



Modelling and interpreting pre-evacuation decision-making using machine learning

Xilei Zhao^{a,*}, Ruggiero Lovreglio^b, Daniel Nilsson^c

^a Department of Civil and Coastal Engineering, University of Florida, United States

^b School of Built Environment, Massey University, New Zealand

^c Department of Civil and Natural Resources Engineering, University of Canterbury, New Zealand

ARTICLE INFO

Keywords:

Random forest
Machine learning
Interpretation
Nonlinearity
Human behaviour in fire
Pre-evacuation

ABSTRACT

The behaviour of building occupants in the first stage of an evacuation can dramatically impact the time required to evacuate buildings. This behaviour has been widely investigated by scholars with a macroscopic approach fitting random distributions to represent the pre-evacuation time, i.e. time from noticing the first cue until deliberate movement. However, microscopic investigations on how building occupants respond to several social and environmental factors are still rare in the literature.

This paper aims to leverage machine learning as a possible solution to investigate factors affecting building occupants' decision-making during pre-evacuation stage. In particular, we focus on applying interpretable machine learning to reveal the interactions among the input variables and to capture nonlinear relationships between the input variables and the outcome. As such, we use a well-established machine-learning algorithm—random forest—to model and predict people's emergency behaviour pre-evacuation. We then apply tools to interpret the black-box random forest model to extract useful knowledge and gain insights for emergency planning. Specifically, this algorithm is applied here to investigate the behaviour of 569 building occupants split between five unannounced evacuation drills in a cinema theatre. The results indicate that both social and environmental factors affect the probability of responding. Several independent variables, such as the time elapsed after the alarm has started and the decision-maker's group size, are presenting strong nonlinear relationships with the probability of switching to the response stage. Furthermore, we find interactions exist between the row number where the decision-maker sits and the number of responding occupants visible to her; the complex relationship between the outcome and these two variables can be visualized by using a two-dimensional partial dependence plot. An interesting finding is that a decision-maker is more sensitive to the proportion of responding occupants than the number of them; hence, the people sitting in the back are often responding more slowly than the people in the front.

1. Introduction

The behaviour of people in the initial stages of evacuation (i.e. pre-movement time) can have a significant impact on the time required to evacuate a building. Many investigations have been carried out to investigate this behaviour using fire evacuation drills and fire accident data as illustrated by [1–4]. Despite the large number of studies, the representation of this behaviour is often oversimplified in the most of the existing computer evacuation models [5].

From a modelling point of view, there are three modelling approaches to representing the pre-evacuation time in computer evacuation models [6]. The first approach focuses on the assignment of a pre-defined time to agents or a pseudo random number drawn from a

distribution. The second approach consists of the assignment of sequences of pre-evacuation actions having a specific duration for each agent. The final approach is to predict agent decision-making in accordance with different external and internal factors affecting their response. From an implementation point of view, the first two approaches are strongly affected by the model users who are supposed to provide these inputs to compensate for model omissions [5]. The third one provides a modelling solution to overcome the weakness of the first two approaches. However, this last approach might introduce the bias of the developers who need to select relevant factors affecting agent behaviour and algorithm to simulate it. Regardless of the advantages of the third approach, it is still rarely implemented in existing evacuation models [7].

* Corresponding author: 504 Weil Hall, 1949 Stadium Rd, Gainesville, FL 32611, USA.

E-mail addresses: xilei.zhao@essie.ufl.edu (X. Zhao), r.lovreglio@massey.ac.nz (R. Lovreglio), daniel.nilsson@canterbury.ac.nz (D. Nilsson).

<https://doi.org/10.1016/j.autcon.2020.103140>

Received 1 November 2019; Received in revised form 11 February 2020; Accepted 15 February 2020

Available online 26 February 2020

0926-5805/ © 2020 Elsevier B.V. All rights reserved.

