


Area and Feature Guided Regularised Random Forest: a novel method for predictive modelling of binary phenomena. The case of illegal landfill in Canary Island

Lorenzo Carlos Quesada-Ruiz, Victor Francisco Rodriguez-Galiano, Raúl Zurita-Milla & Emma Izquierdo-Verdiguier


To cite this article: Lorenzo Carlos Quesada-Ruiz, Victor Francisco Rodriguez-Galiano, Raúl Zurita-Milla & Emma Izquierdo-Verdiguier (2022) Area and Feature Guided Regularised Random Forest: a novel method for predictive modelling of binary phenomena. The case of illegal landfill in Canary Island, International Journal of Geographical Information Science, 36:12, 2473-2495, DOI: [10.1080/13658816.2022.2075879](https://doi.org/10.1080/13658816.2022.2075879)


To link to this article: <https://doi.org/10.1080/13658816.2022.2075879>

 View supplementary material [↗](#)

 Published online: 09 Jun 2022.

 Submit your article to this journal [↗](#)

 Article views: 283



 View related articles [↗](#)

 View Crossmark data [↗](#)

RESEARCH ARTICLE



Area and Feature Guided Regularised Random Forest: a novel method for predictive modelling of binary phenomena. The case of illegal landfill in Canary Island

Lorenzo Carlos Quesada-Ruiz^a , Victor Francisco Rodriguez-Galiano^a , Raúl Zurita-Milla^b  and Emma Izquierdo-Verdiguier^c 

^aDepartment of Physical Geography and Regional Geographical Analysis, University of Seville, Seville, Spain; ^bFaculty of Geoinformation Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands; ^cInstitute of Geomatics, University of Natural Resources and Life Sciences, Vienna (BOKU), Vienna, Austria

ABSTRACT

This paper presents a novel method, Area and Feature Guided Regularised Random Forest (AFGRRF), applied for modelling binary geographic phenomenon (occurrence versus absence). AFGRRF is a wrapper feature-selection method based on a previous modification of Random Forest (RF), namely the Guided Regularised Random Forest (GRRF). AFGRRF produces maps that minimise the affected area without a significant difference in accuracy. For this, it tunes the GRRF hyper-parameters according to a trade of between True Positive Rate and the affected area (Success Rate). AFGRRF also addresses the ‘Rashomon effect’ or the multiplicity of good models. The proposed method was tested to model illegal landfills in Gran Canaria Island (Spain). AFGRRF performance was compared to that of other RF-based methods: (i) standard RF; (ii) Area Random Forest (ARF); (iii) Feature Random Forest (FRF); (iv) Area Feature Random Forest (AFRF) and (v) GRRF. AFGRRF predicted the smallest affected area, 19% of the island, at a similar True Positive Rate. This percentage is substantially smaller than the one predicted by RF (27.43%), ARF (26%), FRF (27.78%), AFRF (23%) and GRRF (29.67%).

ARTICLE HISTORY


Received 8 August 2021
Accepted 6 May 2022


KEYWORDS

Random Forest; feature selection; predictive modelling; binary phenomena; success rate; illegal landfill

1. Introduction

Predictive modelling algorithms identify and learn patterns between a target feature and other independent features from a given subset of training samples. Hence, predictive modelling enables estimation of said target feature’s behaviour using independent features, regardless of when it occurred (past, present or future), whereas forecasting makes projections into the future (Breiman 2001a, Quesada-Ruiz *et al.* 2019a). Predictive modelling has become an important tool for mapping the distribution of multiple geographical phenomena in Earth sciences (Soares and Pereira 2007, Dahal *et al.* 2008, Tehrany *et al.* 2013). These phenomena are often binary in nature,

CONTACT Lorenzo Carlos Quesada-Ruiz  lquesada@us.es

 Supplemental data for this article can be accessed [here](#).

© 2022 Informa UK Limited, trading as Taylor & Francis Group

i.e. occurrence: the absence or presence of the phenomenon (Breslow and Cain 1988, Schill *et al.* 1993, Carranza *et al.* 2008).

Following Rodriguez-Galiano *et al.* (2012), an occurrence map may be considered appropriate, in addition to being accurate, when: (a) the spatial distribution of the phenomenon is consistent with respect to the most important explanatory features; (b) it is replicable, achieving a certain stability in the predicted values and (c) the phenomenon is not over or underestimated. Ignoring this latter aspect may lead to cost overruns for many tasks, especially when we use GIS methods applied to binary mapping. Examples of this include: landslide prevention (Dahal *et al.* 2008, Harris and Grunsky 2015, Hong *et al.* 2017, Chen *et al.* 2019), where potentially affected areas must be defined to efficiently locate slope-stabilisation actions; flood prevention (Tehrany *et al.* 2013), requiring the construction of containment walls against possible floods; ecosystem conservation (Poulos *et al.* 2016, Huettmann *et al.* 2018, Zhang *et al.* 2019), where estimation of the presence and/or absence of the various habitats of an ecosystem helps to facilitate their respective management; and infectious disease (Cecchi *et al.* 2009, Bhunia *et al.* 2012, Iftimi *et al.* 2015) and agricultural pest control (Wittmann *et al.* 2001, Porretta *et al.* 2013, Kumar *et al.* 2016), which require locating the area that contributes to the spread of a given pathogen, as well as predicting the potential area that might be affected in future. In this sense, the concept of potentially affected area (hereafter affected area) is central to the predictive modelling of binary phenomena and refers to areas that may suffer or withstand potential damage or risk (i.e. areas with non-zero probability).

The accuracy of predictive modelling is largely grounded on the classification method and its optimisation using the training data (Foody 2004, Visser and Nijs 2006). Methods for predictive modelling have different abilities to learn patterns, sometimes needing a specific statistical distribution in the features (i.e. normality) (Rodriguez-Galiano *et al.* 2014, Leuenberger and Kanevski 2015, Arabameri *et al.* 2019). However, other aspects that have traditionally received less attention, such as the metrics for assessing performance or Feature Selection, are also important, and might have an impact on both the area and the spatial distribution (Rodriguez-Galiano *et al.* 2018). Predictive models are typically built using large sets of explanatory features (e.g. information about geology, biology or socioeconomic factors, etc.). However, even if the number of samples is notably larger than the number of features, high feature space dimensionality can overwhelm the method's learning capacity (the curse of dimensionality; Chen 2009). Also, selecting a large number of features would lead to models that are difficult to both interpret and replicate. Thus, dimensionality reduction is often needed. Dimensionality reduction is primarily achieved by Feature Extraction or Feature Selection. Feature extraction methods reduce data down to a smaller representative set, projecting these into the most relevant directions of a lower feature space, as in the case of the Principal Component Analysis method (Lucas and Jauzein 2008, Menció and Mas-Pla 2008, Canela *et al.* 2011). Conversely, feature selection methods do not modify the features of the original data; rather, they select a reduced yet meaningful feature subset, improving both interpretability and the accuracy of the model (Blum and Langley 1997, Dash and Liu 1997, Guyon and Elisseeff 2003). Some negative effects could thus be averted using feature selection, such as (Rodriguez-Galiano *et al.* 2018): (i) model overfitting; (ii) limitation of the model's interpretability due to high complexity; (iii) loss of generalisation capacity

and (iv) a significant increase in computational time. A controversial aspect of feature selection is the multiplicity of good models, which is also common in statistical algorithms, such as multiple regression or logistic regression. Different feature subsets might share good and similar accuracy, thus resulting in a non-unique solution or physical model explaining a phenomenon (Rashomon effect, Breiman 2001b).

