged to about 5% of the forest (Cortez and Morais, 2007). The total burned areas of the incidents in this dataset constitute a highly positive skewed, asymmetrical distribution requiring that metric to be better displayed and evaluated in terms of its natural logarithm (Fig. 2). About 48% of the recorded incidents resulted in small total burned areas of less than 100 m^2.

The data variables recorded in this dataset include spatial, temporal and meteorological factors. The four widely used wildfire indices FFMC, DMC, DC and ISI (defined in Table 1) are included as input variables in this dataset. These indices are components of the Fire Weather Index (Stocks et al., 1989; Taylor and Alexander, 2006). FFMC combines measurements of rainfall, relative humidity, temperature and wind speed to provide an indicator of the degree of moisture in forest floor litter, its likelihood to ignite, and the rate of likely spread of a fire once ignited. DMC combines measurements of rainfall, relative humidity and temperature, while DC combines just rainfall and temperature. DMC and DC are typically used as indicators of prevailing moisture content in the forest floor organic layers. DMC and DC record conditions over much longer periods, 12 past days and 52 past days, respectively, compared to the past 16 h involved in FFMC. ISI involves a correlation between FFMC and wind velocity and is used as an indicator of how quickly a fire is likely to spread.

This dataset is important for the current study because previous studies have published prediction accuracy of the total burned area (BA) applying several regression and/or statistically focused machine learning methods (Cortez and Morais, 2007; Castelli et al., 2015). The method evaluated include:

- Geometric Semantic Genetic Programming Decision Tree (GS-GP)
- Standard Genetic Programming Decision Tree (ST-GP)
- Support Vector Machine polynomial kernel (second degree) (SVM)
- Random forests (RF)
- Multi-layer Perceptron Neural Networks (MLP-NN)
- Radial basis function neural network (RBF-NN)
- Linear regression (LR)

Table 1
Range, mean and units for the Montesinho Natural Park (Portugal) forest fire dataset for 2000 to 2003 (Cortez and Morais, 2007; UCI Machine Learning Repository, 2008) recording 517 separate fire incidents.

Variable code	Description	Unit	Min	P25	P50	Mean	P75	Max
X	East-West Coordinate (1 to 9)	Integer series	1	3	4	4.67	7	9
Y	North-South Coordinate (1 to 9)	Integer series	2	4	4	4.30	5	9
Month	(1 to 12)	Integer series	1	7	8	7.48	9	12
Day	(1 to 7)	Integer series	1	2	5	4.26	6	7
FFMC	Fine Fuel Moisture Code	Index	18.7	90.2	91.6	90.64	92.9	96.2
DMC	Duff Moisture Code	Index	1.1	68.6	108.3	110.87	142.4	291.3
DC	Drought Code	Index	7.9	437.7	664.2	547.94	713.9	860.6
ISI	Initial Spread Index	Index	0.0	6.5	8.4	9.02	10.8	56.1
Temp	Temperature	°C	2.2	15.5	19.3	18.89	22.8	33.3
RH	Relative Humidity	%	15	33	42	44.29	53	100
Wind	Wind Speed	km h^{-1}	0.4	2.7	4	4.02	4.9	9.4
Rain	Rainfall	mm	0	0	0	0.02	0	6.4
BA	Total Burned Area	Ha	0	0	1.52	3.04	7.57	1091.8
BA (Ln)	Natural Logarithm of BA	Ln(Ha)	0	0	0.42	1.11	2.02	7.00

Note: A burned area of zero refers to any event for which the BA is less 100 m^2.

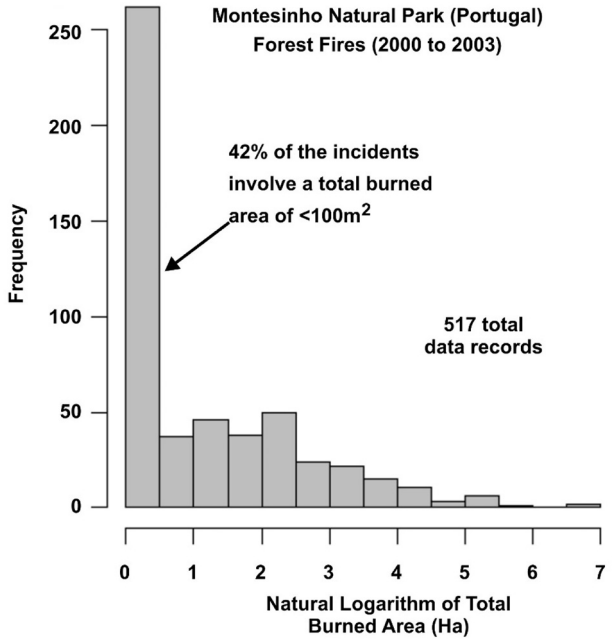


Fig. 2. The positively skewed distribution of total burned areas associated with forest fires in the Montesinho Natural Park (Portugal) from January 2000 to December 2003.

- Multiple regression (MR)
- Isotonic regression (IR)

The correlation matrices between these variables (Table 2) and some of their ratios (Table 3) reveal that they a very poorly correlated with BA. Correlation coefficients with BA are the highest for variables Month (0.1143), DMC (0.0672) and Wind (0.0670). These low correlations explain, to an extent, why correlation-based machine learning algorithm struggle to produce highly accurate predictions for this dataset.

The ratios between input variables also show very low correlations with BA. The correlation coefficients with BA for ratios DC/Temp (0.0760), Wind/RH (0.0737), RH/Month (−0.0725) and Temp/Month (−0.0717) are slightly higher than those of their component variables (compare Tables 1 and 2) suggesting that they should be evaluated as part of feature selection.

Variable relationships between burn area and the input variables are highly non-linear displaying substantial scatter. A key characteristic of the data set is that it is substantially positively skewed (Fig. 3). For this reason, previous prediction analyses of this dataset have elected to express the burn areas as their natural logarithms for the machine learning calculations. The natural logarithm of the total burned area is also used as the dependent variable in the TOB analysis presented here, although the prediction accuracy statistics achieved by those models are expressed in actual burned area (hectare) value terms.

Table 2
Correlation matrix for twelve independent variables associated with 517 forest fire data records with dependent variable (total burned area; BA). For input abbreviations see Table 1.

Correlation coefficient (R)	X	Y	Month	Day	FFMC	DMC	DC	ISI	Temp	RH	Wind	Rain	BA	Rank for R with BA	
X	1.0000	0.5395	−0.0650	−0.0249	−0.0210	−0.0484	−0.0859	0.0062	−0.0513	0.0852	0.0188	0.0654	**0.0620**	5	**X**
Y		1.0000	−0.0663	−0.0055	−0.0463	0.0078	−0.1012	−0.0245	−0.0241	0.0622	−0.0203	0.0332	**0.0388**	9	**Y**
Month			1.0000	−0.0508	0.2915	0.4666	0.8687	0.1866	0.3688	−0.0953	−0.0864	0.0134	**0.1143**	1	**Month**
Day				1.0000	−0.0411	0.0629	0.0001	0.0329	0.0522	0.0922	0.0325	−0.0483	**0.0002**	12	**Day**
FFMC					1.0000	0.3826	0.3305	0.5318	0.4315	−0.3010	−0.0285	0.0567	**0.0468**	8	**FFMC**
DMC						1.0000	0.6822	0.3051	0.4696	0.0738	−0.1053	0.0748	**0.0672**	2	**DMC**
DC							1.0000	0.2292	0.4962	−0.0392	−0.2035	0.0359	**0.0664**	4	**DC**
ISI								1.0000	0.3943	−0.1325	0.1068	0.0677	**−0.0103**	11	**ISI**
Temp									1.0000	−0.5274	−0.2271	0.0695	**0.0535**	7	**Temp**
RH										1.0000	0.0694	0.0998	**−0.0537**	6	**RH**
Wind											1.0000	0.0611	**0.0670**	3	**Wind**
Rain												1.0000	**0.0233**	10	**Rain**
Total Burned Area	0.0620	0.0388	0.1143	0.0002	0.0468	0.0672	0.0664	−0.0103	0.0535	−0.0537	0.0670	0.0233	**1.0000**		**BA**

Table 3
Correlation matrix for selected ratios between the twelve independent variables associated with 517 forest fire data records with dependent variable (total burned area; BA). For input abbreviations see Table 1.

Correlation Coefficient (R)	DCM/Month	DC/Month	Wind/Month	RH/Month	Temp/Month	X/Month	Wind/RH	Wind/Temp	DCM/Wind	DC/Wind	X/Wind	RH/Temp	DCM/Temp	DC/Temp	X/Temp	BA	Selected Variable Ratios
DCM/Month	1.000	0.533	−0.251	−0.235	−0.039	−0.254	−0.144	−0.297	0.541	0.203	0.001	−0.220	0.760	0.154	−0.295	**0.036**	**DCM/Month**
DC/Month		1.000	−0.630	−0.364	−0.439	−0.520	−0.260	−0.418	0.461	0.515	0.061	−0.263	0.533	0.579	−0.362	**0.024**	**DC/Month**
Wind/Month			1.000	0.522	0.474	0.611	0.442	0.518	−0.495	−0.594	−0.303	0.302	−0.317	−0.470	0.340	**−0.030**	**Wind/Month**
RH/Month				1.000	0.322	0.683	−0.212	0.233	−0.254	−0.307	0.074	0.667	−0.197	−0.299	0.429	**−0.073**	**RH/Month**
Temp/Month					1.000	0.508	0.180	−0.232	−0.168	−0.309	0.025	−0.252	−0.473	−0.781	−0.171	**−0.072**	**Temp/Month**
X/Month						1.000	0.000	0.215	−0.298	−0.375	0.315	0.378	−0.336	−0.464	0.640	**−0.016**	**X/Month**
Wind/RH							1.000	0.457	−0.427	−0.467	−0.435	−0.236	−0.227	−0.135	0.016	**0.074**	**Wind/RH**
Wind/Temp								1.000	−0.409	−0.441	−0.301	0.614	0.011	0.266	0.662	**0.104**	**Wind/Temp**
DCM/Wind									1.000	0.861	0.485	−0.202	0.455	0.217	−0.253	**0.003**	**DCM/Wind**
DC/Wind										1.000	0.533	−0.224	0.238	0.331	−0.264	**−0.015**	**DC/Wind**
X/Wind											1.000	0.003	−0.023	−0.021	0.266	**−0.023**	**X/Wind**
RH/Temp												1.000	0.169	0.295	0.731	**−0.002**	**RH/Temp**
DCM/Temp													1.000	0.632	0.004	**0.047**	**DCM/Temp**
DC/Temp														1.000	0.189	**0.076**	**DC/Temp**
X/Temp															1.000	**0.069**	**X/Temp**
Total Burned Area	0.036	0.024	−0.030	−0.073	−0.072	−0.016	0.074	0.104	0.003	−0.015	−0.023	−0.002	0.047	0.076	0.069	1.000	

4. Results and discussion

4.1. TOB Machine Learning analysis to predict burn area of Forest fire dataset

The dataset evaluated by the TOB algorithm includes all twelve independent variables for which measurements are recorded (Table 1) plus the Wind/RH ratio, making thirteen input variables in total. The ratio Wind/RH was selected because it showed the highest correlation coefficient with BA (albeit a very low one) making it second to the "Month" variable in its degree of correlation with BA.