To date, there are only a few studies that have proposed solutions to model pre-evacuation decision-making during the pre-evacuation time. One of the first attempts was done by Proulx and Hadjisophocleous [8] who develop a probabilistic sub-model simulating human response to information indicating the presence of a fire. A more recent approach is the Evacuation Decision Model proposed by Reneke [9] who proposed a differential equation model to predict the evacuation state of agents. This model was expanded by other authors. For instance, Retana and Spearpoint [10] adapted Reneke's model to represent 11 evacuation scenarios proposed in the New Zealand fire safety code. Lovreglio et al. [6] proposed a couple of binary logit model (or logistic regression model) to predict the decision to start investigating and evacuating while Lovreglio et al. [11] expanded the original model by Reneke and proposed an ordered logit solution to calibrate it. These random logit solutions have the advantages to identify factors affecting the decision-making process and predict such a decision. However, the possible linear or nonlinear trends of each factor on the prediction outcomes need to be specified by the modellers—a priori introducing a modeller's bias. As such, this may reduce the possibility to investigate the actual trends. To date, machine learning has shown the potential to reduce such a bias and investigate the actual impact of an independent variable on a dependent variable by applying machine-learning interpretation methods (e.g. [12–14]).

Several authors have used machine-learning-based solutions to simulate human evacuations (e.g. [15–18]). However, only a few studies have used machine-learning techniques to investigate evacuation behaviour on large-scale disasters, such as tsunami evacuations, and on small-scale disasters, such as building evacuations. Song et al. [19] used Markov Decision Process to study 'big' human population movements and evacuations during the Great East Japan Earthquake using GPS records from mobile devices of 1.6 million people. Wang et al. [20] investigated pedestrian movement in evacuation laboratory experiments using several machine-learning methods. Zheng et al. [21] implemented a multi-objective particle swarm optimization method to classify evacuee population based on a combination of demographic and environmental factors. Finally, two pioneering studies were carried out by Liu and Lo [23] and Lo et al. [22] to investigate people's pre-evacuation behaviour in Hong Kong building fires using Support Vector Machine and Artificial Network based Fuzzy Inference System respectively. However, these two studies focus only on human response in fires of high-rise domestic buildings in Hong Kong. As such, this scenario does not allow a comprehensive investigation of how social influences and different alarm cues affect pre-evacuation decision-making. Moreover, the data used in these previous studies is from interviews after the event. This data type does not allow researchers to have clear measurements of the timing required to take decisions.

In this study, we propose a new approach to investigating pre-evacuation decision-making by using a well-established machine-learning algorithm: random forest (RF). This algorithm was first introduced by Ho in 1995 and later extended by Breiman and Cutler [24]. RF is an ensemble learning method for classification and regression by constructing a wide range of decision trees when training and outputting the majority class (for classification) or mean prediction (for regression) of the individual trees [25]. On the one hand, RF can reduce the overfitting problem of the individual decision tree by using bootstrapping. On the other hand, compared to traditional statistical models (e.g., logit models), RF can enhance the predictive accuracy by automatically capturing the nonlinearities and interactions within the input data. RF has been applied in several recent studies to model and interpret people's travel mode choices, such as Hagenauer and Helbich [26] and Zhao et al. [13]; however, few work has modelled and explained people's emergency behaviour using interpretable machine learning. In this work, we use the RF model as well as machine-learning interpretation tools to investigate the factors affecting the decision to evacuate by using the data collected in the evacuation drill carried out in Sweden in a cinema-theatre by Bayer and Rejnö [27] and analyzed

by Nilsson and Johansson [28]. This allows us to investigate if selected factors affect the choice to respond to emergency scenarios and the possible nonlinear trends of these factors and the interactions between them.

2. Methodology

In this paper, we are primarily interested in two research questions: How to forecast decision makers' emergency behaviour more accurately than using simple random distributions? What factors and how they influence their choices? The paper aims to tackle these two questions by applying random forest (RF) and machine-learning interpretation tools. It is expected to accurately predict people's evacuation choices and to interpret how the contributing factors drive the decision-making.

2.1. Model assumptions

The following assumptions are made for modelling pre-evacuation decision-making:

- Occupants are assumed to have two different behaviour stages: normal stage (NS) and response stage (RS). Occupants are in NS when they are performing their pre-emergency activities. Occupants in RS are those who have responded to the emergency either investigating or evacuating. This assumption is based on the model proposed by Reneke [9] who showed that evacuation behaviours can be classified into different behavioural states.
- Occupants involved in evacuation behave rationally and their passages from NS to RS is ruled by a binary decision-making process. This assumption is in line with many experimental and theoretical studies showing that non-rational behaviour is extremely rare during emergencies [29]. Occupants make their decisions based on the available information and cues during the emergency following a series of steps: Perception, Interpretation, and Decision-Making [30]. As such, based on the interpreted information and cues, occupants can decide whether to pass from NS to RS.
- The decision-making process is influenced by both environmental (external) and occupant (internal) factors. This assumption implies that occupants' decision-making depends on the social and physical information they perceived, named external factors [28]. However, occupants characteristics (e.g. previous experience, physical and mental condition, and alertness) can play a key role since these internal factors can influence the way in which an occupant perceives, interpret and make a decision [30].
- The decision-making process used in this study is modelled by RF. Given the binary structure of the decision-making process, RF represents a suitable modelling approach which allows the investigation on how several internal and external factors affect occupants' decisions.