Among the different approaches for feature selection, filters, embedded and wrapper methods stand out (Hall and Smith 1997, Tuv 2009). Filters select features regardless of the predictive model and accuracy of predictions (e.g. linear correlation) (Guyon and Elisseeff 2003, Dixon 2005). Current approaches include embedded methods, which are algorithms that include an internal estimate of a feature's importance based on different metrics, such as gain or mean decrease in accuracy. Some examples of this algorithm type are decision trees or Random Forest (RF). However, embedded methods only provide a ranking of a feature's importance and do not determine the optimal number of features (Bazi and Melgani 2006, Tuv 2009, Pal and Foody 2010, Rodriguez-Galiano *et al.* 2012). Finally, wrapper-based approaches select an optimal subset of features, repeatedly and automatically training the model with different subsets (Guyon and Elisseeff 2003). The design of the wrapper algorithm for feature selection requires three components: a predictive algorithm (i.e. RF, support vector machines or neural networks), a method for searching in the feature space (i.e. forward or backward deterministic search, exhaustive search, genetic algorithms etc.) and a metric for evaluating performance (i.e. RMSE in the case of regression, Receiver Operating Curve (ROC) or overall accuracy in the case of classification) (Rodriguez-Galiano *et al.* 2018). Wrappers are thus very computationally intense algorithms (Hall and Smith 1997, Navin Lal *et al.* 2006). RF-based algorithms are well-suited to building wrappers because of their low sensitivity to hyperparameter tuning and their robustness and speed from a computational standpoint (Breiman 2001a). Various RF-based wrapper methods have been proposed in Earth sciences, using either sequential search (Rodriguez-Galiano *et al.* 2018) or exhaustive grid search, such as Guided Regularised Random Forest (GRRF) (Deng and Runger 2013, Izquierdo-Verdiguier and Zurita-Milla 2020). This paper presents the Area and Feature Constrained Random Forest (AFGRRF) binary classification method. The proposed method is a new machine learning feature selection method that can also be used for predictive modelling. Other specific objectives include: (i) assessing the application of different Random Forest based algorithms to binary mapping; (ii) reducing the affected area and therefore the environmental management costs for binary phenomena.

2. Afgrrf classification method

2.1. Modelling background

AFGRRF is a modification of the GRRF algorithm that prevents an overestimation of the affected area by optimising both the True Positive Rate (TPR) and the affected area via Success Rate (SR) application (see Section 2.3). AFGRRF may be a novel way to address the Rashomon effect, by selecting the feature subset from among multiple good predictive models that leads to a smaller affected area. The proposed method is tested in a case study to predict the possible distribution of illegal landfills (ILs) on Gran Canaria island in Spain. Gran Canaria is an island within the Canary archipelago, which are an

outermost region of the European Union and a Spanish autonomous region. Gran Canaria has an area of 1560 km² and is the second most populated among the islands (845,000 residents) after Tenerife (891,000 residents) (INE 2016a). The population of Gran Canaria is mainly located in coastal areas, while the interior is less populated. The Canary Islands rank eight within the Spain's gross domestic product. According to Cruz *et al.* (2011), the major driver of economic activity on Gran Canaria is the tourism, which has led to a strong boost in the construction sector. Tourism on Gran Canaria is fundamentally beach-related, being concentrated in the South of the island. Around 4.2 million people visited the island in 2016 (INE 2016b). The Canary Islands comprises a small and fragmented territory where space is a lack of resource, limiting and hampering territorial planning and land-use management. Hence, the creation of waste-management infrastructures (GOBCAN 2015, 2008) and the containment of ILs is an important challenge (GOBCAN 2015 2008, Quesada-Ruiz *et al.* 2019a, 2019b). IL are an environmental management problem for the Canary Islands as in many countries, harming the environment, human health and local economies (Quesada-Ruiz *et al.* 2019b).

The primary impacts of IL are local landscape deterioration, air pollution, aquifer pollution and increased risk to human health (Bridges *et al.* 2000, Monteiro Santos *et al.* 2006, Ichinose and Yamamoto 2011). The cost associated with locating and remediating IL has been estimated per year, for example by (i) the Environment Agency of the United Kingdom, at 120–175 million euros in the UK; (ii) The Queensland Government (Australia), at 4 million euros (EUR 420 per tonne) (Glanville and Chang 2015); (iii) The Pennsylvania Department of Transportation in the United States, with an annual tax cost for waste clean-up of approximately 8.6 million euros (EUR 710 per tonne) (PPRC 2016). Moreover, waste management on the Canary Islands is more challenging than in other places due to a lack of waste facilities (Quesada-Ruiz *et al.* 2018). Gran Canaria has experienced an increase of 317.7 ha in areas affected by IL between 2000 and 2012 due to urban sprawl and the housing bubble (Quesada-Ruiz *et al.* 2019a). Previous studies have identified 'construction and demolition waste' as the most abundant IL typology in Gran Canaria (Quesada-Ruiz *et al.* 2018). Additionally, the lack of dissuasive measures in more than 95% of IL cases reflects the urgent need for monitoring and prevention policies (Quesada-Ruiz *et al.* 2018). Hence, an accurate delimitation of IL-affected areas would reduce control and monitoring costs, supporting the implementation of deterrence measures such as environmental control patrols or installation of video cameras and posters, optimising and delimiting areas where prior intervention was implemented. On the other hand, it could help to local government to create citizen participation programs, encouraging the prevention of IL by the citizen participation and increasing their opportunities to utilise waste treatment infrastructures, with the objective of meliorate waste collection process and environmental education policies in those areas (Quesada-Ruiz *et al.* 2018).

This case study was selected because ILs are clearly binary in nature, i.e. they either exist or do not exist. Hence, they are suitable to test the applicability of the proposed method. Moreover, ILs represent a problem that requires significant economical resources and manpower from local authorities in order to control and manage them. Thus, further optimisation in modelling them helps to reduce the environmental management costs (Ichinose and Yamamoto 2011, Glanville and Chang 2015), which are mainly waste disposal and site remediation, and surveillance costs of landfilling (Tasaki *et al.* 2007).

2.2. Modelling principles

AFGRRF is a new algorithm that can be applied in the GIS framework for selecting feature subsets for mapping binary phenomena, applied in this case to predictive modelling of ILs. AFGRRF carries out a regularisation and an exhaustive grid search identical to the GRRF method (Deng and Runger 2013, Izquierdo-Verdiguier and Zurita-Milla 2018, 2020). AFGRRF generate multiple models based on different feature subsets spatially related with the occurrence of IL according to the 100 possible combinations of the gamma and lambda values (Figure 1), being their corresponding values between 0.1 and 1 by intervals of 0.1. AFGRRF trains multiple soft classification models with RF using the different feature subsets generated from the Guided Grid search regularisation. Each soft map built from different feature subsets is reclassified iteratively based on the SR. The results of SR can be shown in a graph where the TPR for different IL affected area percentages is represented (see Figure 7). The TPR (true positives/(true positives + false negatives)) is computed by finding the binary class probability membership threshold values that split the map in different affected areal quantiles. The TPR value is computed for each map reclassified as affected and unaffected by IL using an independent test. The model that is finally selected by AFGRRF is the one that is obtained from the feature subset that leads to the minimum IL affected area at a TPR equal to or greater than 90%. This TPR reference value can also be adjusted and modified according to the needs. Therefore, AFGRRF is based on optimising SR and minimising the IL potential affected area, serving as an alternative to traditional wrappers, which are based on overall accuracy. In that sense, the method proposed use a widely feature subset of possible features related to the ILs problem, such as distance to coast or distance to industrial areas, and selected the features or possible combinations of features according to their spatial distribution and relation with ILs occurrence. Hence, the method tries to map the minimal affected areas of ILs in a most accurate way, considering the ILs sample distribution, for reducing the cost of surveillance, recovery and restauration of the new possible potential affected areas. AFGRRF pseudocode could be summarised as follows:

1. Train a RF model.
2. Obtain the embedded RF importance.
3. Guided Grid search regularisation
 - a. Initialise an empty subset of selected features and a threshold gain ($G^* = 0$)
 - b. Fix the values of λ and γ to calculate α .
 - c. Computation of G_{GRRF}
 - d. If, $G_{GRRF}(x_j, \nu) > G^*$ the feature j is selected and the threshold gain is updated to the GRRF gain. Otherwise, the feature is not selected.
4. Multiple soft RF models are built from the various feature subsets.
5. Feature subset selection based on SR
 - a. Each soft map is reclassified into multiple binary hard maps considering different percentages of affected area (pixel quantiles)
 - b. TPR is computed for all binary maps at increasing areal percentages for each feature subset.
6. Model selection based on a trade-off between TPR and minimal area from SR.