4.2. Data matching feature selection and ranking of Most influential variables

In complex datasets, with non-linear relationships between multiple input variables careful feature selection is essential. The highly skewed distribution of the dependent variable means that just a few data records with extreme values are located in the distribution tail. The TOB method offers three distinct prediction-performance insights to feature selection based on the input variables it includes in its data matching procedure.

1. TOB Stage 1 BA predictions can involve data matching using all available input variables, or in this case between 1 and 13 variables. The squared differences for the input variables selected are evaluated with equal variable error and Q fixed at 10.
2. TOB Stage 1 BA prediction constrained to use the squared differences for all 13 variables and Q fixed at 10. This provides a useful benchmark comparison for prediction insight 1
3. TOB Stage 2 BA optimized predictions allows the number of best data matches and variable weights to vary ($2 \leq Q \leq 10$ and $0 \leq W_N \leq 1$) and considers all 13 variables. However, this prediction is constrained to use only the best record matches selected by prediction insight 1.

Sensitivity analysis comparing the predictions for these helps the feature selection to establish the best combination of variables to use for data matching in TOB Stage 1. Using all data records, variables are sequentially eliminated from influencing the data matching process to establish an order of influence for prediction insight 1. Table 4 compares the sequence of elimination of the input variables by this sensitivity analysis with the ranking of input variables in terms of their correlation coefficient with BA.

Tables 4 and 5 reveal that the 1-Var sensitivity-case model (with DC only used for TOB Stage 1 data matching) generates the lowest RMSE value for BA prediction by TOB Stage 1. It also achieves the lowest RMSE and MAE values for BA by TOB Stage 2. Fig. 3 illustrates the results of the feature selection analysis for TOB data matching for all thirteen sensitivity-case models evaluated in terms of MAE and RMSE. All model structures evaluated achieve credible performance accuracy in terms of RMSE for TOB Stage 2 analysis. Models 1-Var, 2-Var and 3-Var provide superior TOB Stage 2 prediction performance accuracy for BA in terms of MAE for models with RMSE as the objective function (Fig. 3A). Models 1-Var and 3-Var also provide superior TOB Stage 2 prediction performance accuracy for BA in terms of RMSE (Fig. 3B). For these reasons a 1-Var (DC) model is evaluated in more detail for this study.

4.3. RMSE versus MAE as objective functions for predicting highly skewed variable distributions

RMSE is typically more sensitive to predictions of individual data records that generate high errors than MAE. This is because those relatively few points with high prediction error values are squared as part of the RMSE calculation. Most of those high error points are associated with few high total burned area events located in the right-side tail of

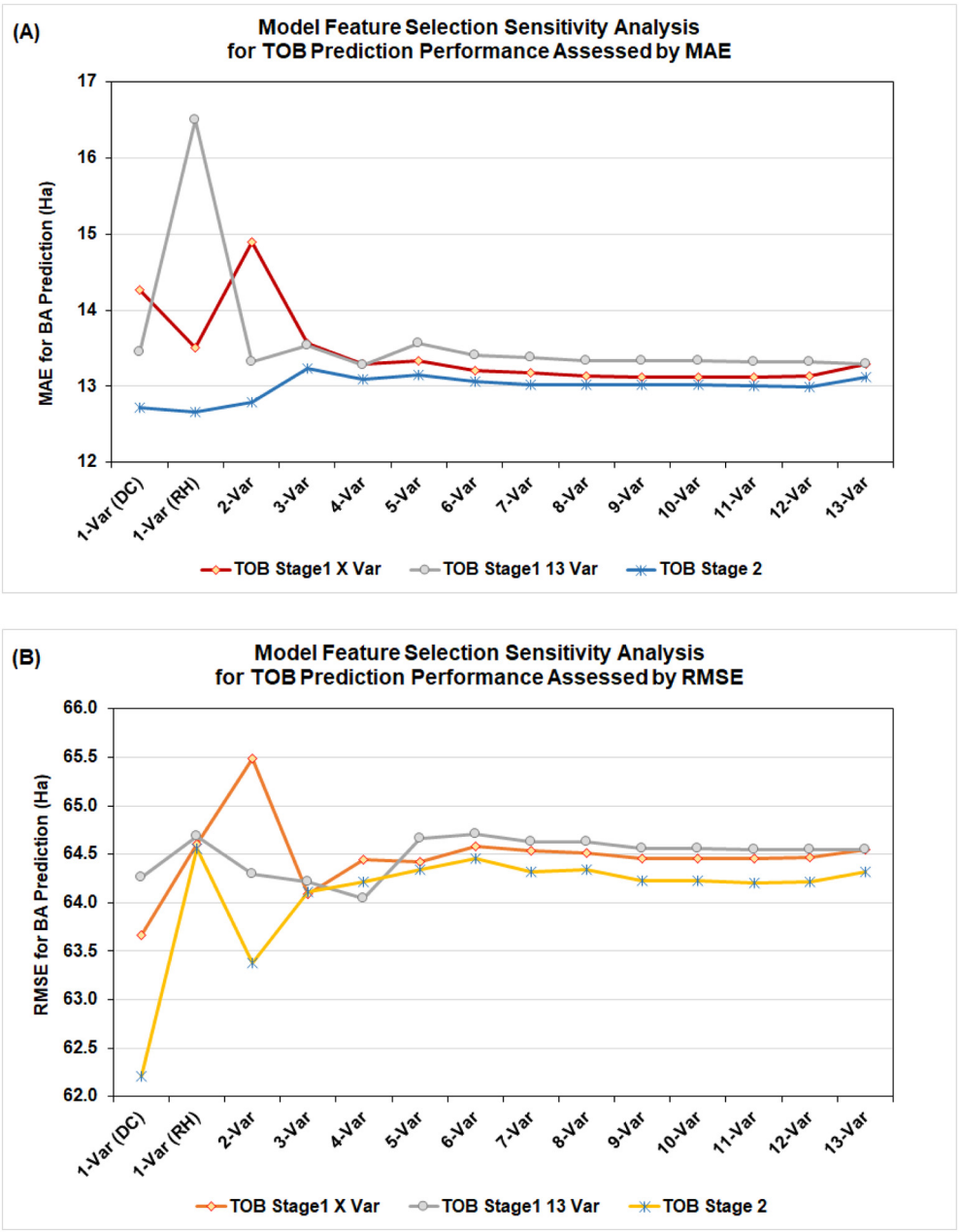


Fig. 3. Comparison of feature selection analysis for TOB data matching for BA predictions: (A) for MAE; (B) for RMSE which is the objective function that is minimized for TOB stage 2 optimization in this sensitivity analysis.

the total burn area distribution in the positively skewed distribution considered here. This fundamental characteristic tends to make RMSE a more effective optimization objective function for such highly skewed datasets than MAE. However, MAE is also a useful objective function to consider as it provides a distinctive insight.

Fig. 4 illustrates the significance of prediction error contributions, in terms of absolute error values, across the highly skewed BA distribution studied. It also highlights the BA prediction improvements achieved by the TOB Stage 2 optimizer. Some70% of the data records are associated with BA prediction errors of ≤5Ha. The TOB stage 2 solution improves on the TOB Stage 1 solution by substantially reducing the absolute errors for a substantial number of data records with prediction errors of less than 5 Ha. It is reassuring that the model is able to predict BA for 70% or so of the data records to an accuracy of 5 Ha. This generates confidence in applying the model for the majority of fires that result in relatively small burn areas. Burn events with prediction errors >12.72 Ha constitute only about ~14% of the data records for this best performing prediction model (Table 4).

4.4. Burned area prediction accuracy comparisons with Machine Learning algorithms

The prediction accuracy of the TOB model compares favourably with that published for this same dataset using a range of regression and statistical based machine learning algorithms applied to all twelve variables with and without feature selection (Cortez and Morais, 2007; Castelli et al., 2015). Cortez and Morais (2007) applied a trial and error feature selection procedure involving the evaluations of four distinct combinations of some of the input variables were considered. Their feature selection analysis identified a high-performing SVM

Table 4
Results of sequential removal of low contributing variables to identify TOB data-matching models with less variables. The earlier an input variable is removed in this sequential analysis generally the less it is contributing to generate high performing data matches. Those variables contributing most to prediction accuracy for specific model structures are also identified and, for the most part, they are the variables with the higher correlations with the BA dependent variable. Although that is not the case for variable Y.

Feature Selection Sensitivity Analysis for TOB Model									
Case	Variables Used for Data Matching	Variable Removed to form Case	Removed Variable's R with BA	Removed Variable's R Rank with BA	Variable Most Valuable for Contribution to BA Prediction	MAE for BA Predictions TOB Stage 1 with Only Matched Variables	RMSE for BA Predictions TOB Stage 1 with Only Matched Variables	MAE for BA Predictions TOB Stage 2 Applied All Variables	RMSE for BA Predictions TOB Stage 2 Applied to All Variables
13-Var	13	None				13.292	64.55	13.124	64.32
12-Var	12	X	0.0620	6	Y then Month	13.128	64.47	12.991	64.22
11-Var	11	ISI	-0.0103	12	Y then Wind	13.117	64.46	13.003	64.20
10-Var	10	FFMC	0.0468	9	Y then Wind	13.119	64.46	13.015	64.23
9-Var	9	Rain	0.0233	11	Y then Wind	13.118	64.46	13.019	64.23
8-Var	8	Wind/RH	0.0737	2	Y then Wind	13.136	64.52	13.016	64.34
7-Var	7	Month	0.1143	1	Y then Wind	13.181	64.53	13.014	64.32
6-Var	6	DMC	0.0672	3	Y then RH	13.212	64.58	13.057	64.45
5-Var	5	Day	0.0002	13	Y then RH	13.329	64.42	13.152	64.34
4-Var	4	Wind	0.0670	4	Y then DC	13.287	64.44	13.088	64.21
3-Var	3	Temp	0.0535	8	DC then RH	13.571	64.09	13.240	64.11
2-Var	2	Y	0.0388	10	DC then RH	14.904	65.49	12.797	63.38
1-Var	1	DC	0.0664	5	RH (remains)	13.502	64.60	12.664	64.56
1-Var	1	RH	-0.0537	7	DC (remains)	14.265	63.67	12.717	62.21

Notes: (1) This analysis helps in the selection of the most promising feature selection for the TOB data matching model that minimizes RMSE (and MAE) for all 517 data records in the dataset. This is achieved by removing the lowest contributing feature from one case to the next. (2) Based on TOB sensitivity analysis data matching with a 1-Var model (DC only) is the most promising for generating low RMSE (and MAE) BA predictions for the dataset. For high performing models, the backgrounds of their superior results are highlighted in this table.