2.2. Problem formulation

This paper assumes the availability of a longitudinal dataset that records individual emergency behaviour through time. The data thus can be described as $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]$ is a vector of p variables for individual i and y_i is the response variable. The input variables usually include important information of the decision-maker and the social and physical environment around him/her. The response variable y_i is binary, representing two different behavioural stages: Value 0 indicates *normal stage* (NS) and value 1 indicates *response stage* (RS). In other words, $y_i \in \{0, 1\}$.

2.3. Pre-evacuation decision modelling

Applying machine-learning techniques to model people's choices is typically approached as a *classification* problem [12, 13, 26, 31].

Machine-learning algorithms try to “classify” or recognize people's choice patterns from the observed data. More precisely, the goal of machine learning is to learn a target function f that maps input variables \mathbf{x} to the output variable \mathbf{y} as

$$\mathbf{y} = f(\mathbf{x}|\theta), \quad (1)$$

where θ is the unknown parameter or hyperparameter vector for the machine-learning model.

As discussed in Liu et al. [32], machine-learning classifiers can be divided into two major categories, i.e., hard classification and soft classification. The *hard classification* aims to directly predict the target label (in this case, 0 or 1), while the *soft classification* predicts the conditional probabilities for different classes and outputs the predicted label with the highest probability. In this study, we consider each choice option independently and treat the choice modelling as a soft classification problem. By conducting soft classification, we will be able to estimate the choice probability of each option at the individual level, which provides much more information than a hard predicted label. That is to say, we are trying to estimate $g_k(\mathbf{x}_i|\theta) = P(y_i = k), k \in \{0, 1\}$, and the relationship between f and g can be summarized as:

$$f(\mathbf{x}_i|\theta) = \underset{k \in \{0,1\}}{\operatorname{argmax}} g_k(\mathbf{x}_i|\theta) \quad (2)$$

2.4. Random Forest formulation

RF is a widely-used machine-learning algorithm that has been applied to model people's choices (e.g. [12, 13, 26]). RF is among the most accurate general-purpose classifiers so far with the capability of handling high dimensional data [33]. RF is also sufficiently robust: the input variables for RF can be of any type (numerical, categorical, continuous, or discrete) and RF is insensitive to skewed distributions, outliers, missing values, and the inclusion of irrelevant variables [24]. In addition, RF requires fairly minor efforts in tuning hyperparameters (i.e., two major hyperparameters) and is usually not very sensitive in their values [25, 34]. It also needs relatively short training time [25]. More importantly, as a tree-based ensemble learning algorithm, RF is able to model complex nonlinear relationships between the input variables and the response variable and capture high-order interactions among variables due to its flexible modelling structure [24]. In recent years, RF has shown to be effective in various fields. To name a few instances, the RF has been successfully applied to predict construction injury [35], model travel mode choice [13, 26], forecast Alzheimer's disease [36], detect credit card fraud [37], and classify earthquake building damage [38]. Therefore, in this paper, we choose to use RF to model people's pre-*evacuation* decision-making.

Before explaining RF, we first introduce decision trees upon which RF is built. One of the most popular algorithms to build a decision tree is called the classification and regression trees (CART). Here, we mainly focus on CART. For a classification problem, the CART model builds classification trees to predict a categorical dependent variable. In particular, the CART model creates classification trees where each internal node of the tree recursively partitions the data based on the value of a single predictor. Leaf nodes represent the category (i.e., NS or RS) predicted for that occupant [39]. The tree structure is ideal for capturing interactions between variables in the data [40] and is essentially a nonlinear mapping of \mathbf{x} to \mathbf{y} . However, the decision trees are often unstable and easy to overfit the data.