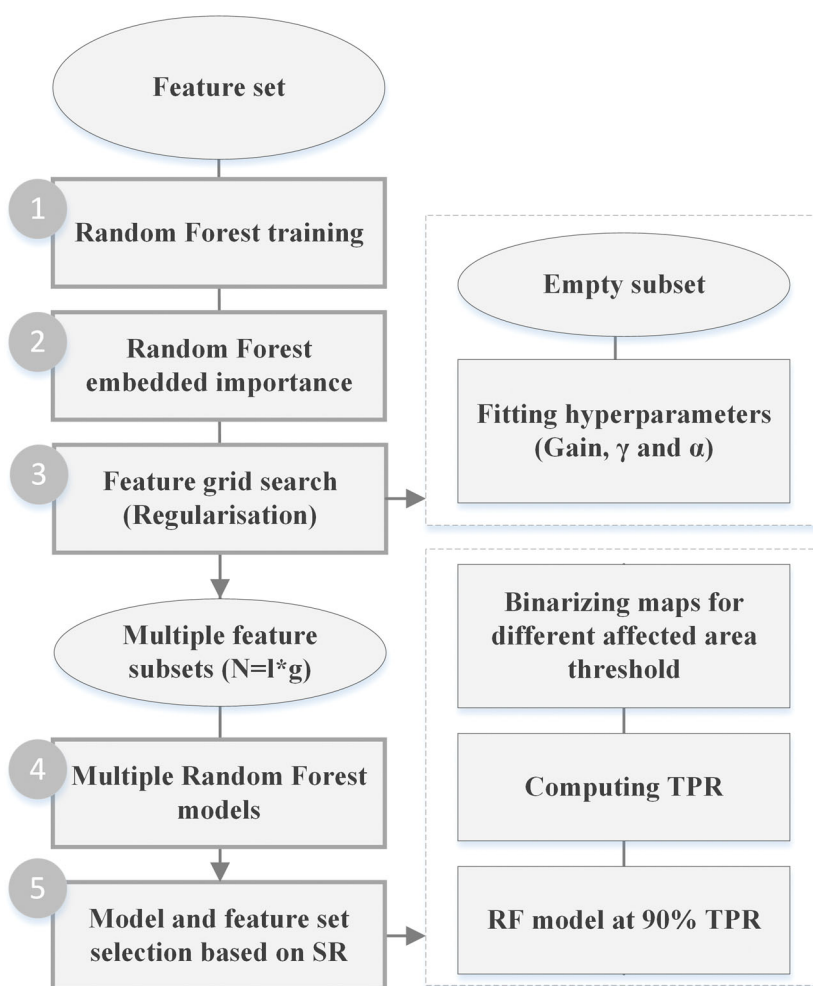


Figure 1. Area Feature Guide Regularised Random Forest (AFGRRF) flowchart. AFGRRF is a wrapper feature-selection method based on a modification of Guide Regularised Random Forest (GRRF), applied to the mapping of binary geographic phenomenon (occurrence versus absence) and exemplified in the case of illegal landfills (IL) occurrence mapping. AFGRRF trains multiple soft classification models using the different feature subsets (i.e. distance to coast line, communications routes density, industrial activity index, etc.) generated from the guided grid search regularisation. AFGRRF tunes the hyper-parameters of the GRRF for selecting a feature subset spatially related with the IL occurrence according to a trade of between True Positive Rate (proportion of ILs that are correctly classified over the total number of IL;TPR) and the IL affected area (Success Rate). The model selected is the one that leads to the minimum affected area by ILs at a TPR equal to or greater than 90%, producing maps that minimise the affected area without a significant difference in accuracy and allowing the cost reduction of environmental management actions.

2.3. Accuracy assessment metrics

Besides the wrapper's general performance metrics (overall accuracy or Kappa), other metrics are also used for binary classification, such as: the percentage of true positives

and true negatives, and the percentage of false positives and false negatives (Fawcett 2006, Powers 2007). The true positives and true negatives represent the number of successes between the predicted values and real values for locations where a phenomenon is present and absent, respectively. The false positives and false negatives measure the percentage of errors between the predicted values and real values for locations where a phenomenon is present and absent, respectively. A good model will thus be one that contains a high percentage of true positives and true negatives and a low percentage of false positives and false negatives. Therefore, analysing false positives to avoid overestimations is just as important as analysing false negatives to avoid underestimations. However, none of these metrics take into consideration both the accuracy of the classifications and the extent of the area affected by a phenomenon. Binary maps, such as the occurrence of ILs, might be improved from the standpoint of economic cost of remediation and monitoring, if feature selection was optimised using metrics that consider both the accuracy and the extent of the affected area, such as the success rate (Chung and Fabbri 1999).

SR represents the TPR for binary predictive maps with increasing affected area (Chung and Fabbri 1999). The SR is represented in a graph with the TPR on the y axis and different affected area percentage on the x axis (see Figure 7). The maps for increasing areal percentages are computed by reclassifying taking into consideration the classification probability threshold values at different quantiles. The TPR is computed for each map using an independent test. This way, a map at a good accurate level (TPR) that minimises the affected area can be chosen when success rate function converges.

3. Experimental validation

3.1. Experimental data

An IL database for GC (Figure 2) was used for the experimental design. It was generated by interpreting digital orthophotos for the years 2012 and 2015 and through complementary field work in which 387 potential locations were visited (Quesada-Ruiz *et al.* 2018). 286 IL locations were obtained after filtering out IL that were less than two years old and with an area smaller than 2000 m² with a view to rejecting temporary and small dump sites (Quesada-Ruiz *et al.* 2018). Information on socioeconomic aspects obtained from the Spanish National Institute of Statistics (e.g. per capita income, population, industrial and tourism activity indices), as well as geomorphology was obtained for the study area from Spanish National Institute of Geography. After preliminary process, 117 features (see supplementary material: Table 1a and Table 1b) that could be linked to IL occurrence were derived from this information (Biotto *et al.* 2009, Alexakis and Sarris 2014, Quesada-Ruiz *et al.* 2019b), such as population size and density, per capita income, industrial and touristic activity indices, elevation and slope, etc. New features were extracted from this initial feature set using different GIS analysis procedures (Şener *et al.* 2011, Demesouka *et al.* 2014, Uyan 2014, Akbari and Rajabi 2017): interpolating socioeconomic information aggregated by population centres; considering the calculation of Euclidean distance between the IL location and elements of interest, such as infrastructure, equipment, population centres, coast, land use etc. (Biotto *et al.* 2009, Tasaki *et al.* 2007) computing kernel densities of elements of interest, such as

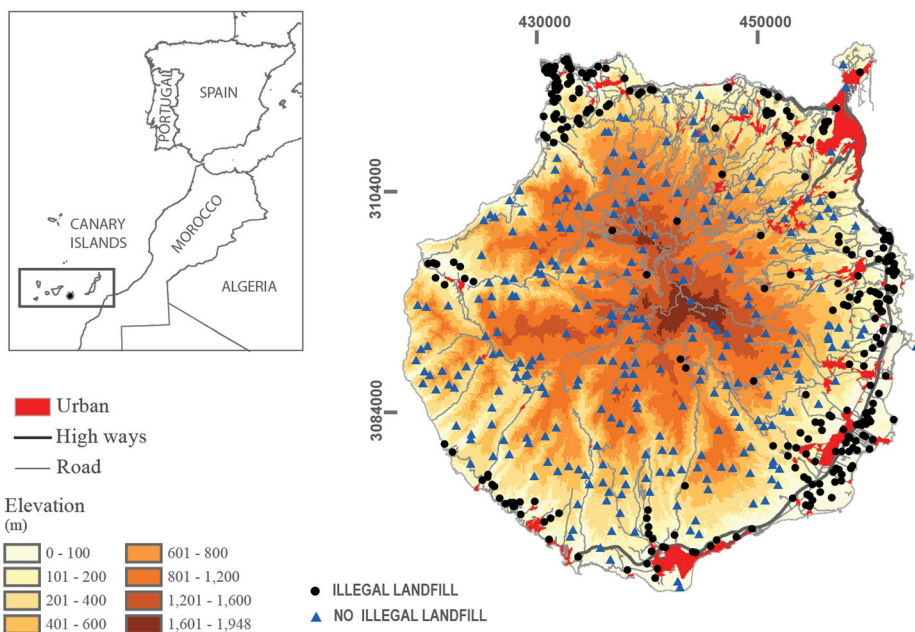


Figure 2. Study area.

communication routes or buildings, and other distance-based search functions for different radio (250 m, 500 m, 1500 m) (Silverman 1986). Additionally, the Normalised Difference Vegetation Index (NDVI) (Silvestri and Omri 2008) was obtained from a SPOT-5 summer image for 31st August with 11% of cloud coverage and 10 m of spatial resolution. The primary features were rasterised, standardised and resampled at a spatial resolution of 10 m. Table 1 shows the main features used in the experimental design grouped by typology. Following Carranza *et al.* 2008, the database was completed by including no-IL locations (i.e. places free of IL) to distinguish areas of negative IL occurrence, carrying out a stratified random sampling (Quesada-Ruiz *et al.* 2018). The negative and positive IL occurrence locations were coded as 0s and 1s, respectively, with an overall result of 286 negative samples and 286 positive samples. All feature values were obtained for both negative and positive IL locations.