Table 5
Results of sequential removal of low contributing variables to form TOB stage 1 data matching models with less variables: TOB Stage 1 and Stage 2 prediction performance compared in terms of MAE and RMSE. The TOB Stage 2 analysis uses RMSE as its objective function for these model results.

RMSE and MAE prediction accuracies compared for TOB solutions to aid feature selection							
Case	Variables used for data matching	MAE values of solutions optimized with RMSE			RMSE optimization (as TOB objective function)		
		MAE for BA predictions TOB stage 1 with only matched variables	MAE for BA predictions TOB stage 1 with using all variables on matched records	MAE for BA predictions TOB stage 2	RMSE for BA predictions TOB stage 1 with only matched variables	RMSE for BA predictions TOB stage 1 with using all variables on matched records	RMSE for BA predictions TOB stage 2
13-Var	13	13.292	13.292	13.124	64.550	64.550	64.315
12-Var	12	13.128	13.323	12.991	64.473	64.544	64.219
11-Var	11	13.117	13.321	13.003	64.462	64.550	64.200
10-Var	10	13.119	13.341	13.015	64.461	64.561	64.227
9-Var	9	13.118	13.340	13.019	64.461	64.561	64.231
8-Var	8	13.136	13.341	13.016	64.519	64.624	64.336
7-Var	7	13.181	13.384	13.014	64.532	64.626	64.323
6-Var	6	13.212	13.407	13.057	64.582	64.711	64.454
5-Var	5	13.329	13.570	13.152	64.423	64.665	64.336
4-Var	4	13.287	13.275	13.088	64.442	64.047	64.214
3-Var	3	13.571	13.539	13.240	64.085	64.212	64.114
2-Var	2	14.904	13.324	12.797	65.485	64.298	63.379
1-Var (RH)	1 (RH)	13.502	16.501	12.664	64.600	64.689	64.563
1-Var (DC)	1 (DC)	14.265	13.443	12.717	63.667	64.262	62.213

Notes: (1) These results identify the solutions that are best able to minimize RMSE and MAE for all 517 data records in the dataset. (2) Based on TOB data matching sensitivity analysis the 1-Var model (DC only for record matching) is the most promising for generating low RMSE for BA predictions. On the other hand, the 1-Var model (RH only) for record matching is the most promising for generating low MAE for BA predictions for the dataset.

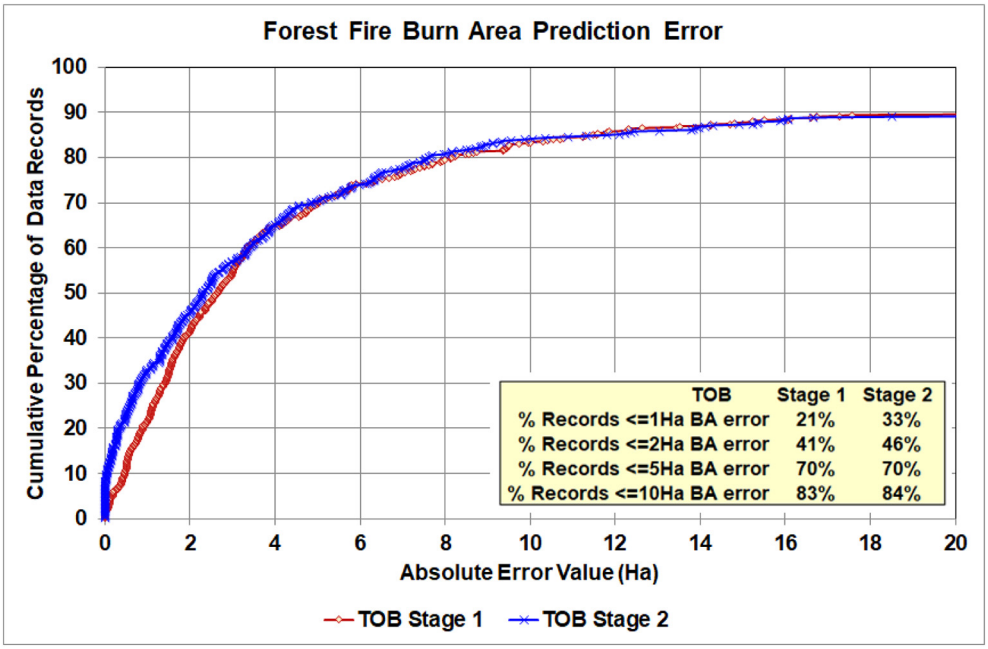


Fig. 4. Cumulative prediction error curve for the best TOB solutions optimized for MAE (MAE = 12.717 Ha; 1-Var (DC) model applied to all 517 data records) distinguishing TOB Stage 1 and Stage 2 solutions.

solution involving just four direct weather inputs (i.e. temperature, rain, relative humidity and wind speed) that performed well when considering just wildfires with small-burned areas.

The optimum TOB model achieves a more accurate best RMSE (62.21 Ha; 1-Var (DC) model), when applied to the full dataset, compared to the best RMSE value for published results for other machine learning algorithms (i.e., RMSE = 63.7 Ha for the SVM model of Cortez and Morais, 2007). That optimum TOB model achieves a similar best MAE of 12.717 Ha applied to the full dataset. This almost matches the best MAE value for published results for other machine learning algorithms (i.e., MAE = 12.71 Ha for their SVM model of Cortez and Morais, 2007). Other machine-learning models evaluated in that study only achieved MAE values ranging from 13.08 Ha to 18.61 Ha.

These prediction accuracy comparisons confirm that the TOB model can generate prediction accuracy for the total burned area for the dataset studied comparable to the best prediction accuracy generated by a range of regression and/or statistical based machine learning algorithms. Moreover, the TOB model has the additional benefit of transparency and forensic access to the components involved in each of its predictions. This additional attribute, which is considered a vital ecological benefit for forest-fire datasets, because it facilitates data mining and yields greater insight to the similarities of specific events with other burn events recorded in the dataset. It also reveals over-fitting as and

when it occurs. These additional attributes and their value are explored in Section 4.5. The regression and/or statistical based machine learning algorithms, for the most part, generate individual predictions in a "black-box" manner, which limits their usefulness in some of these key data mining tasks.

4.5. Employing small tuning and testing subsets to find high performing prediction solutions

A key question to evaluate for highly skewed distributions is, can the tuning of multiple small subsets of data records reveal solutions that can be meaningfully applied to predict the dataset as a whole, avoiding both overfitting and under fitting? The TOB 1-Var (DC) model has been configured to tune and test fifty cases involving tuning subsets of 100 data records (about 19% of the full dataset) and tested with fifty independent testing subsets (100 data records extracted distinctly from the dataset without replacement). This sampling of the dataset means that only 60% of the data records remain as the training data against which the tuned subsets are matched and optimized. Each relatively small tuning and testing subset selected in each case evaluated is spread across the entire BA value range. This is achieved so that it is impossible for the same records to be present in more than one of the testing, training or tuning subsets in each case. Each tuned subset is optimized twice, once with RMSE as the objective function and once with MAE as the objective function.

Tables 6 and 7 display the key prediction performance statistics for the optimized solutions (both MAE and RMSE) found for the 50 cases and evaluated against the 100-data-record testing subsets and the full 517-data-record dataset, respectively. It apparent that there is less dispersion of prediction accuracy for those models optimized with RMSE as the objective function (left-side Tables 6 and 7), compared with those optimized with MAE as the objective function. However, the MAE optimized cases find slightly lower MAE and RMSE best cases. Overall, MAE and RMSE find similar minimum (best) solutions for both MAE and RMSE suggesting that both are viable objective functions for TOB to optimize.

Table 8 lists the variable weights applied to the best TOB Stage 2 solutions found by sensitivity analysis (Table 5) and the tuning subset analysis (Table 7). The TOB Stage 1 data matching is restricted to just one variable (DC) for establishing the ten closest matching data records. Nevertheless, the TOB Stage 2 solutions have then gone on to exploit most of the thirteen input variables in the best-matching data records available to improve their prediction performance.

The best MAE values achieved by the TOB model (12.72 to 13.56 Ha, Table 8) straddle the the mean value achieved by 50 testing subsets (MAE = 12.9 Ha) of the best performing GS-GP decision tree model reported by Castelli et al. (2015). The other machine learning models evaluated in that study achieved mean MAE values ranging from 13.6 to 33.8 Ha.

4.6. Data mining of the high-performing burned area prediction solutions

4.6.1. Data mining to assess data records established as best prediction matches

In data mining terms, the TOB model can go far beyond total burn area predictions with this and other forest fire datasets. It can provide information concerning the broader similarities between conditions and outcomes associated with specific fire incidents alongside each prediction. For the TOB algorithm to data mine the forest fire dataset, sequential sample numbers (identifiers) 1 to 517 have been assigned to each data record in the order in which they are listed in the UCI database file (UCI Machine Learning Repository, 2008).

Some details of the TOB prediction analysis for two representative data records (#144 in Table 9; #369 in Table 10) illustrate the some of the detailed information that TOB makes available for each stage of its prediction analysis. The upper half of Tables 9 and 10 shows the TOB Stage 1 prediction with the BA(Ln) prediction calculations with the values expressed in normalized terms (−1,+1) before being converted to actual BA(Ln) values (0 to 7 scale). The lower half of Tables 9 and 10 shows the TOB Stage 2 BA (Ln) prediction calculations. The last three lines of the table convert the dependent variable values (actual and predicted) into non-logarithmic terms. The TOB Stage 2 solution used for both #144 and #369 is the " Best Solution Found" column of Table 8 (i.e. optimizing all 517 dataset records for 1-Var (DC)).

The sum of the squared error term (SumE) in the fourth column from the right in Tables 9 and 10 combines the differences for each input variable between the records being matched. The individual variable differences can be easily accessed in the TOB model but are not shown in these tables. It is the last three columns on the right side of these tables that contain the key calculations that establish what fractional contribution each matching data record makes to the data record being predicted. Letter formula are shown above these three columns in Tables 9 and 10 to explain these calculations.

There is nothing unusual about the two data records considered (#144 and #369) and the TOB algorithm provides predictions with relatively low errors for both of them. There is an improvement on the initial TOB Stage 1 predictions in both cases by the adjusted weights (defined by the TOB Stage 2 solution applied with Q = 6) for the variable errors used in the TOB Stage 2 predictions. The TOB Stage 1 predictions list the top-ten matching data records and display them in descending rank order. The column second from the right shows the fractional contribution of each matching data record to the BA (Ln) prediction and those contributions are arranged in descending order for the

Table 6
Prediction accuracy achievements for the best-case solutions established using 100-data-record tuning subsets, matched with the optimum TOB model structure (TOB Stage 1 data matching restricted to variable DC) and tested against independent 100-data-record testing subsets.