To tackle the issues of CART, the tree-based ensemble techniques were proposed to form more robust, stable, and accurate models than a single decision tree [41, 42]. One of the ensemble methods is RF. It trains multiple decision trees in parallel by bootstrapping the training data, i.e., sampling with replacement [24]. When training the trees, RF selects a random subset of all the input variables. More precisely, the trees in RF use all the variables, but every node in each tree only uses a random subset of them. By doing so, RF can overcome the overfitting

problems of a single decision tree, and reduce variance between correlated trees. RF makes (hard) choice predictions by determining the majority voting among all the classification trees. In this study, we predict the class probabilities by using the proportion of votes for different classes, see Fig. 1. More formally, the probability of choosing $k, k \in \{0, 1\}$ for individual i can be indicated as

$$g_k(\mathbf{x}_i|\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B I_0(\hat{T}_b), \quad (3)$$

where $\hat{\theta}$ is the estimated hyperparameter¹ vector for RF, B is the total number of decision trees in RF (one of the hyperparameters), \hat{T}_b is predicted class label for decision tree $b, b = 1, \dots, B$, and $I_0(\hat{T}_b)$ is an indicator that equals to 1 if $\hat{T}_b = 0$.

In this paper, the R package *randomForest* is used to train RF models [44].

2.5. Model interpretation

Interpretable or explainable machine learning has become increasingly more important in the broad field of machine learning [40]. The machine-learning interpretation tools can be roughly divided into two major categories, including model-specific and model-agnostic. Specifically, *model-specific* tools are designed for specific types of machine-learning models, while *model-agnostic* methods can be applied to any machine-learning models of choice [45]. Therefore, model-agnostic tools are usually more flexible and applicable, and, more importantly, may offer a consistent criteria for various machine-learning models.

In the paper, we mainly apply two widely-adopted model-agnostic methods, i.e., variable importance and partial dependence plots [46], to explain individual pre-*evacuation* decision-making.

2.5.1. Variable importance

Variable importance measures the significance of each variable with respect to its influence on predicting the response variable. The more a model depends on a variable to make predictions, the more important it is for the trained model. For the RF model, there are two different approaches to compute variable importance.

The first approach is computed by permuting the out-of-bag (OOB) samples, in order to measure the prediction strength of each variable [24]. For each decision tree in the RF model, the OOB sample data are passed down the tree to generate a predictive accuracy. The same process is conducted after permuting an input variable. The difference between the two values (predictive accuracy) is averaged over all the trees, which is used as a measure of the variable importance for this variable. Even though this approach was first proposed specifically for RF by Breiman [24], the core idea (i.e., the variable importance is the decrease in prediction error after permuting the variable's values) was extended to a model-agnostic version of the variable importance by Fisher et al. [47].

The second approach is conducted by computing the mean decrease in node impurities (usually measured by Gini index for classification problems) from splitting on a variable. However, this approach is specific to tree-based models, such as decision trees, RF, and boosting trees.

The relative variable importance plot constructed using the first approach will be presented in Subsection 3.3.1.

2.5.2. Partial dependence plots

Partial dependence plots (PDPs), as one of the most popular model-agnostic tools for interpretable machine learning, can graphically reveal the relationship between the input variable(s) and the predicted

¹ Hyperparameters are parameters whose values are set before the learning process begins and they can be tuned to control the behaviour of a machine learning algorithm [43].

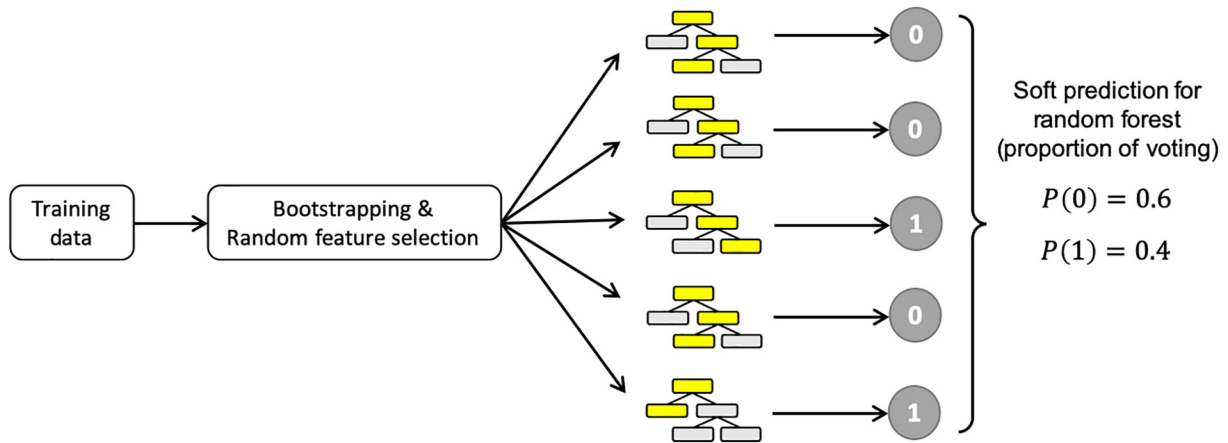


Fig. 1. Illustration of random forest model with soft prediction mechanism.