3.2. Experimental design

The performance of AFGRRF was compared to a baseline composed of five different RF methods: (i) Random Forest (RF); (ii) Area Random Forest (ARF); (iii) Feature Random Forest (FRF); (iv) Area Feature Random Forest (AFRF) and (v) GRRF. The hard classification methods (RF, FRF and GRRF) produce categorical maps, considering by default an arbitrary threshold value of 0.5 in the class conditional probability. The soft classification methods (ARF, AFRF and AFGRRF), which predict a class conditional probability, were assessed in terms of the smallest affected area at a TPR equal or higher than 90%. In this sense, the soft classification methods estimate the class conditional probabilities and after perform classification based on estimated probabilities. Each method used 500 trees and default *mtry* parametrisation (square root of the number

Table 1. Features used in the experimental design, grouped by typology.

Feature typology	Units
Socioeconomic	
Population density	km ⁻²
Mining and extraction activity index	%
Industrial activity index	%
Distance	
Distance to pit zones	m
Distance to transport infrastructure	m
Distance to pit zones with different kernels	m
Distance to transport infrastructures	m
Distance to element of interest	m
Distance to educational equipment	m
Distance to coast	m
Distance to protected areas	m
Distance to cultural equipment	m
Distance to agricultural areas	m
Visibility	
Visibility from the coastline	Unitless
Physiographic	
Slope	%
Altitude	m
NDVI index	Unitless
Density	
Buildings density	km ⁻²
Land use transitions density from 1990 to 2000	km ⁻²
Land use transition density from 1990 to 2012	km ⁻²
Impervious cover transitions density from 1990 to 2012	km ⁻²
Greenhouses density	km ⁻²
Communication routes density	km ⁻²

of features) to ensure the stability of the results. RF was an embedded method without feature selection to generate a hard classification model. FRF was a wrapper for feature selection that used RF embedded importance and a forward sequential search. Forward sequential search starts from the empty feature space and adds by steps the most important features until the value of a given performance metric decreases (Rodriguez-Galiano *et al.* 2018). Instead of a sequential search, GRRF was a wrapper that used a regularisation based on a grid search. In this sense, the GRRF model used different gamma and lambda values to obtain multiple feature subsets, enabling us to obtain multiple hard classification models and choose the one with the highest overall accuracy. On the other hand, ARF, AFR and AFGRRF used the same procedure as RF, FRF and GRRF, respectively, to obtain soft classification models. Nevertheless, ARF, AFR and AFGRRF maps are derived from the SR function, by reclassifying iteratively the class-conditional probabilities map for different affected area percentages (Figure 3) and choosing the best map for every method as that with the smallest affected area at a TPR higher than 90% threshold.

Three subsets were generated from the initial IL database to train and assess the method's performance: training (60%), test 1 (20%) and test 2 (20%) (Ng 2018). We used this percentage in order to maintain a reasonable number of test samples. Test 1 was used as an internal validation for GRRF and AFGRRF, and test 2 to compare GRRF and AFGRRF with other RF-based methods. The McNemar test was applied between the best map generated by each method (ARF, AFRF and AFGRRF) (Foody 2004) to evaluate whether the differences between model accuracies were significant. It should



Figure 3. Flowchart with the Random Forest (RF) based methods used in this study to benchmark the proposed Area and Feature Guided Regularised Random Forest (AFGRRF). The performance of our proposed method was compared to (i) standard RF; (ii) Area Random Forest (ARF); (iii) Feature Random Forest (FRF); (iv) Area Feature Random Forest (AFRF) and (v) GRRF. The hard classification methods (RF, FRF and GRRF) produce categorical maps, considering by default an arbitrary threshold value of 0.5 in the class conditional probability. The soft classification methods (ARF, AFRF and AFGRRF), which predict a class conditional probability, were assessed in terms of the smallest affected area at a true positive rate equal or higher than 90%.

be noted that, in this case the interpretation of the McNemar test is backwards when compared to traditional studies on the evaluation of new classifiers, where it expected that the algorithm significantly outperforms a baseline. We therefore formulated two

hypotheses: H_0) the models induced significant changes in the responses, i.e. the changes seen in the sampling were not due to chance; and H_1) the models did not induce significant changes in the responses, i.e. the changes observed in the sampling were due to chance. Results with statistical confidence above 95% were considered. Values lower than 1.96 in the McNemar test would imply that the maps are not significantly different (Foody, 2004), thus rejecting H_0 .

4. Experimental results

The GRRF and AFGRRF methods (Table 2) were used to build 100 models (all possible combinations between lambda and gamma values). The most accurate GRRF model obtained lambda and gamma values of 1 and 0.3, respectively, with an overall accuracy of 94.59%. The best AFGRRF model obtained lambda and gamma values of 0.9 and 0.2, respectively, with an overall accuracy of 93.62%. Models with higher lambda values and lower gamma values outperformed the rest (Figures 4(B,C)). Nevertheless, exclusively considering overall accuracy did not minimise the affected area. In this sense, GRRF and AFGRRF estimated an affected area of 29.67% (462.80 km²) and 19.00% (296.40 km²), respectively. Thus, AFGRRF reduces the affected area by 166.4 km² while accuracy is only reduced by 3.52% when compared to the GRRF method. The RF, ARF, FRF and AFRF methods obtained 91.49%, 86.67%, 92.85% and 89.28% of overall accuracy, and affected areas of 27.43% (427.90 km²), 26.00% (405.60 km²), 27.78% (436.20 km²) and 23.00% (358.80 km²), respectively (Table 3). This means that the AFGRRF method reduced the affected area by 131.5 km², 109.2 km², 139.8 km² and 62.4 km² compared to the other methods (see Table 3). Therefore, considering a SR above 90%, AFGRRF reduced the affected area without drastically decreasing overall accuracy compared to GRRF. Furthermore, the differences between models were subtle according to the spatial distribution of values on the maps (Figure 5). Furthermore, as we can see in Figure 5, the spatial distribution of the misclassified sites are similar for the hard and soft models even when the threshold condition was a SR above 90%. Therefore, the results showed a significant reduction in affected areas for the AFGRRF method without a significant impact on performance, which may improve management and reduces the costs associated to environmental monitoring and protection activities.

The McNemar test revealed no significant differences between models (see supplementary material, Table 2). The map produced using the AFGRRF method was significantly similar to those methods at a higher accuracy level (RF, FRF and GRRF). In this sense, a subtle decrease in accuracy could lead to a reduction in the affected area. It should be noted that, McNemar is interpreted differently in this case study. This test is commonly used to know whether a proposed method outperforms a reference method. This is a statistically significant increase in accuracy. However, in this case, the objective was to test whether AFGRRF, a wrapper that selects a subset of features that minimise the affected area, has a similar performance. This is a non-statistically significant decrease in accuracy. ARF was the only statistically different method; it was less accurate than the rest and less capable in reducing the affected area without a significant decrease in the accuracy level (see Table 3).

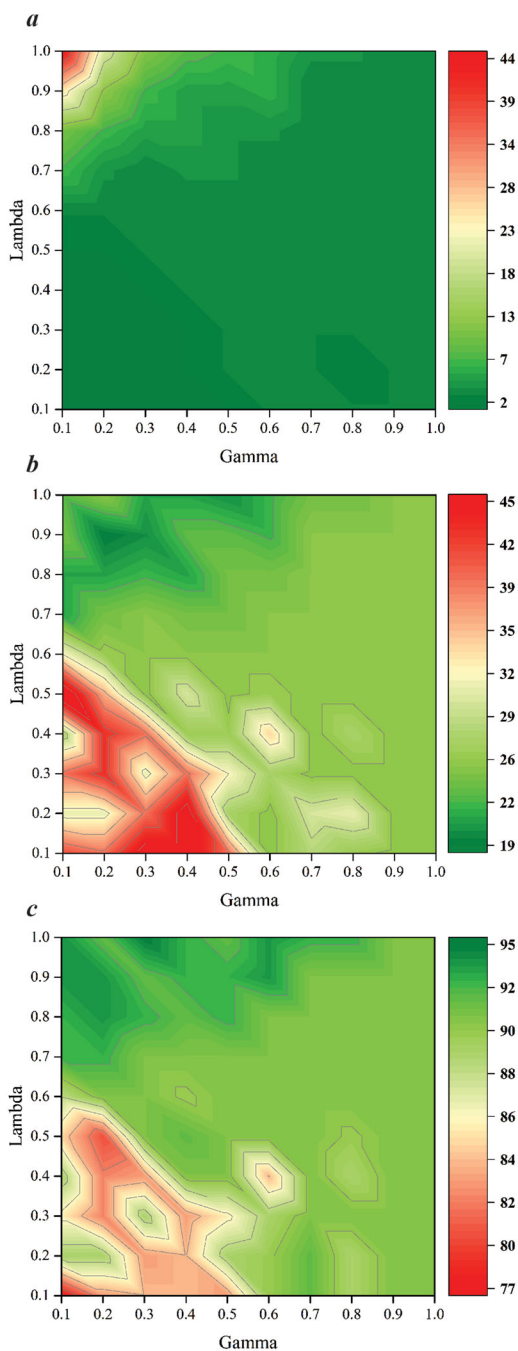


Figure 4. AFGRRF models for lambda and gamma: (a) Number of features selected; (b) Potential affected area for 90% success rate; (c) Overall accuracy.