Prediction accuracies achieved for independent testing subsets						
Statistics for solutions from 50 (100-record) tuned subsets	TOB Stage 2 solution: Optimized for RMSE		TOB stage 1 solution (1-Var (DC only) mean		TOB stage 2 solution: optimized for MAE	
	MAE (Ha)	RMSE (Ha)	MAE (Ha)	RMSE (Ha)	MAE (Ha)	RMSE (Ha)
			5.992	**11.75**		
Minimum	**4.603**	**9.11**	**← Best Case →**		**4.433**	**9.06**
25 Percentile	4.854	9.86	All solutions tested with 100-data record independent testing subsets		5.317	11.57
50 Percentile	5.597	11.95			6.344	12.72
Mean	5.758	11.79			6.037	13.08
75 Percentile	6.378	13.42			6.856	15.27
Maximum	7.576	17.92			7.623	19.89
Standard deviation	0.849	2.097			0.965	2.763

Notes: (1) Assessed using testing subsets with 100 data records not used in the TOB tuning process. (2) Accuracies refer to total burned area (BA) predictions. (3) The tuned TOB solution used is that Initially matched using 1-Var (DC only).

Table 7
Prediction accuracy achievements for the best-case solutions established using 100-data-record tuning subsets, matched with the optimum TOB model structure (TOB Stage 1 data matching restricted to variable DC) and tested against all 517 data records.

Prediction accuracies achieved for all data records						
Statistics for solutions from 50 (100-record) tuned subsets	TOB stage 2 solution: optimized for RMSE		TOB stage 1 solution:		TOB Stage 2 Solution: Optimized for MAE	
	MAE (Ha)	RMSE (Ha)	MAE (Ha)	RMSE (Ha)	MAE (Ha)	RMSE (Ha)
			14.265	**63.67**		
Minimum	**13.086**	**63.26**	**← Best Cases →**		**13.026**	**61.88**
25 Percentile	13.288	64.35	All solutions applied to all data records in the ful dataset		13.491	62.90
50 Percentile	13.407	64.49			13.578	64.61
Mean	13.447	64.33			13.697	64.08
75 Percentile	13.561	64.57			13.770	64.74
Maximum	14.153	65.25			15.848	70.88
Standard deviation	0.239	0.502			0.444	1.435

Notes: (1) Assessed using all 517 data records in the studied dataset. (2) Accuracies refer to total burned area (BA) predictions. (3) The tuned TOB solution used is that Initially matched using 1-Var (DC only).

Table 8
TOB Stage 1 and Stage 2 solutions achieving the highest BA prediction accuracy. Note that TOB Stage 1 itself achieves high prediction accuracy.

Details of high-performing TOB solutions						
	For data matching TOB stage 1 solution	Low RMSE achieved (Evol 2_6)	Low MAE achieved (Evol2_24)	RMSE optimized for all 517 records	Low MAE achieved (GRGM4_10)	Lowest RMSE achieved (EvolM4_15)
MAE (Ha)	14.265	13.098	13.086	12.717	13.026	13.561
RMSE (Ha)	63.667	63.25619	63.539	62.213	63.019	61.877
Q	10	10	10	6	9	4
	Only DC Used	Best Solutions Optimized for RMSE		Best Solution Found	Best Solutions Optimized for MAE	
X Wt	0	0	0.0025950	0.0435660	0.0321675	0.0846351
Y Wt	0	0.921672	0.9260039	0.7865996	0.9999999	0.6746566
Month Wt	0	0.554625	0.4137009	0.0036408	0.1949935	0.2615751
Day Wt	0	0.040349	0.0147467	0.1553836	0.0588744	0.7062877
FFMC Wt	0	0.535292	0.4969708	1	0.3405357	0.6766137
DMC Wt	0	0.142199	0	0.0000491	0	0
DC Wt	0.5	0.392132	0.3890379	0	0.6387197	0.4658374
ISI Wt	0	0.666055	0.6605516	0.3231543	0.1588633	0.5557298
Temp Wt	0	0.030889	0	0.0008815	0.0141902	0.0088463
RH Wt	0	0.629962	0.3742943	0.4107742	0.9999999	0.7566496
Wind Wt	0	0.379060	0.4062201	0.1439708	0.1378218	0.0113238
Rain Wt	0	0.522364	0.5894075	1	0.9671547	0.7104992
Wind/RH Wt	0	0	0.0040449	0.0749174	0.0425287	0.4493523

Notes: (1) Assessed using all 517 data records in the studied dataset. (2) Accuracies refer to total burned area (BA) predictions. (3) 1-Var (DC only) solutions derived by TOB tuning subsets with just 100 data records.

TOB Stage 1 calculation. For data record #144 the top three matching records contribute ~77% to the BA (Ln) TOB Stage 1 prediction. This changes to ~63% for the three matching solutions with the lowest squared errors contributing to the BA (Ln) TOB Stage 2 prediction. On the other hand, for data record #369 the top three matching records contribute ~91% to the BA (Ln) TOB Stage 1 prediction. This decreases to 88% for the BA (Ln) TOB Stage 2 prediction.

4.6.2. Identifying the Most influential data Records for Specific Prediction Solutions

It is insightful to conduct comparative analysis of the cumulative MAE and RMSE solutions for specific high-performing prediction solutions, such as those displayed in Table 8. For the data record sequence 1 to 517 (arranged in ascending order of BA values) each MAE (or RMSE) absolute error is progressively added to the previous cumulative value through the BA ranked (lowest to highest) sequence. The "Best Solution Found" (Table 8) is used as the benchmark and the cumulative values for that solution are subtracted from the cumulative values of each of the other high-performing solutions identified. The cumulative solution differentials are shown in Fig. 7 in terms of MAE for three of the solutions displayed in Table 8.

Table 9
Characteristics of data records that are difficult to fit for the high-performing 1-Var (DC) TOB solutions, or substantially influence those fits.

Problematic data records for high-performing TOB solutions				
Ascending BA value sequence number	Dataset data record number	DC	BA	BA (Ln)
21	21	692.6	0	0
94	94	601.4	0	0
104	104	674.4	0	0
197	399	715.1	0	0
213	433	698.6	0	0
287	154	692.3	2.46	0.90
308	164	674.4	2.95	1.08
497	228	480.8	59.30	4.08
502	230	480.8	72.30	4.28
509	235	674.4	155.88	5.05
517	239	674.4	1091.84	7.00
Some Influential Matching Records Selected for Problematic Predictions				
507	233	692.6	104.39	4.65
513	237	674.4	201.95	5.31
512	236	601.4	197.48	5.29

Note: The listed data records are identified as those leading to increased prediction errors when selected as high matches for just a few data records in the high-performing TOB solutions.

Solution EvolM4_15 shows relatively poor MAE prediction performance (Fig. 7; Table 8) and its differential MAE increases progressively across the record sequence relative to the benchmark "best" solution. Solutions Evol_6 (optimized for RMSE) and GRGM4_10 (optimized for MAE) show better MAE prediction performance overall but their performances are subtly different. Solution Evol2_6 substantially outperforms GRGM4_10 over the lower half of the fire incident sequence (mostly small-burned-area fires). It is only the poorer performance in predicting the last few data records (highest BA values) in the sequence that leads solution Evol2_6 ending up with a lower MAE (13.098) than solution GRGM4_10 (13.026). Both of these solutions show jumps at the same points in their cumulative solution differentials relative to the benchmark solutions. Specifically, they predict data records in sequence positions 21, 94, 197, 497, 502, 513 and 517 with substantially higher errors than the benchmark solution. It is, particularly, the predictions of data records 497, 502, 513 and 517 that have the greatest influence on the MAE and RMSE of all these solutions. The predictions that the benchmark solution makes for those data records are themselves not that good but they are better than for most of the other high-performing solutions (Tables 9 and 10).

The cumulative solution differentials are shown in Fig. 8 for squared errors (influencing RMSE) for three of the solutions displayed in Table 8.

Solution EvolM4_15 shows poor squared-error prediction performance over most of the latter 80% of the data record sequence (Fig. 8). However, by managing to predict the data record in the highest sequence position (517, data record 239 with the highest BA value) substantially more accurately than the other solutions, it ends up with the lowest RMSE overall, even better than that of the benchmark. Solution EvolM4_15's success hinges entirely on its prediction performance of the largest burned area data record in the sequence. In the opposite sense, the fate of solution Evol2_6 for its squared-error cumulative differentials (RMSE) is also determined adversely by data record #517. Over most of the sequence, up to record #502, solution Evol 2_6 outperforms the benchmark and solution GRGM4_15 (displaying a negative cumulative differential). However, the relatively poorer prediction performance of solution Evol2_6 for data records 497 to 517 (particularly #517) compared to the benchmark solution and GRGM4_15 results in it ending up with a significantly higher RMSE than the other solutions displayed in Fig. 8. These trends suggest that for the earlier part of the sequence (small total burned areas) solution Evol2_6 would likely provide the best BA predictions (better than the benchmark solution). On the other hand, for the sequence above data record #486 (largest burned area events) the benchmark solution or solution GRGM4_15 would provide better BA predictions.

It is worthwhile distinguishing some of the characteristics of the problematic (difficult to fit for the TOB 1-Var DC solutions) data records. Table 9 does this and distinguishes their sequence order numbers in the ranked BA distribution from their actual data record numbers (as defined by the order of data records in the dataset repository).

Table 10 explains why some of the problematic data records, those responsible for substantial jumps in the trends of Figs. 7 and 8, are predicted quite differently by the benchmark solution and solution Evol2_6 (Table 8).

For data record #21 and #94 (Tables 10 and 11) it is the lower Q value of the benchmark solution (Q = 6) that leads to it not including high BA value data records #233 and #236 (Table 11) in its predictions. This results in a substantially lower absolute prediction error for the benchmark solution for these data records. The problem for the DC-based record matching is data records #21 and #94 have high DC values but zero BA values. This causes data records #233 (and #238) and #236, with very high BA values, to be ranked in the top ten data matches against data records #21 and #94, respectively. The lower Q value of the benchmark solution works against it in the prediction of data record #104 (Table 10, measured BA value =0) causing it to assign a higher weight that solution Evol2_6 to matching record #235 (high BA value). The BA predictions for data records #228 and #230 (high measured BA values) shown in Tables 10 and 11 are also impacted by favourably by the Q value of the benchmark solution. The DC data matching leads to data records #6 and #93 (BA = 0) to be selected as top-ten matching data records (ranks 7 and 8). These are both excluded by the benchmark solution in its forecast due to the Q = 6 constraint leading to much lower absolute BA prediction errors.