class probabilities [46]. Furthermore, as discussed in Zhao and Hastie [48], PDPs may be used to draw causal inference if these three conditions are satisfied:

- A good predictive model;
- Domain knowledge about the causal structure;
- Visualization tools such as PDPs.

To properly define PDPs in our application, suppose we want to evaluate the impact of x_S , $S \subset \{1, \dots, p\}$, on the soft prediction outcomes (i.e., choice probabilities), and let C be the complement set of S . Given the choice probability output g_k , $k \in \{0, 1\}$ of the RF model, the partial dependence of g_k on x_S is defined as

$$g_{kS}(x_S) = \mathbb{E}_{x_C} g_k(x_S, x_C | \theta). \quad (4)$$

In practice, Eq. (4) can be estimated by

$$\hat{g}_{kS}(x_S) = \frac{1}{N} \sum_{i=1}^N g_k(x_S, x_{C_i} | \hat{\theta}), \quad (5)$$

where $x_{C_i} (i = 1, \dots, N)$ is the values of x_C for each instance in the training set. A PDP reveals that, for the given value(s) of feature(s) S , what the average marginal effect on the predicted choice probability is.

In many previous studies, PDPs can be used to readily reveal the nonlinear relationships between the input variable(s) and the response variable for black-box machine-learning models, such as RF (e.g. [13, 49, 50]). Note that PDPs do not require any effort to assume the linear/nonlinear trends between input variables and the response variable by the modellers, which is needed by the logit models.

3. Case study

The case study uses the behavioural data collected during unannounced evacuation drills carried out in a cinema theatre in Sweden by Bayer and Rejnö [27]. They conducted 18 experiments in the same cinema theatre to investigate the impact of different alarm systems on the pre-evacuation time. Each participant only took part once at the experiments to avoid bias on the response to the alarm. The geometry of the environment is illustrated in Fig. 2; there are two exits, one in the front and the other in the back.

3.1. The data

In this work, we use the data collected during experiments aimed at testing the impact of different types of alarm [28]. These drills were selected by the authors since they have the highest number of participants. These data were previously used by Lovreglio et al. [6] to identify the behavioural state of the participants based on the types of

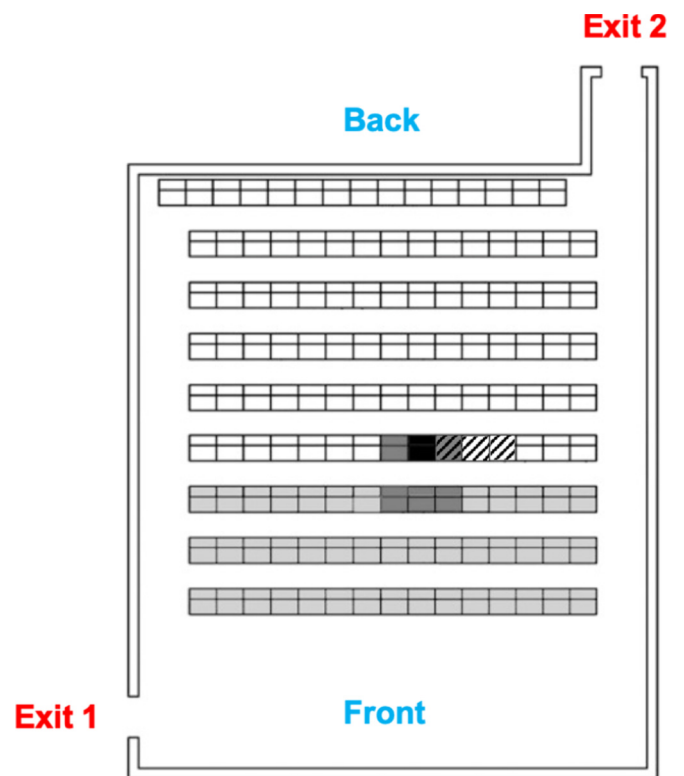


Fig. 2. Geometry of the cinema theatre. The grey squares (including light and dark grey squares) represent the occupants visible to the participant under investigation (i.e. decision-maker). The dark grey squares are the occupants considered close to the decision-maker. The squares with diagonal lines represent the occupants belonging to the decision-maker's personal group.