For this study, where embedded and wrapper feature selection was applied, the selected features differed between methods in terms of number and typology (Table 4). The RF and AFR methods considered all features, while FRF and ARF

Table 2. List of acronyms.

AFGRRF	Area Feature Guide Regularised Random Forest	IL	Illegal Landfills
AFRF	Area Feature Random Forest	RF	Random Forest
ARF	Area Random Forest	ROC	Receiver Operating Curve
FRF	Feature Random Forest	SR	Success rate
GRRF	Guide Regularised Random Forest	TPR	True Positive Rate

Table 3. Overall results.

Method	Minimum affected area (%)	Minimum affected area (km ²)	Overall accuracy	Features selected
RF	27.43	427.9	91.49	113
ARF ^a	26	405.6	86.67	113
FRF	27.78	436.2	92.85	7
AFRC ^a	23	358.8	89.28	7
GRRF	29.67	462.8	94.59	11
AFGRRF ^a	19	296.4	93.62	12

^aMinimum affected area for success rate greater than 90%.

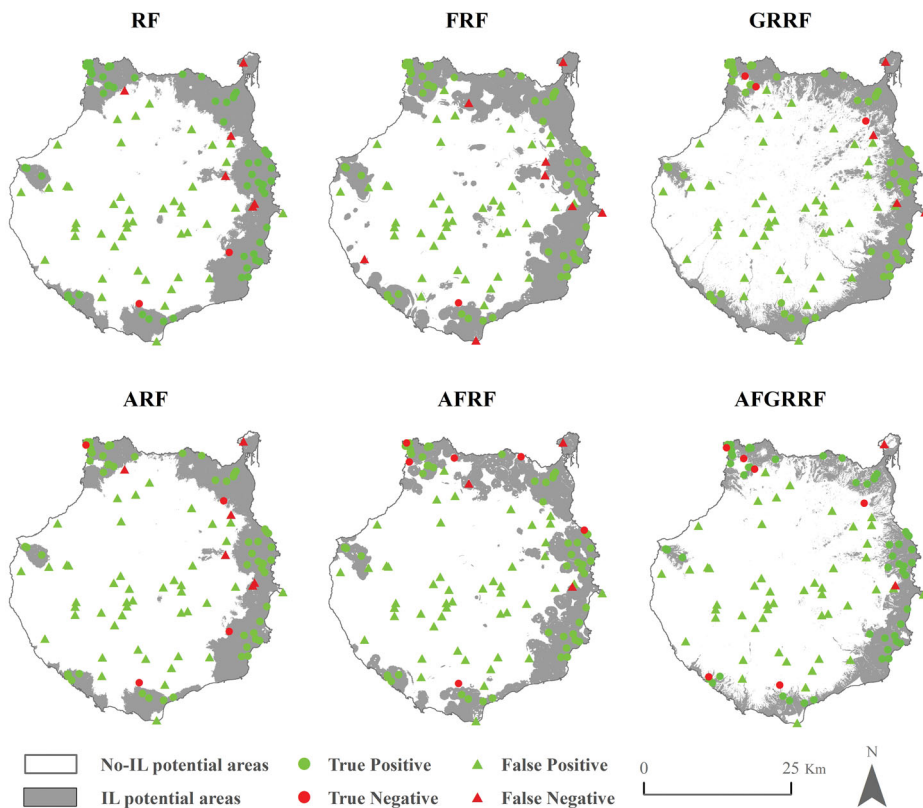


Figure 5. Map of illegal landfill potential occurrence for hard methods (RF: Random Forest; FRF: Feature Random Forest; GRRF: Guide Regularised Random Forest) and reclassified soft methods (ARF: Area Random Forest; AFRF: Area Feature Random Forest; AFGRRF: Area Feature Guide Regularised Random Forest).

Table 4. Feature selected by Random Forest (RF), Area Random Forest (ARF), Feature Random Forest (FRF), Area Feature Random Forest (AFRF), Guide Regularised Random Forest (GRRF), Area Feature Guide Regularised Random Forest (AFGRRF).

RF	ARF	FRF	AFRF	GRRF	AFGRRF
Whole feature space		Mining and extraction activity index Distance to pit zones Distance to transport infrastructures		Distance to coast Slope Distance to pit zones with different kernels Distance to element of interest	
		Land use transitions density from 1990 to 2000 Land use transition density from 1990 to 2012 Impervious cover transitions density from 1990 to 2012 Greenhouses density		Distance to educational equipment Distance to transport infrastructures	
				Visibility from the coastline Population density Industrial activity index Communication routes density	

methods selected seven features: mining and extraction activity index, distance to pit zones, distance to transport infrastructures, land use transitions density from 1990 to 2000, land use transition density from 1990 to 2012, impervious cover transitions density from 1990 to 2012 and greenhouses density (Quesada-Ruiz *et al.* 2018). The GRRF method selected eleven features: communication routes density, distance to pit zones with different kernels, distance to transport infrastructures, distance to element of interest, distance to educational equipment, distance to coast, visibility from the coastline (Gorr and Kurkand 2020) and population density. Finally, the AFGRRF method selected twelve features: buildings density, distance to transport infrastructures, distance to protected areas, distance to pit zones, distance to coast, distance to agricultural areas, distance to cultural equipment, slope, altitude, industrial activity index and population density. It should be noted that the affected area was larger for smaller feature subsets in AFGRRF (Figure 4).

Despite the selected features being different among methods, all of the methods considered proximity to the coast, agricultural areas, pit zones and transport infrastructures as important for IL occurrence, as in previous studies (Quesada-Ruiz *et al.* 2018, 2019b). Physiographic features were also particularly relevant, likely due to the rugged terrain of the island, especially for the proposed method, as shown by the Gini index values for the selected features (see [supplementary material, Figure 1](#)). This explains the visual similarities between the hard and soft maps (see [Figures 6 and 7](#)). The map in [Figure 6\(b\)](#) has a distinctive appearance because it takes into consideration the 'population density' feature. Feature selection-based methods obtained higher probability values for IL ([Figures 6\(b,c\)](#)). [Figure 5](#) shows how the methods without feature selection (RF and ARF) produced coarser maps, distinguishing the general patterns in affected areas, but unable to identify finer patterns further inland. In contrast, when feature selection was carried out (FRF, ARF, GRRF and AFGRRF) new affected areas were revealed, producing maps with finer spatial detail, especially in the case of GRRF and AFGRRF. Furthermore, methods that permit application of SR enabled spurious affected areas with lower probability to be filtered out (see [Figures 6 and 7](#)).

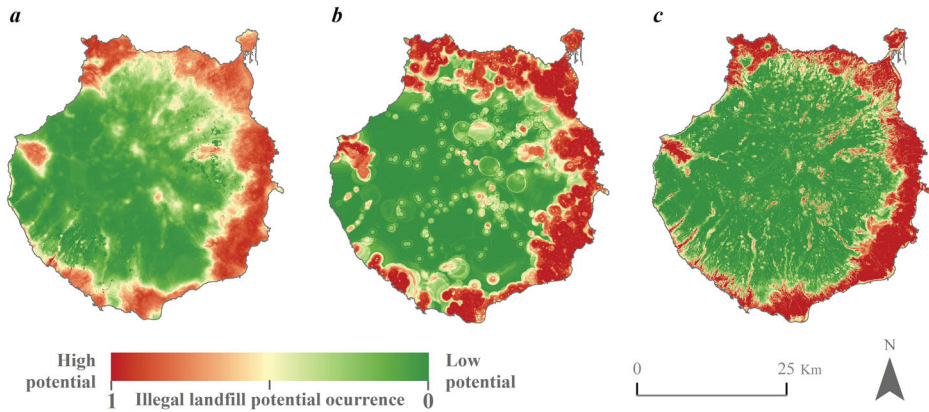


Figure 6. Map of illegal landfill occurrence probability: (a) Random Forest; (b) Feature Random Forest; (c) Guide Regularised Random Forest.