Data record #235 is accurately by all the solutions considered in Tables 10 and 11 but solution GRGM4_10 generates the most accurate solution. The involvement of data record #239 in the top-ten data records contributes significantly to the BA predictions of data record #235 by these solutions. On the other hand, the predictions of data record 239 (with the highest measured BA value) are not good for any of the solutions considered in Tables 10 and 11, despite them including data records #235 and #237 in their top-ten matches. Although the measured BA values of #235 and #237 (Table 9) are high, they are five-times lower than the BA value of #239 when compared on a non-logarithmic scale. Although the benchmark solution does not predict #239 accurately it does so better than the other solutions considered.

5. Discussion

The results presented demonstrate that optimized data matching has significant benefits for predicting total burned areas of forest fires and for data mining the associated datasets. Indeed, the transparency achieved significantly exceeds what is achievable with regression-based machine-learning methods, and this can provide beneficial insights for agricultural, ecological, environmental and forestry datasets. In particular, the transparency provided by the predictions of the TOB algorithm provide insights that are ecologically significant regarding the pros and cons of particular high-performing prediction solutions. It clearly identifies the most influential combination of data records (historical burn events) contributing to each prediction. However, the analysis presented raises some important issues related to highly positively skewed and generally asymmetrical distributions that are typically associated with the burned areas of data sets compiling historical forest fire incidents in different geographic areas. It is difficult to develop machine learning solutions that accurately predict BA for the bulk of the data records (small to modest total burned area incidents) and, at the same time, provide accurate BA prediction for those few burn events located in the right-side tail of the distribution (largest total burned area incidents).

The feature selection sensitivity analysis process described for TOB in this study also has significant potential for other forest fire datasets. Its ability to identify the most and least influential input variables for data matching provides useful information (Table 5 and Fig. 3). Ranking input features in this way typically complements the information that is available by comparison of correlation coefficients or generated by multi-variate analysis such as factor analysis and principal component analysis. This approach can help focus attention on which variables are influential and how the information they provide could be refined, complemented or improved. It can also help identify and assess the benefits of potential future collection of new input variables that better describe the characteristics of specific forests or forests in certain climatic zones. All these benefits should contribute positively to the fundamental ecological requirements of modern wildfire management decision support systems (Martell, 2015).

The results and analysis presented justify the combined use of prediction accuracy metrics MAE and RMSE to root out high-performing solutions. RMSE as an objective function focuses more on the right-side tail of the BA distribution. On the other hand, MAE can reveal important information about the absolute prediction errors in the parts of the distribution containing most data records. Optimizing models separately using both MAE and RMSE as objective functions minimizes errors and provides more ecological insight to those predictions.

Table 10
Prediction comparisons for some of the problematic data records associated with jumps in the trends of Figs. 7 and 8. The cells with shaded background distinguish the matched data records that dominate the predictions differently for the two solutions compared (benchmark solution and solution Evol2_6 optimized for RMSE).

BA (Ln) Prediction Comparisons for Problematic Data Records (RMSE Optimized)												
Record being Predicted		**21 (sq#=21)**			**94 (sq#=94)**			**104 (sq#=104)**			**399 (sq#=197)**	
Measured BA (Ln)	**0.000**			**0.000**			**0.000**			**0.000**		
All 517 Optimum BA (Ln)			**1.856**			**0.331**			**4.745**			**0.858**
Evol2_6 BA (Ln)		**3.294**			**4.111**			**4.086**			**3.635**	
Absolute Error (BA Ln)		**3.294**	**1.856**		**4.111**	**0.331**		**4.086**	**4.745**		**3.635**	**0.858**
Squared Error (BA Ln)		**10.852**	**3.446**		**16.897**	**0.110**		**16.695**	**22.515**		**13.217**	**0.736**
Rank Stage 1	**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**	
1	71	0.0010	0.0365	81	0.0103	0.0565	164	0.0391	0.0578	414	0.0493	0.2306
2	140	0.2683	0.2795	26	0.0314	0.2035	235	0.2221	0.3674	461	0.0305	0.1868
3	186	0.0223	0.1694	99	0.0130	0.0863	237	0.0502	0.0869	462	0.0206	0.1103
4	199	0.0055	0.0983	100	0.0410	0.2014	239	0.3348	0.3146	459	0.0206	0.1103
5	201	0.0206	0.1366	101	0.0410	0.2014	129	0.0864	0.1118	425	0.0278	0.1344
6	208	0.0242	0.2798	173	0.0410	0.2509	204	0.0544	0.0614	439	0.0397	0.2275
7	233	0.6412	0	236	0.7607	0	377	0.0341	0	378	0.6454	0
8	238	0.0066	0	432	0.0179	0	437	0.0397	0	306	0.0523	0
9	31	0.0065	0	490	0.0172	0	442	0.0441	0	307	0.0451	0
10	46	0.0039	0	491	0.0264	0	428	0.0950	0	308	0.0686	0
		Q=10	**Q=6**		**Q=10**	**Q=6**		**Q=10**	**Q=6**		**Q=10**	**Q=6**
Record being Predicted		**228 (sq#=497)**			**230 (sq#=502)**			**235 (sq#=509)**			**239 (sq#=517)**	
Measured BA (Ln)	**4.083**			**4.281**			**5.049**			**6.996**		
All 517 Optimum BA (Ln)			**3.391**			**3.197**			**5.218**			**4.171**
Evol2_6 BA (Ln)		**1.502**			**2.114**			**4.452**			**3.600**	
Absolute Error (BA Ln)		**2.581**	**0.692**		**2.167**	**1.084**		**0.597**	**0.169**		**3.396**	**2.824**
Squared Error (BA Ln)		**6.661**	**0.479**		**4.696**	**1.175**		**0.357**	**0.028**		**11.532**	**7.977**
Rank Stage 1	**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**	
1	230	0.3404	0.7456	463	0.0116	0.0229	164	0.0310	0.0378	164	0.0322	0.0444
2	288	0.0251	0.0496	228	0.4957	0.7473	104	0.1118	0.1414	235	0.6370	0.7691
3	289	0.0251	0.0529	289	0.0365	0.0529	237	0.0159	0.0187	237	0.0167	0.0191
4	290	0.0249	0.0526	290	0.0353	0.0512	239	0.5826	0.7091	374	0.0468	0.0637
5	291	0.0322	0.0543	291	0.0746	0.0754	129	0.0500	0.0496	129	0.0588	0.0521
6	292	0.0202	0.0450	292	0.0327	0.0503	204	0.0560	0.0435	204	0.0576	0.0515
7	6	0.3301	0	6	0.1425	0	377	0.0243	0	377	0.0262	0
8	93	0.1608	0	93	0.1109	0	437	0.0379	0	437	0.0362	0
9	63	0.0095	0	63	0.0135	0	442	0.0456	0	442	0.0431	0
10	120	0.0317	0	120	0.0467	0	428	0.0448	0	428	0.0455	0
		Q=10	**Q=6**		**Q=10**	**Q=6**		**Q=10**	**Q=6**		**Q=10**	**Q=6**

Note: The fractions of two high-performing solution are shown for each case: the first is for solution Evol2_6 (RMSE Optimized); the second is for the benchmark optimum solution.

Table 11
Prediction comparisons for some of the problematic data records associated with jumps in the trends of Figs. 7 and 8 The cells with shaded background distinguish the matched data records that dominate the predictions differently for the two solutions compared (benchmark solution and solution GRGM4_10 optimized for MAE).

BA (Ln) Prediction Comparisons for Problematic Data Records (MAE Optimized)									
Record being Predicted	**21 (sq#=21)**			**94 (sq#=94)**			**228 (sq#=497)**		
Measured BA (Ln)	**0.000**			**0.000**			**4.083**		
All 517 Optimum BA (Ln)			**1.856**			**0.331**			**3.291**
GRGM4_10 BA (Ln)		**4.046**			**4.757**			**1.003**	
Absolute Error (BA Ln)		**4.046**	**1.856**		**4.757**	**0.331**		**3.079**	**0.792**
Squared Error (BA Ln)		**16.370**	**3.446**		**22.634**	**0.110**		**9.482**	**0.627**
Rank Stage 1	**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**	
1	71	0.0021	0.0365	81	0.0054	0.0565	230	0.2221	0.7456
2	140	0.0723	0.2795	26	0.0157	0.2035	288	0.0222	0.0496
3	186	0.0295	0.1694	99	0.0053	0.0863	289	0.0275	0.0529
4	199	0.0129	0.0983	100	0.0199	0.2014	290	0.0271	0.0526
5	201	0.0246	0.1366	101	0.0199	0.2014	291	0.0258	0.0543
6	208	0.0438	0.2798	173	0.0227	0.2509	292	0.0240	0.0450
7	233	0.7919	0	236	0.8910	0	6	0.3504	0
8	238	0.0105	0	432	0.0091	0	93	0.2933	0
9	31	0.0125	0	490	0.0110	0	63	0.0077	0
10	46	0	0	491	0	0	120	0	0
		Q=9	**Q=6**		**Q=9**	**Q=6**		**Q=9**	**Q=6**
Record being Predicted	**230 (sq#=502)**			**235 (sq#=509)**			**239 (sq#=517)**		
Measured BA (Ln)	**4.281**			**5.049**			**6.996**		
All 517 Optimum BA (Ln)			**3.197**			**5.218**			**4.171**
GRGM4_10 BA (Ln)		**1.882**			**4.950**			**3.859**	
Absolute Error (BA Ln)		**2.399**	**1.084**		**0.099**	**0.169**		**3.136**	**2.824**
Squared Error (BA Ln)		**5.754**	**1.175**		**0.010**	**0.028**		**9.836**	**7.977**
Rank Stage 1	**Matches**	**Fractions**		**Matches**	**Fractions**		**Matches**	**Fractions**	
1	463	0.0136	0.0229	164	0.0197	0.0378	164	0.0228	0.0444
2	228	0.4290	0.7473	104	0.1137	0.1414	235	0.7030	0.7691
3	289	0.0476	0.0529	237	0.0130	0.0187	237	0.0136	0.0191
4	290	0.0449	0.0512	239	0.6687	0.7091	374	0.0510	0.0637
5	291	0.0826	0.0754	129	0.0412	0.0496	129	0.0464	0.0521
6	292	0.0499	0.0503	204	0.0494	0.0435	204	0.0594	0.0515
7	6	0.1486	0	377	0.0172	0	377	0.0197	0
8	93	0.1694	0	437	0.0350	0	437	0.0385	0
9	63	0.0145	0	442	0.0420	0	442	0.0455	0
10	120	0	0	428	0	0	428	0	0
		Q=9	**Q=6**		**Q=9**	**Q=6**		**Q=9**	**Q=6**

Note: The fractions of two high-performing solution are shown for each case: the first is for solution GRGM4_10 (MAE Optimized); the second is for the benchmark optimum solution.