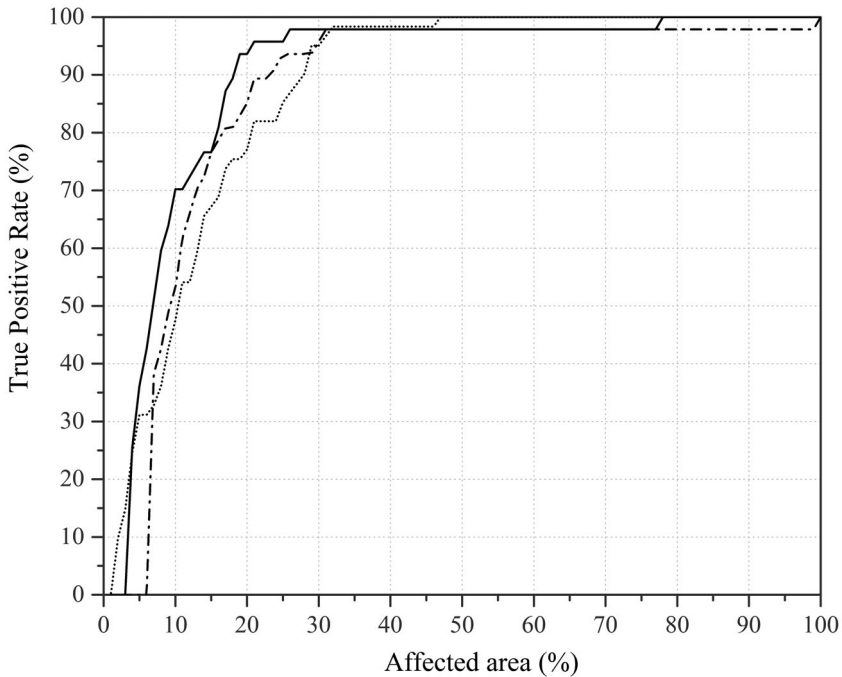


Figure 7. Success rate. Solid line (AFGRRF), dash-dotted line (ARF), and dotted line (AFRF).

5. Discussion

A majority of the studies focused on IL modelling applying weighted methods without feature extraction or feature selection, despite the accuracy of predictive modelling depending on feature selection, among other factors (Rodriguez-Galiano *et al.* 2018). Furthermore, these weighted methods usually rely exclusively on expert knowledge (Biotto *et al.* 2009, Matos *et al.* 2012, Chu *et al.* 2013) or data-driven approaches, such as Logistic Regression (Keser *et al.* 2012, Lucendo-Monedero *et al.* 2015) or

Discriminant Analysis (Quesada-Ruiz *et al.* 2019b). Other studies applied feature extraction, primarily Principal Component Analysis (Tasaki *et al.* 2007, Glanville and Chang 2015). In terms of GIS, expert knowledge methods are characterised by the combination and integration of multiple datasets. The intervention of an analyst with domain knowledge is thus indispensable to, for instance, determine the parameters of the method (Saaty 1980). In contrast, data-driven methods require less supervision to integrate multiple data layers to solve a geospatial problem.

While feature selection-based studies are scarce, there are a few examples of the application of embedded and wrapper algorithms using logistic regression with forward or backward search (Quesada-Ruiz *et al.* 2018, 2019b). Weighted methods build models that assign different importance to each feature of the feature space, considering subjective expert knowledge or filter-based, such as the Analytic Hierarchy Process of Saaty Method applied widely in GIS science (Saaty 1980). On the other hand, feature extraction and feature selection have less expert intervention than weighted methods, removing spurious or redundant features and reducing the feature space, either by combining the most relevant features or selecting them in an unbiased manner. Therefore, weighted methods consider the whole feature set, even when the statistical significance is low. Feature extraction allows removal of the least significant new features, but requires a subsequent selection process (selecting the most informative components based on the percentage of explained variance in Principal Component Analysis), with wrapper-based feature selection being the only fully automatable approach.

The definition of IL probability thresholds should be considered an important phase in the process of obtaining accurate binary/hard maps. Weighted methods (Biotto *et al.* 2009, Matos *et al.* 2012) and data-driven methods (Lucendo-Monedero *et al.* 2015, Quesada-Ruiz *et al.* 2019b) for mapping IL reclassified continuous values (i.e. between 0 and 1) consider a threshold value of 0.5. This threshold definition is arbitrary, as it relies on a symmetrical statistical distribution of probability values without considering spatial distribution or accuracy metrics. There are alternatives to arbitrarily choosing thresholds, such as analysis of ROC curves where a trade-off is sought between True Positive and False Negative rates to avoid overestimation and underestimation, respectively (Chu *et al.* 2013, Rodriguez-Galiano *et al.* 2014). The application of ROC is widely used in many scientific fields, such as bioinformatics, where a positive or negative diagnosis for certain diseases might be equally relevant (Beck and Shultz 1986). However, geoscience studies focusing on the spatial distribution of a binary phenomenon are different. Including negative cases for optimising threshold values could lead to underestimation of IL when negative occurrences are more frequent (i.e. there are more locations without IL than with IL). Therefore, our study or other spatially driven studies, such as landslides (Dahal *et al.* 2008, Hong *et al.* 2017, Chen *et al.* 2019) or mining (Carranza *et al.* 2008, Rodriguez-Galiano *et al.* 2015), focus on the positive cases. All of these studies are characterised by their interest in predicting a minimal area with the highest accuracy in positive cases, thus reducing costs associated with prospecting or monitoring. Hence, the method proposed not only could improve the delimitation of potentially affected areas by ILs but it could also facilitate the evaluation of the possible costs of recovery, or the implementation of dissuasive

and surveillance measures by minimizing the area (Quesada-Ruiz *et al.* 2019b). This paper proposes using SR as an alternative method to ROC for mapping binary problems, considering the TPR together with the area instead of the false positive rate. Figure 7 presents the results obtained from soft models, showing the percentage of cases correctly classified regarding affected area. SR allowed identifying AFGRRF as the model with smallest affected area for a TPR above 90%. SR also facilitated distinguishing affected areas, maximising the accuracy of positive occurrences while minimising the affected area (see Figure 5). The role of features to minimise the affected area was reinforced by using SR in a feature selection approach inside a wrapper. Modifying the GRRF algorithm to build a wrapper with SR as the accuracy metric may offer new methodological perspectives for feature selection when the phenomenon being studied has a binary behaviour, considering not just the overall accuracy metric but also the spatial criterion. Nevertheless, the application of AFGRRF has some limitations and requirements: (i) there must be a sufficiently large geospatial database with a large sampling size to assess and compare its application with respect to other feature selection methods; (ii) sampling must be separated into training, test 1 and test 2; (iii) an additional test (T1 in our case) is needed to optimise the affected area, that it is different from the test (T2 in our case) used to evaluate the overall accuracy of the models; (iv) a balanced sampling between negative and positive cases. In this sense, AFGRRF offers new perspectives for its application to other binary phenomenon such as: landslide prevention, flood prevention, ecosystem conservation, infectious disease or agricultural pest control. The sensitivity of the method to noise could also be studied, attending the errors in the positive or negative occurrences of the binary phenomenon, as well as its sensitivity to the reduction of the training data.