Many of the subsets used to establish the mean values for the regression-based and statistically-based machine learning models evaluated in multi-case studies are likely to have involved a degree of overfitting. However, evaluating their BA prediction performance with relatively small testing subsets of data records is unable to reveal this due to a lack of transparency with repsect to each data record prediction. On the other hand, the two-stage prediction process involved in the TOB models clearly reveals those cases with solutions generated from small tuning and testing subset samples for which some degree of overfitting has occurred. These are the cases where the TOB stage 2

Dependent Variable Prediction Value Calculated for Data Record # 144

Rank of Top-Matching Data Records	Top-ranking Matched Records in Training Subset	Dependent Variable Normalized Value (BA Ln)	Sum of Weighted Squared Errors (SumE)	Relative Magnitude of Contribution to Prediction	Relative Contribution fraction	Contributions to Prediction Components
TOB Stage 1 Equal Weights & Q=10		W_N =0.5	**TOB Stage 1 Prediction for Test Record # 144**			
Testing Subset Record Case #1A:	**144**	**-0.8466**	**SumE**	**Y = Total (X)/ Sum E**	**F= Y/ Total (Z)**	**Predicted Value F*BA(Ln)**
1 (1st ranking match)	400	-1.0000	5.391E-06	1835460	0.2974	-0.2974
2	401	-0.3668	5.391E-06	1835460	0.2974	-0.1091
3	301	-1.0000	8.912E-06	1110340	0.1799	-0.1799
4	302	-0.5687	8.912E-06	1110340	0.1799	-0.1023
5	298	-1.0000	8.321E-05	118926	0.0193	-0.0193
6	297	-0.8165	8.321E-05	118926	0.0193	-0.0157
7	224	0.0401	5.088E-04	19450	0.0032	0.0001
8	139	-0.9121	8.043E-04	12303	0.0020	-0.0018
9	475	-0.3124	1.145E-03	8645	0.0014	-0.0004
10 (10th rankingMatch)	474	0.2200	5.253E-03	1884	0.0003	0.0001
			9.896E+00	**6171734**	**1.0000**	**-0.7258**
The top three ranking matches contribute 77% to the TOB Stage 1 predicted value			**Total (X)**	**Total (Z)**	**Sum of F**	**Normalized Prediction = Sum (F*BA Ln)**
					Min Dependent Variable Actual Value (BA Ln):	0.0000
					Min Dependent Variable Actual Value (BA Ln):	6.9956
					Stage 1 Provisional Prediction of Dependent Variable (BA Ln) Value for Data Record: #144	**0.9592**
					Actual Recorded Dependent Variable Value (BA Ln) for Data Record: #144	0.5365
					Difference between Actual and TOB Stage 1 Predicted BA (Ln) Value:	-0.4227
TOB Stage 2 with Optimized Weights						
Stage 2 TOB Optimized Q = 6		**BA (Ln)**	**SumE**	**Y = Total (X)/ Sum E**	**F= Y/ Total (Z)**	**Predicted Value F*BA(Ln)**
2	400	-1.0000	1.017E+00	6.0624	0.1486	-0.1486
2	401	-0.3668	1.017E+00	6.0624	0.1486	-0.0545
5	301	-1.0000	1.276E+00	4.8317	0.1184	-0.1184
4	302	-0.5687	1.126E+00	5.4753	0.1342	-0.0763
6	298	-1.0000	1.273E+00	4.8404	0.1186	-0.1186
1 (best match)	297	-0.8165	4.554E-01	13.5321	0.3316	-0.2708
N/A	224	0.0401	0.000E+00	0.0000	0.0000	0.0000
N/A	139	-0.9121	0.000E+00	0.0000	0.0000	0.0000
N/A	475	-0.3124	0.000E+00	0.0000	0.0000	0.0000
N/A	474	0.2200	0.000E+00	0.0000	0.0000	0.0000
			6.163E+00	**40.8043**	**1.0000**	**-0.7872**
The three records with the lowest SumE contribute 63% to the TOB Stage 2 predicted value		**Solution 1-Var (DC) All 517 Records Applied**	**Total (X)**	**Total (Z)**	**Sum of F**	**Normalized Prediction = Sum (F*BA Ln)**
					Min Dependent Variable Actual Value (BA Ln):	0.0000
					Min Dependent Variable Actual Value (BA Ln):	6.9956
					Stage 2 Optimized Prediction of Dependent Variable Value (BA Ln) for Data Record: #144	**0.7443**
					Actual Recorded Dependent Variable Value (BA Ln) for Data Record: #144	0.5365
					Difference between Actual and TOB Stage 2 Predicted BA(Ln) Value:	-0.2079
					Actual Recorded Dependent Variable Value (BA) for Data Record: #144 = $(e^{(BALn)})$-1:	**0.71**
					Predicted TOB Stage 1 Dependent Variable Value (BA) for Data Record: #144 = $(e^{(BALn)})$-1:	**1.61**
					Predicted TOB Stage 2 Dependent Variable Value (BA) for Data Record: #144 = $(e^{(BALn)})$-1:	**1.11**

Fig. 5. Some details of the intermediate calculations involved in TOB prediction calculations for data record #144. the matching data records involved are revealed.

solutions generate lower prediction accuracy than TOB stage 1 solutions when applied to the full dataset (Tables 6 and 7).

The TOB method provides forensic auditing capabilities for each prediction made by transparently providing access to the data matching and variable error weighting in each prediction calculation (Tables 9 and 10). It is this identification of the best matching data records, and their contributions to each data record predictions, that represents the true data mining benefit of the TOB algorithm. This enables it to provide

Dependent Variable Prediction Value Calculated for Data Record # 369						
Rank of Top-Matching Data Records	Top-ranking Matched Records in Training Subset	Dependent Variable Normalized Value (BA Ln)	Sum of Weighted Squared Errors (SumE)	Relative Magnitude of Contribution to Prediction	Relative Contribution fraction	Contributions to Prediction Components
TOB Stage 1 Equal Weights & Q=10		W_N =0.5	**TOB Stage 1 Prediction for Test Record # 369**			
Testing Subset Record Case #1A:	**369**	**-0.2530**	**SumE**	**Y = Total (X)/ Sum E**	**F= Y/ Total (Z)**	**Predicted Value F*BA(Ln)**
1 (1st ranking match)	506	-0.4520	1.000E-07	158.9626	0.3017	-0.1364
2	424	-0.8450	1.000E-07	158.9626	0.3017	-0.2550
3	376	0.0571	1.000E-07	158.9626	0.3017	0.0172
4	348	-1.0000	2.228E-06	7.1346	0.0135	-0.0135
5	349	-1.0000	2.228E-06	7.1346	0.0135	-0.0135
6	357	-0.7669	2.228E-06	7.1346	0.0135	-0.0104
7	350	-0.7225	2.228E-06	7.1346	0.0135	-0.0098
8	354	-0.7139	2.228E-06	7.1346	0.0135	-0.0097
9	353	-0.6831	2.228E-06	7.1346	0.0135	-0.0093
10 (10th rankingMatch)	351	-0.5570	2.228E-06	7.1346	0.0135	-0.0075
The top three ranking matches contribute 91% to the TOB Stage 1 predicted value			**1.590E-05** **Total (X)**	**526.8304** **Total (Z)**	**1.0000** **Sum of F**	**-0.4478** **Normalized Prediction = Sum (F*BA Ln)**
					Min Dependent Variable Actual Value (BA Ln):	0.0000
					Min Dependent Variable Actual Value (BA Ln):	6.9956
					Stage 1 Provisional Prediction of Dependent Variable (BA Ln) Value for Data Record: #369	**1.9315**
					Actual Recorded Dependent Variable Value (BA Ln) for Data Record: #369	2.6130
					Difference between Actual and TOB Stage 1 Predicted BA (Ln) Value:	0.6815
TOB Stage 2 with Optimized Weights **Stage 2 TOB Optimized Q = 6**		**BA (Ln)**	**SumE**	**Y = Total (X)/ Sum E**	**F= Y/ Total (Z)**	**Predicted Value F*BA(Ln)**
6	506	-0.4520	8.232E-01	2.0910	0.0184	-0.0083
5	424	-0.8450	3.319E-01	5.1867	0.0458	-0.0387
1 (Best)	376	0.0571	2.339E-02	73.5913	0.6492	0.0371
4	348	-1.0000	2.819E-01	6.1067	0.0539	-0.0539
3	349	-1.0000	1.322E-01	13.0179	0.1148	-0.1148
2	357	-0.7669	1.288E-01	13.3687	0.1179	-0.0904
N/A	350	-0.7225	0.000E+00	0.0000	0.0000	0.0000
N/A	354	-0.7139	0.000E+00	0.0000	0.0000	0.0000
N/A	353	-0.6831	0.000E+00	0.0000	0.0000	0.0000
N/A	351	-0.5570	0.000E+00	0.0000	0.0000	0.0000
The three records with the lowest SumE contribute 88% to the TOB Stage 2 predicted value		**Solution 1-Var (DC) All 517 Records Applied**	**1.721E+00** **Total (X)**	**113.3623** **Total (Z)**	**1.0000** **Sum of F**	**-0.2691** **Normalized Prediction = Sum (F*BA Ln)**
					Min Dependent Variable Actual Value (BA Ln):	0.0000
					Min Dependent Variable Actual Value (BA Ln):	6.9956
					Stage 2 Optimized Prediction of Dependent Variable Value (BA Ln) for Data Record: #369	**2.5567**
					Actual Recorded Dependent Variable Value (BA Ln) for Data Record: #369	2.6130
					Difference between Actual and TOB Stage 2 Predicted BA(Ln) Value:	0.0563
					Actual Recorded Dependent Variable Value (BA) for Data Record: #369 = $(e^{(BALn)})$-1:	**12.64**
					Predicted TOB Stage 1 Dependent Variable Value (BA) for Data Record: #369 = $(e^{(BALn)})$-1:	**5.90**
					Predicted TOB Stage 2 Dependent Variable Value (BA) for Data Record: #369 = $(e^{(BALn)})$-1:	**11.89**

Fig. 6. Some details of the intermediate calculations involved in TOB prediction calculations for data record #369. the matching data records involved are revealed.

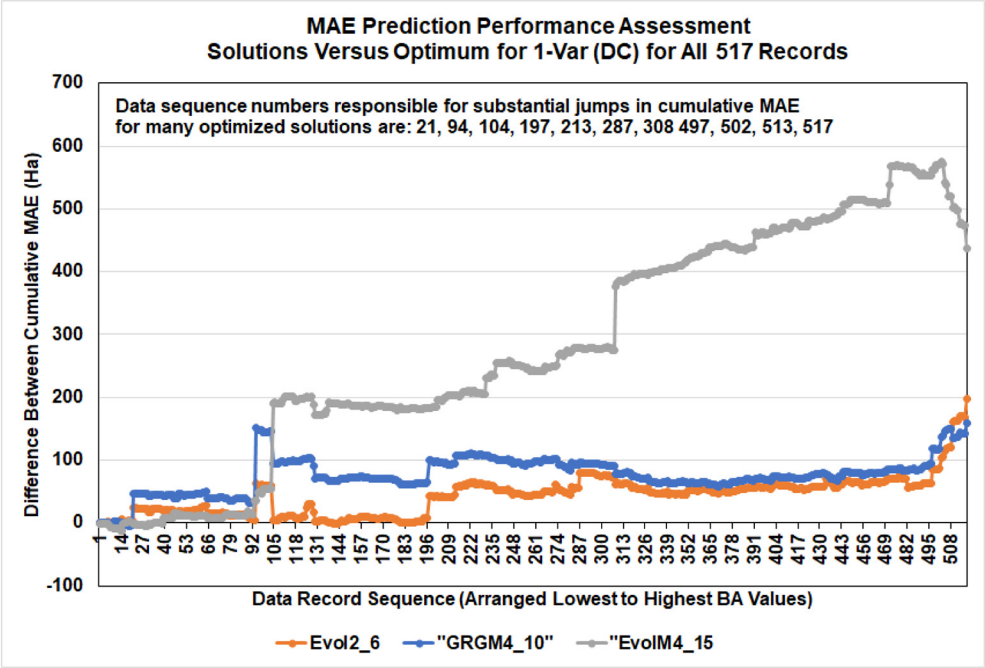


Fig. 7. Cumulative solution MAE differentials for Table 8 solutions versus "Best Solution Found" (MAE = 12.717 Ha; 1-Var (DC) model applied to all 517 data records). Noticeable jumps in these trends are linked to the prediction differences for a small group of data records for which the sequence order is listed in the graph.

ecological insight for each of its predictions. This capability differentiates the TOB method, in a positive way, from the machine learning algorithms that use regression, statistical or probabilistic relationships between the variables, which inhibit transparency and ready-access to the underlying details for individual predictions.

Graphical presentations of cumulative prediction error (Fig. 3) help to provide insight to a models prediction performance. It is useful to compare cumulative prediction error trends for the two predictions generated for each data record by the TOB algorithm. This reveals exactly for what parts of the total burned area distribution the predictions are being improved by the optimization process. The cumulative absolute error and squared error differentials between specific solutions are also usefully displayed as graphical trends (Figs. 5 and 6). Such displays help to make choices concerning which prediction model is most useful for specific ecological purposes. The ability to drill down into the specific data record predictions (Table 10 and 11) and understand the data records influencing each prediction, in positive or negative ways, is a valuable feature unique to the TOB method. For forest fire

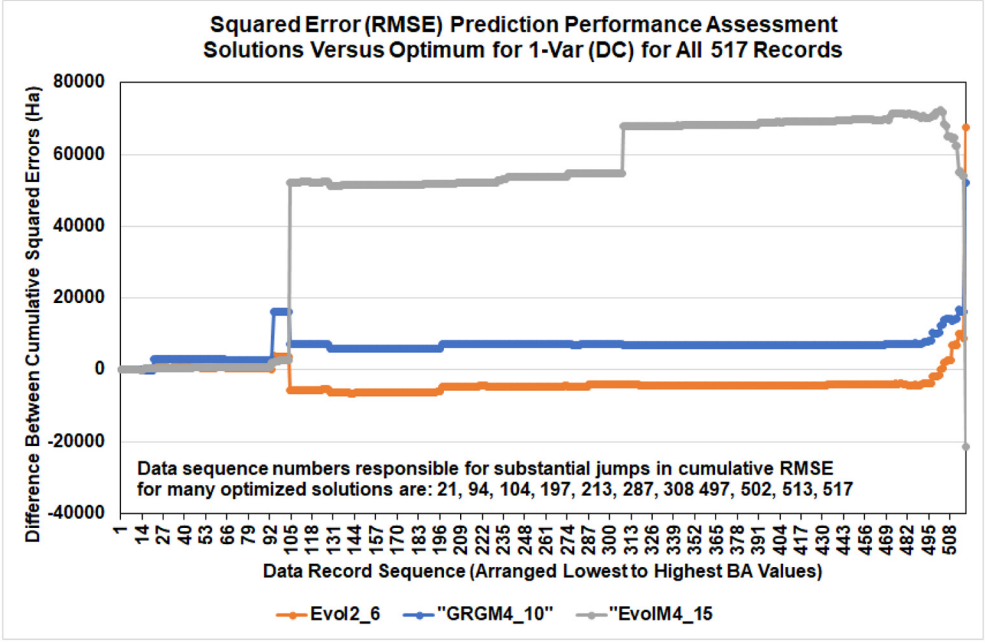


Fig. 8. Cumulative solution RMSE differentials for Table 8 solutions versus "Best Solution Found" (MAE = 12.717 Ha; 1-Var (DC) model applied to all 517 data records). Noticeable jumps in these trends are linked to the prediction differences for a small group of data records which for which the sequence order is listed in the graph.

predictions from datasets with many recorded input variables it makes ecological sense to move beyond regression-based machine learning and benefit from the auditability of optimized data matching to provide valuable insights from closely related historical records.

The detailed data mining made possible with the TOB solutions reveals a good deal of useful information about each data record (burn event). Information about the influences on accurate and inaccurate predictions associated with specific optimized solutions. This can help to surface important ecological information that is useful in making appropriate response decisions to specific burn incidents. It identifies which are the most useful TOB models and the most influential variables that provide, and consistently justify, BA predictions. Such improved confidence in BA predictions, and the greater understanding of the factors influencing each burn event derived from associated data mining, should lead to more appropriate responses to burn events as they happen. Moreover, it should lead to more informed follow-up mitigation responses that help to limit agricultural, ecological, environmental and forestry damage from similar incidents that might occur in the future.

6. Conclusions

An optimized data-matching machine learning algorithm can predict the total burned areas of a well-studied forest fire dataset from Portugal (Montesinho Natural Park) to a high degree of accuracy. While doing so it provides exceptional data mining insights (that are not immediately obvious) with which to interrogate each historical burn event. The transparent open box (TOB) algorithm avoids the use of regression, correlation and statistical distribution assumptions in making its predictions and does not involve any hidden layers or complex calculations.

For the dataset studied, TOB feature selection sensitivity analysis identified drought code (DC) and relative humidity (RH) as the most influential, of the thirteen variables considered, on TOB's total burned area predictions. The use of mean absolute error (MAE) and root mean squared error (RMSE) of the total burned area predictions separately as objective functions to minimize provides complementary information with respect to specific burned-area predictions from the highly positively skewed total burned area distributions of forest-fire datasets. The results indicate that transparent optimized data matching machine learning reveals more in wildfire datasets than regression-based machine learning algorithms.

The consideration of cumulative absolute error (and squared error) differentials, displaying all data records, aid the comparisons between the detailed prediction analysis achieved by high-performing TOB models. Graphical representations of these cumulative error trends facilitate data mining and reveal the few problematic data records that negatively influence the overall prediction accuracy achievable by specific model structures. This enables model structures to be selected that focus prediction accuracy on specific sectors of the total burned area distributions and/or more broadly across the entire distributions. This ability provides knowledge that is beneficial for sound agricultural, ecological, environmental and forestry responses to better respond to and manage specific burn events and potentially limit the impacts of future similar events. The method proposed could be readily adapted to predict and data mine generic agricultural systems that are dependent on complex interactions between weather and environmental variables going beyond regression- and correlation-based methods.

Funding

This study received no funding.

Credit author statement

I am the sole author of this manuscript and responsible for every aspect of its study and preparation.

Declaration of Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alkhatib, A.A., 2014. A review on forest fire detection techniques. Int. J. Distrib. Sens. Netw. 10, 597368–597380. https://doi.org/10.1155/2014/597368.

Angayarkkani, K., Radhakrishnan, N., 2009. Efficient forest fire detection system: a spatial data mining and image processing-based approach. Int. J. Comput. Sci. Netw. Secur. 9, 100–107.

Artes, T., Cortes, A., Margalef, T., 2016. Large forest fire spread prediction: data and computational science. Procedia Comp. Sci. 80, 909–918. https://doi.org/10.1016/j.procs.2016.05.330.

Castelli, M., Vanneschi, L., Popovič, A., 2015. Predicting burned areas of forest fires: an artificial intelligence approach. Fire Ecol. 11, 106–118. https://doi.org/10.4996/fireecology.1101106.

Caton-Kerr, S.E., Ali Tohidi, A., Gollner, M.J., 2019. Firebrand generation from thermally-degraded cylindrical wooden dowels. Front. Mech. Eng. 5, 32 12 pages. https://doi.org/10.3389/fmech.2019.00032.

Coffield, S.R., Graff, C.A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., Randerson, J.T., 2019. Machine learning to predict final fire size at the time of ignition. Int. J. Wildland Fire 28, 861–873. https://doi.org/10.1071/WF19023.

Cortez, P., Morais, A., 2007. A data mining approach to predict forest fires using meteorological data. In: Neves, J., Santos, M.F., Machado, J. (Eds.), New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512–523.

De Gennaro, M., Billaud, Y., Pizzo, Y., Garivait, S., Loraud, J.-C., El Hajj, M., Porterie, B., 2017. Real-time wildland fire spread modeling using tabulated flame properties. Fire Saf. J. 91, 872–881. https://doi.org/10.1016/j.firesaf.2017.03.006.

De Souza, F.T., Koerner, T.C., Chlad, R., 2015. A data-based model for predicting wildfires in Chapada das mesas National Park in the state of Maranhao. Environ. Earth Sci. 74, 3603–3611. https://doi.org/10.1007/S12665-015-4421-8.

Di Giuseppe, F., Remy, S., Pappenberger, F., Wetterhall, F., 2018. Using the fire weather index (FWI) to improve the estimation of fire emissions from fire radiative power (FRP) observations. Atmos. Chem. Phys. 18, 5359–5370. https://doi.org/10.5194/ACP-18-5359-2018.

Fehrmann, S., Jahn, W., de Dios Rivera, J., 2017. Permeability comparison of natural and artificial *Pinus radiata* forest litters. Fire. Technol 53, 1291–1308. https://doi.org/10.1007/s10694-016-0631-1.

Finney, M.A., 1998. Farsite: fire area simulator-model development and evaluation. Technical Report RMRS-4. USDA Forest Service, Rocky Mountain Research Station, Missoula, MT.

Finney, M.A., Cohen, J.D., McAllister, S.S., Jolly, W.M., 2013. On the need for a theory of wildland fire spread. Int. J. Wildland Fire 22, 25–36. https://doi.org/10.1071/WF11117.

Forsell, N., Garcia, F., Sabbadin, R., 2009. Reinforcement learning for spatial processes. 18th World IMACS/MODSIM Congress, Cairns, Australia, pp. 755–761.

Frontline Solvers, 2020. Standard Excel Solver - Limitations of Nonlinear Optimization. https://www.solver.com/standard-excel-solver-limitations-nonlinear-optimization [Accessed 3 January 2021].