6. Conclusions

Predictive modelling of binary phenomena such as presence or absence focuses on the application of numerical methods to estimate the probability of occurrence of a phenomenon. This paper proposes a new method for feature selection that modifies the GRRF algorithm for use inside a wrapper, improving the mapping and modelling of binary phenomena and the accuracy of the affected area mapping to reduce environmental management costs of binary phenomena. AFGRRF addressed the 'Rashomon effect' or the multiplicity of good models. This new method, AFGRRF, uses a new metric for feature selection (SR), selecting the model built from a feature subset that minimises the affected area within multiple accurate models. This approach is an alternative to previously applied overall accuracy-based feature selection methods. Its novelty resides in selecting a feature subset that optimises both the True Positive Rate (TPR) and the potentially affected area using the SR. Hence, AFGRRF may offer new GIS methodological perspectives for feature selection in GIScience when the phenomenon being studied has a binary behaviour, considering not just the TPR metric but also the spatial criterion. In this sense, AFGRRF achieve to obtain a spatial distribution of the binary phenomenon without overestimation or underestimation consistent with respect to the most important explanatory features and allowing it replicability with certain stability. Probability maps are usually transformed into hard maps to facilitate

management actions. Hard maps are obtained using arbitrary thresholds that assume a symmetrical statistical distribution of probability values or other more sophisticated approaches, such as ROC. However, these approaches do not take into consideration the accuracy of the classifications with the extent of the affected or affected area. In this sense, geoscience studies are interested in predicting the minimal distribution area with the highest accuracy in positive cases in order to reduce the costs of prospecting or monitoring. Hence, our method proposed facilitated distinguishing affected areas, maximising the accuracy of positive occurrences while minimising the affected area and identifying the model with smallest affected area for a TPR above 90%. AFGRRF was tested on the predictive modelling of ILs in the Canary Islands. The performance of AFGRRF was compared to five different RF-based methods, showing the capability of AFGRRF to reduce the affected area without a drastic decrease in overall accuracy.

Author contributions

All authors contributed to the conceptualization of the work and devised the methods. Rodríguez-Galiano, V. together with Zurita-Milla, R. and Izquierdo-Verdiguier, E. supervised the work. All authors discussed the results. Quesada-Ruiz, L. and Rodríguez-Galiano, V. wrote most of the original draft. All authors reviewed and edited the manuscript.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

LQR is the holder of a Margarita Salas postdoctoral Fellow grant awarded by the Spanish Ministry of Universities [reference AH/20001]. The authors are grateful for the financial support given by the projects [RTI2018-096561-A-I00 and US-1262552], funded by 'Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación/FEDER - Junta de Andalucía (Consejería de Economía y Conocimiento)', respectively.

Notes on contributors

Lorenzo Carlos Quesada-Ruiz received the B.Sc. degree in geography from the University of Las Palmas de Gran Canaria, M.Sc. Geographical Information Sciences and Remote sensing from the University of Zaragoza, and Ph.D. in Geography from the University of Seville. He is Margarita-Salas postdoctoral fellowship at the University of Seville. His current research focuses on the spatial analysis applied to environmental problems.

Victor Francisco Rodríguez-Galiano received the B.Sc. degree in environmental sciences, the M. Eng. degree in geodesy and cartography and Ph.D. degree in remote sensing from the University of Granada, Spain. He is associate professor at the Department of Geography of the University of

Seville. His current research focuses on machine learning for addressing environmental problems and satellite derived land surface phenology and its validation with ground data.

Raul Zurita-Milla received the Agricultural Engineering degree from the University of Cordoba (Spain), and the M.Sc. and Ph.D. degrees in Geo-information Science and Earth observation from Wageningen University. He is full professor and head of the Geo-Information Department at the Faculty ITC of the University of Twente. His current research focuses on the use of data-driven approaches for modelling seasonal processes.

Emma Izquierdo-Verdiguier received the B.Sc. degree in physics and the M.Sc. and Ph.D. degrees in remote sensing from the University of Valencia, Spain. She is assistant postdoc in BOKU and a Google Developer Expert. Her research interests are the use of machine learning for Earth Observation data analysis and cloud computing environment for land surface monitoring.

Data and codes availability statement

The data and codes that support the findings of this study are available at <https://github.com/AFGRRF/Area-Feature-Guide-Regularised-Random-Forest>. The proposed AFGRRF code requires the following R libraries: RRF, Raster, Rgdal, and ROC written by others who are not affiliated with the research.

ORCID

Lorenzo Carlos Quesada-Ruiz  <http://orcid.org/0000-0001-7886-5678>

Victor Francisco Rodriguez-Galiano  <http://orcid.org/0000-0002-5422-8305>

Raúl Zurita-Milla  <http://orcid.org/0000-0002-1769-6310>

Emma Izquierdo-Verdiguier  <http://orcid.org/0000-0003-2179-1262>

References

- Akbari, M.A., and Rajabi, S.H.C. 2017. Landfill site selection by combining GIS and fuzzy multi criteria decision analysis, case study: Bandar Abbas, Iran. *World Applied Sciences Journal*, 3 (1), 39–47.
- Alexakis, D.D., and Sarris, A., 2014. Integrated GIS and remote sensing analysis for landfill siting in Western Crete, Greece. *Environmental Earth Sciences*, 72 (2), 467–482.
- Arabameri, A., et al., 2019. Science of the total environment GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *The Science of the Total Environment*, 658, 160–177.
- Bazi, Y., and Melgani, F., 2006. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44 (11), 3374–3385.
- Beck, J.R., and Shultz, E., 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology & Laboratory Medicine*, 110 (1), 13–20.
- Bhunja, G.S., et al., 2012. Localization of kala-azar in the endemic region of Bihar, India based on land use/land cover assessment at different scales. *Geospatial Health*, 6 (2), 177–193.
- Biotto, G., et al., 2009. GIS, multi-criteria and multi-factor spatial analysis for the probability assessment of the existence of illegal landfills GIS, multi-criteria and multi-factor spatial analysis for the probability assessment of the existence of illegal landfills. *International Journal of Geographical Information Science*, 23 (10), 1233–1244.
- Blum, A.L., and Langley, P., 1997. Artificial intelligence selection of relevant features and examples in machine. *Artificial Intelligence*, 97 (1–2), 245–271.
- Breiman, L., 2001a. Random Forest. *Machine Learning*, 45 (1), 5–32.
- Breiman, L., 2001b. Statistical modeling: the two cultures. *Statistical Science*, 16 (3), 199–231.

- Breslow, N.E., and Cain, K.C., 1988. Logistic regression for two-stage case-control data. *Biometrika*, 75 (1), 11–20.
- Bridges, O., Bridges, J.W., and Potter, J.F., 2000. A generic comparison of the airborne risks to human health from landfill and incinerator disposal of municipal solid waste. *The Environmentalist*, 20 (4), 325–334.
- Canela, M., Lora, M., and Estrella, , 2011. Cómo hacer una Regresión Logística binaria «paso a paso» (II): análisis multivariante. *Docuweb FABIS*, 34, 1–16.
- Carranza, E.J.M., Hale, M., and Faassen, C., 2008. Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. *Ore Geology Reviews*, 33 (3–4), 536–558.
- Cecchi, G., et al., 2009. Mapping sleeping sickness in Western Africa in A context of demographic transition and climate change. *Parasite*, 16 (2), 99–106.
- Chen, L., 2009. Curse of dimensionality. In: L. Liu and M.T. Özsu, eds. *Encyclopedia of database systems*. Boston: Springer.
- Chen, W., et al., 2019. Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China). *Bulletin of Engineering Geology and the Environment*, 78 (1), 247–266.
- Chu, T.-H., Lin, M.-L., and Shiu, Y.-S., 2013. Risk assessment mapping of waste dumping through a GIS-based certainty factor model combining remotely sensed spectral unmixing model with spatial analysis. In: Paper presented at the 7th international conference on renewable energy sources and the 1st international conference on environmental informatics, 2–4 April, Kuala Lumpur, 367–372.
- Chung, C.J., and Fabbri, A., 1999. Probabilistic prediction models for landslide hazard mapping. *Photogrammetric Engineering and Remote Sensing*, 65 (12), 1389–1399.
- Cruz, Y., et al., 2011. El turismo en Canarias. Fundación ed. Available from: http://www3.gobier-nodecanarias.org/aciisi/obidic/files/fyde_el_turismo_en_canarias.pdf
- Dahal, R.K., et al., 2008. Predictive modelling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights-of-evidence. *Geomorphology*, 102 (3–4), 496–510.
- Dash, M., and Liu, H., 1997. Feature selection for classification. *Intelligent Data Analysis*, 1 (3), 131–156.
- Demesouka, O.E., Vavatsikos, A.P., and Anagnostopoulos, K.P., 2014. GIS-based multicriteria municipal solid waste landfill suitability analysis: a review of the methodologies performed and criteria implemented. *Waste Management & Research*, 32 (4), 270–296.
- Deng, H., and Runger, G., 2013. Gene selection with guided regularized random forest. *Pattern Recognition*, 46 (12), 3483–3489.
- Dixon, B., 2005. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis. *Journal of Hydrology*, 309 (1–4), 17–38.
- Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), 861–874.
- Foody, G.M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, 70 (5), 627–633.
- Glanville, K., and Chang, H.C., 2015. Mapping illegal domestic waste disposal potential to support waste management efforts in Queensland, Australia. *International Journal of Geographical Information Science*, 29 (6), 1042–1058.
- GOBCAN, 2008. Residuos. *Informe de Coyuntura*, 2008, 168–185.
- GOBCAN, 2015. Generación & tratamiento de residuos en Canarias. *Gobierno de Canarias*. Available from: https://www.gobiernodecanarias.org/medioambiente/temas/residuos/mas_informacion/enlaces_y_documentos_de_interes/
- Gorr, W.L., and Kurkand, K.S., 2020. *GIS tutorial for ArcGIS desktop 10.8*. Redlands, CA: ESRI Press.
- Guyon, I., and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Hall, M., and Smith, L., 1997. Analysis of magnetic field effects on distributed heat sources in a nanofluid-filled enclosure by natural convection. *Journal of Applied Fluid Mechanics*, 9, 1175–1187.