Garzón, M.B., Blazek, R., Neteler, M., De Dios, R.S., Ollero, H.S., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* l. In the Iberian Peninsula. Ecol. Model. 197, 383–393. https://doi.org/10.1016/j.ecolmodel.2006.03.015.

Hedayati, F., Bahrani, B., Zhou, A., Quarles, S.L., Gorham, D.J., 2019. A framework to facilitate firebrand characterization. Front. Mech. Eng. 5, 43 14 pages. https://doi.org/10.3389/fmech.2019.00043.

Houtman, R.M., Montgomery, C.A., Gagnon, A.R., Calkin, D.E., Dietterich, T.G., McGregor, S., et al., 2013. Allowing a wildfire to burn: estimating the effect on future fire suppression costs. Int. J. Wildland Fire 22, 871–882. https://doi.org/10.1071/WF12157.

Jaafari, A., Zenner, E.K., Panahi, M., Shahabi, H., 2019. Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. Agric. For. Meteorol. 266–267, 198–207. https://doi.org/10.1016/j.agrformet.2018.12.015.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.

Liu, Z., Wimberly, M.C., 2016. Direct and indirect effects of climate change on projected future fire regimes in the western United States. Sci. Total Environ. 542, 65–75. https://doi.org/10.1016/J.SCITOTENV.2015.10.093.

Malarz, K., Kaczanowska, S., Kuakowski, K., 2002. Are forest fires predictable? Int. J. Mod. Phys. C 13, 1017–1031. https://doi.org/10.1142/S0129183102003760.

Manzello, S.L., Suzuki, S., 2017. Generating wind-driven firebrand showers characteristic of burning structures. Proc. Combust. Inst. 36, 3247–3252. https://doi.org/10.1016/j.proci.2016.07.009.

Martell, D.L., 2015. A review of recent forest and wildland fire management decision support systems research. Curr. For. Rep. 1, 128–137. https://doi.org/10.1007/s40725-015-0011-y.

McAllister, S., 2019. The role of fuel bed geometry and wind on the burning rate of porous fuels. Front. Mech. Eng. 5, 11. https://doi.org/10.3389/fmech.2019.00011.

McAllister, S., Finney, M., 2016. Burning rates of wood cribs with implications for wildland fire. Fire. Technol 52, 1755–1777. https://doi.org/10.1007/s10694-015-0543-5.

Mcgregor, S., Houtman, R., Buckingham, H., Montgomery, C., Metoyer, R., Dietterich, T.G., 2016. Fast simulation for computational sustainability sequential decision-making problems. Proceedings of the 4th International Conference on Computational Sustainability, New York, pp. 5–7.

Miller, C., Tang, W., Finney, M., McAllister, S., Forthofer, J., Gollner, M., 2017. An investigation of coherent structures in laminar boundary layer flames. Combust. Flame 181, 123–135. https://doi.org/10.1016/j.combustflame.2017.03.007.

Montgomery, C., 2014. Chapter 13: fire: an agent and a consequence of land use change. In: Duke, J.M., Wu, J. (Eds.), The Oxford Handbook of Land Economics. Oxford University Press, pp. 281–301.

Moon, K., Duff, T.J., Tolhurst, K.G., 2019. Sub-canopy forest winds: understanding wind profiles for fire behaviour simulation. Fire Saf. J. 105, 320–329. https://doi.org/10.1016/j.firesaf.2016.02.005.

Peng, G., Li, J., Chen, Y., Norizan, A.P., Tay, L., 2006. High-resolution surface relative humidity computation using modis image in peninsular Malaysia. Chin. Geogr. Sci. 16, 260–264. https://doi.org/10.1007/s11769-006-0260-6.

Quill, R., Sharples, J.J., Wagenbrenner, N.S., Sidhu, L.A., Forthofer, J.M., 2019. Modeling wind direction distributions using a diagnostic model in the context of probabilistic fire spread prediction. Front. Mech. Eng. 5, 5 16 pages. https://doi.org/10.3389/fmech.2019.00005.

Raposo, J.R., Viegas, D.X., Xie, X., Almeida, M., Figueiredo, A.R., Porto, L., Sharples, J., 2018. Analysis of the physical processes associated with junction fires at laboratory and field scales. Int. J. Wildland Fire 27, 52–68. https://doi.org/10.1071/WF16173.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146, 303–310. https://doi.org/10.1016/S0304-3800(01)00316-7.

Rios, O., Valero, M.M., Pastor, E., Planas, E., 2019. A data-driven fire spread simulator: validation in Vall-llobrega's fire. Front. Mech. Eng. 5, 8 11 pages. https://doi.org/10.3389/fmech.2019.00008.

Rodrigues, A., Ribeiro, C., Raposo, J., Viegas, D.X., André, J., 2019. Effect of canyons on a fire propagating laterally over slopes. Front. Mech. Eng. 5, 41 9 pages. https://doi.org/10.3389/fmech.2019.00041.

Sanjuan, G., Margalef, T., Cortés, A., 2016. Applying domain decomposition to wind field calculation. Parallel Comput. 57, 484–490. https://doi.org/10.1109/HPCSim.2015.7237080.

Santoso, M.A., Christensen, E.G., Yang, J., Rein, G., 2019. Review of the transition from smouldering to flaming combustion in wildfires. Front. Mech. Eng. 5, 49 20 pages. https://doi.org/10.3389/fmech.2019.00049.

Saranya, N.N., Hemalatha, M., 2012. Integration of machine learning algorithm using spatial semi supervised classification in FWI data. 2012 International Conference on Advances in Engineering, Science and Management (ICAESM). IEEE, Nagapattinam, pp. 699–702.

Sehgal, V., Getoor, L., Viechnicki, P.D., 2006. Entity resolution in geospatial data integration. Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems. ACM, Arlington, Virginia, pp. 83–90.

Sitanggang, I.S., Ismail, M.H., 2011. Classification model for hotspot occurrences using a decision tree method. Geomat. Nat. Hazards Risk 2, 111–121. https://doi.org/10.1080/19475705.2011.565807.

Stocks, B.J., Lynham, T., Lawson, B., Alexander, M., Wagner, C.V., McAlpine, R., et al., 1989. Canadian forest fire danger rating system: an overview. For. Chron. 65, 258–265. https://doi.org/10.5558/tfc65258-4.

Subramanian, G.S., Crowley, M., 2018. Using spatial reinforcement learning to build forest wildfire dynamics models from satellite images. Front. ICT 5, 6. https://doi.org/10.3389/fict.2018.00006.

Subramanian, S.G., Crowley, M., 2017. Learning forest wildfire dynamics from satellite images using reinforcement learning. Conference on Reinforcement Learning and Decision Making. Ann Arbor, MI.

Tang, W., Finney, M., McAllister, S., Gollner, M., 2019. An experimental study of intermittent heating frequencies from wind-driven flames. Front. Mech. Eng. 5, 34. https://doi.org/10.3389/fmech.2019.00034.

Taylor, S., Alexander, M., 2006. Science, technology, and human factors in fire danger rating: the Canadian experience. Int. J. Wildland Fire 15, 121–135.

Tohidi, A., Kaye, N.B., 2017. Comprehensive wind tunnel experiments of lofting and downwind transport of non-combusting rod-like model firebrands during firebrand shower scenarios. Fire Saf. J. 90, 95–111. https://doi.org/10.1016/j.firesaf.2017.04.032.

UCI Machine Learning Repository, 2008. Forest fires machine learning dataset with 517 data records and 13 attributes from Montesinho Natural Park in Portugal, January 2000 to December 2003 (deposited by P. Corez and A.D.J.R. Morais in February 2008). https://archive.ics.uci.edu/ml/datasets/Forest+Fires.

Veraverbeke, S., Rogers, B.M., Goulden, M.L., Jandt, R.R., Miller, C.E., Wiggins, E.B., Randerson, J.T., 2017. Lightning as a major driver of recent large fire years in north American boreal forests. Nat. Clim. Chang. 7, 529–534. https://doi.org/10.1038/NCLIMATE3329.

Wang, H., van Eyk, P.J., Medwell, P.R., Birzer, C.H., Tian, Z.F., Possell, M., Huang, X., 2019. Air permeability of the litter layer in broadleaf forests. Front. Mech. Eng. 5, 53 15 pages. https://doi.org/10.3389/fmech.2019.00053.

Wijayanto, A.K., Sani, O., Kartika, N.D., Herdiyeni, Y., 2017. Classification model for forest fire hotspot occurrences prediction using ANFIS algorithm. IOP Conf. Series: Earth Environ. Sci. 54, 012059. https://doi.org/10.1088/1755-1315/54/1/012059.

Wood, D.A., 2016. Metaheuristic profiling to assess performance of hybrid evolutionary optimization algorithms applied to complex wellbore trajectories. J. Nat. Gas Sci. Eng. 33, 751–768. https://doi.org/10.1016/j.jngse.2016.05.041.

Wood, D.A., 2018. A transparent open-box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. Adv. Geo-Energy Res. 2 (2), 148–162. https://doi.org/10.26804/ager.2018.02.04.

Wood, D.A., 2019a. Transparent open box learning network provides auditable predictions for coal gross calorific value. Model. Earth Syst. Environ. 5, 395–419. https://doi.org/10.1007/s40808-018-0543-9.

Wood, D.A., 2019b. Sensitivity analysis and optimization capabilities of the transparent open box learning network in predicting coal gross calorific value from underlying compositional variables. Model. Earth Syst. Environ. 5, 753–766. https://doi.org/10.1007/s40808-019-00583-1.

Yang, J., Liu, N., Chen, H., Gao, W., 2018. Smoldering and spontaneous transition to flaming over horizontal cellulosic insulation. Proc. Combust. Inst. 37, 4073–4081. https://doi.org/10.1016/j.proci.2018.05.054.

Yongzhong, Z., Feng, Z.-D., Tao, H., Liyu, W., Kegong, L., Xin, D., 2004. Simulating wildfire spreading processes in a spatially heterogeneous landscapes using an improved cellular automaton model. Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International. 5. IEEE, Anchorage, AK, pp. 3371–3374.

Young, A.M., Higuera, P.E., Duffy, P.A., Hu, F.S., 2017. Climatic thresholds shape northern high-latitude fire regimes and imply vulnerability to future climate change. Ecography 40, 606–617. https://doi.org/10.1111/ECOG.02205.

Zhang, G., Wang, M., Liu, K., 2019. Forest fire susceptibility modeling using a convolutional neural network for Yunnan Province of China. Int J Disaster Risk Sci. 10, 386–403. https://doi.org/10.1007/s13753-019-00233-1.

Zhang, J.-H., Yao, F.-M., Liu, C., Yang, L.-M., Boken, V.K., 2011. Detection, emission estimation and risk prediction of forest fires in China using satellite sensors and simulation models in the past three decades – an overview. Int. J. Environ. Res. Public Health 8, 3156–3178. https://doi.org/10.3390/ijerph8083156.