- Harris, J.R., and Grunsky, E.C., 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers & Geosciences*, 80, 9–25.
- Hong, H., et al., 2017. Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in Xing Guo area (China). *Geomatics, Natural Hazards and Risk*, 8 (2), 1997–2022.
- Huettmann, F., et al., 2018. Use of machine learning (ML) for predicting and analyzing ecological and 'presence only' data: an overview of applications and a good outlook. In: G. Humphries, D.R. Magness, and F. Huettmann, eds. *Machine learning for ecology and sustainable natural resource management*. Cham: Springer International Publishing, 27–61.
- Ichinose, D., and Yamamoto, M., 2011. On the relationship between the provision of waste management service and illegal dumping. *Resource and Energy Economics*, 33 (1), 79–93.
- Iftimi, A., et al., 2015. Space-time airborne disease mapping applied to detect specific behaviour of varicella in Valencia, Spain. *Spatial and Spatio-Temporal Epidemiology*, 14–15, 33–44.
- INE, 2016a. Estadística del Padrón Continuo. Madrid: Instituto Nacional de Estadística. Available from: http://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710990.
- INE, 2016b. Movimientos turísticos en fronteras. Madrid: Frontur. Available from: <http://estadisticas.tourspain.es/es-ES/estadisticas/frontur/informesdinamicos/paginas/anal.aspx>.
- Izquierdo-Verdiguier, E., and Zurita-Milla, R., 2018. Use of guided regularized random forest for biophysical parameter retrieval. In: Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain. 22–27 July 2018, 5776–5779.
- Izquierdo-Verdiguier, E., and Zurita-Milla, R., 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 88, 102051.
- Keser, S., Duzgun, S., and Aksoy, A., 2012. Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. *Waste Management*, 32 (3), 359–371.
- Kumar, S., Yee, W.L., and Neven, L.G., 2016. Mapping global potential risk of establishment of *Rhagoletis pomonella* (Diptera: Tephritidae) using MaxEnt and CLIMEX niche models. *Journal of Economic Entomology*, 109 (5), 2043–2053.
- Leuenberger, M., and Kanevski, M., 2015. Extreme learning machines for spatial environmental data. *Computers & Geosciences*, 85 (B), 64–73.
- Lucas, L., and Jauzein, M., 2008. Use of principal component analysis to profile temporal and spatial variations of chlorinated solvent concentration in groundwater. *Environmental Pollution*, 151 (1), 205–212.
- Lucendo-Monedero, A.L., Jordá-Borrell, R., and Ruiz-Rodríguez, F., 2015. Predictive model for areas with illegal landfills using logistic regression. *Journal of Environmental Planning and Management*, 58 (7), 1309–1326.
- Matos, J., Oštir, K., and Kranjc, J., 2012. Attractiveness of roads for illegal dumping with regard to regional differences in Slovenia. *Acta Geographica Slovenica*, 52 (2), 431–451.
- Menció, A., and Mas-Pla, J., 2008. Assessment by multivariate analysis of groundwater-surface water interactions in urbanized Mediterranean streams. *Journal of Hydrology*, 352 (3–4), 355–366.
- Monteiro Santos, F.A., et al., 2006. Mapping groundwater contamination around a landfill facility using the VLF-EM method – a case study. *Journal of Applied Geophysics*, 60 (2), 115–125.
- Navin Lal, T., et al., 2006. Embedded methods In: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh, eds. *Feature extraction*. Berlin, Heidelberg: Springer Berlin, Heidelberg. Available from: https://doi.org/10.1007/978-3-540-35488-8_6
- Ng, A., 2018. Machine learning yearning. *GitHub*. Available from: <https://github.com/ajaymache/machine-learning-yearning>.
- Pal, M., and Foody, G.M., 2010. Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (5), 2297–2307.

- Porretta, D., et al., 2013. Effects of global changes on the climatic niche of the tick *Ixodes ricinus* inferred by species distribution modelling. *Parasites & Vectors*, 6, 271. Available from: <https://doi.org/10.1186/1756-3305-6-271>
- Poulos, D.E., et al., 2016. Distribution and spatial modelling of a soft coral habitat in the Port Stephens-Great Lakes Marine Park: implications for management. *Marine and Freshwater Research*, 67 (2), 256–265.
- Powers, D.M.W., 2007. *Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation*. Adelaide: School of Informatics and Engineering, Flinders University, Technical Report SIE-07-001.
- PPRC, 2016. *Ordinary meeting of council*. France: Port Pirie Regional Council.
- Quesada-Ruiz, L.C., Perez, L., and Rodriguez-Galiano, V., 2019a. Spatiotemporal analysis of the housing bubble's contribution to the proliferation of illegal landfills – the case of Gran Canaria. *The Science of the Total Environment*, 687, 104–117.
- Quesada-Ruiz, L.C., Rodriguez-Galiano, V., and Jordá-Borrell, R., 2019b. Characterization and mapping of illegal landfill potential occurrence in the Canary Islands. *Waste Management*, 85, 506–518.
- Quesada-Ruiz, L.C., Rodriguez-Galiano, V., and Jordá-Borrell, R., 2018. Identifying the main physical and socioeconomic drivers of illegal landfills in the Canary Islands. *Waste Management Research*, 36, 1049–1060.
- Rodriguez-Galiano, V.F., Chica-Olmo, M., and Chica-Rivas, M., 2014. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. *International Journal of Geographical Information Science*, 28 (7), 1336–1354.
- Rodriguez-Galiano, V.F., et al., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.
- Rodriguez-Galiano, V.F., et al., 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *The Science of the Total Environment*, 624, 661–672.
- Rodriguez-Galiano, V., et al., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). *Science of the Total Environment*, 476–477, 189–206.
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, 804–818.
- Saaty, T., 1980. *The analytic hierarchy process*. New York: McGrawHill.
- SCHILL, W., et al., 1993. Logistic analysis in casecontrol studies under validation sampling. *Biometrika*, 80 (2), 339–352.
- Şener, Ş., Sener, E., and Karagüzel, R., 2011. Solid waste disposal site selection with GIS and AHP methodology: a case study in Senirkent – Uluborlu (Isparta) Basin. *Environmental Monitoring and Assessment*, 173 (1–4), 533–554.
- Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. London, New York: Chapman and Hall, 175.
- Silvestri, S., and Omri, M., 2008. A method for the remote sensing identification of uncontrolled landfills: formulation and validation. *International Journal of Remote Sensing*, 29 (4), 975–989.
- Soares, A., and Pereira, M.J., 2007. Space–time modelling of air quality for environmental-risk maps: a case study in South Portugal. *Computers & Geosciences*, 33 (10), 1327–1336.
- Tasaki, T., et al., 2007. A GIS-based zoning of illegal dumping potential for efficient surveillance. *Waste Management*, 27 (2), 256–267.
- Tehrany, M.S., Pradhan, B., and Jebur, M.N., 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology*, 504, 69–79.

- Tuv, E., 2009. Feature selection with ensembles. *Artificial Variables, and Redundancy Elimination*, 10, 1341–1366.
- Uyan, M., 2014. MSW landfill site selection by combining AHP with GIS for Konya, Turkey. *Environmental Earth Sciences*, 71 (4), 1629–1639.
- Visser, H., and Nijs, T.D., 2006. The map comparison kit. *Environmental Modelling & Software*, 21 (3), 346–358.
- Wittmann, E.J., Mellor, P.S., and Baylis, M., 2001. Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe. *Revue Scientifique et Technique (International Office of Epizootics)*, 20 (3), 731–740.
- Zhang, L., et al., 2019. Classification and regression with random forests as a standard method for presence-only data SDMs: a future conservation example using China tree species. *Ecological Informatics*, 52, 46–56.