behaviour identified by Nilsson and Johansson [28] in their video analysis. Each participant's state was identified for every relevant event. The relevant event of a participant is his/her decision to respond to the emergency and the change of state of at least one of other visible participants or belonging to his/her personal group (i.e. the group a participant is attending the movie with). Fig. 2 provides a visual definition of visible participants as well as personal group. In this work, we did not investigate the hearing interactions (i.e. participants can hear people behind them) as it was not possible to measure such an interaction using the CCTV videos. A new record was added whenever there was any relevant event. Each record also includes environmental data concerning both social (e.g. the number of other visible occupants in normal and response states) and physical environmental factors (e.g.

Table 1
Description of all explanatory variables.

Name	Description
AlarmType	Dummy variable equal to 1 if the alarm system is Alarm Bell and 0 if Pre-Recorded Message ^a
Time	Time elapsed after the alarm has started
VisDM_NS	Total number of occupants visible to the decision-maker who are in NS
VisDM_RS	Total number of occupants visible to the decision-maker who are in RS
VisDMClose_NS	Total number of occupants visible to the decision-maker who are close to her and are in NS
VisDMClose_RS	Total number of occupants visible to the decision-maker who are close to her and are in RS
DMPerGroup_NS	Total number of occupants belonging to decision-maker's personal group who are in NS
DMPerGroup_RS	Total number of occupants belonging to decision-maker's personal group who are in RS
GroupSize	The decision maker's personal group size (i.e. number of occupants belonging to the decision maker's personal group)
Row	Number of the row where the decision-maker is sitting
Seat	Number of the seat where the decision-maker is sitting

^a The content of the message translated from the original Swedish version is "Important message, important message. There is a fire in the building. We ask everyone to head for the nearest exit and gather outside the building."

the time elapsed from the beginning of the alarm). There are 5240 data points in total.

In the revealed-preference data, a decision-maker can have two different behavioural stages, including *NS* and *RS*. Therefore, the response variable has two different class labels, i.e., 0 (*NS*) and 1 (*RS*). The independent variables used in this case study are illustrated in Table 1. These variables include measurements related to social factors, i.e. the behaviour of other visible evacuees and personal group, which are fundamental in the early stage of an evacuation as observed by Nilsson and Johansson [28], Lovreglio et al. [6], and Lovreglio et al. [11]. Moreover, the independent variables include the position of the decision-maker in the room as this can be another relevant factor as observed by Galea et al. [51]. As such, this represents the first attempt to investigate the combined effect of these two factors on the pre-evacuation decision process.

In the data pre-processing stage, we first use the variance inflation factor (VIF) to evaluate the multicollinearity among all the independent variables shown in Table 1. VIF provides estimates how much the variance of an estimated regression coefficient is increased because of multicollinearity. After removing all highly correlated ones (i.e., *VisDM_NS* and *DMPerGroup_NS*), the remaining variables all have VIF less than five, which is a common threshold for determining multicollinearity [52]. Hence, there are nine independent variables included for modelling pre-evacuation decision-making. The descriptive statistics of the nine independent variables and the dependent variable are shown in Table 2.

Table 2
Descriptive statistics for independent and dependent variables used for modelling.

Variable	Category	%	Mean	SD	Min	Max
<i>Dependent variable</i>						
Response choice	NS (0)	79.90				
	RS (1)	20.10				
<i>Independent variable</i>						
Time (s)			18.23	6.72	2.00	60.00
GroupSize			2.65	0.98	1.00	5.00
Row			6.02	2.15	1.00	9.00
Seat			7.90	4.20	1.00	15.00
AlarmType	Pre-Recorded Message (denoted by 0)	60.78				
	Alarm Bell (denoted by 1)	39.22				
VisDM_RS			20.56	21.28	0.00	110.00
VisDMClose_NS			3.06	1.69	0.00	6.00
VisDMClose_RS			1.44	1.54	0.00	6.00
DMPerGroup_RS			0.48	0.78	0.00	4.00

3.2. Hyperparameter tuning for random forest

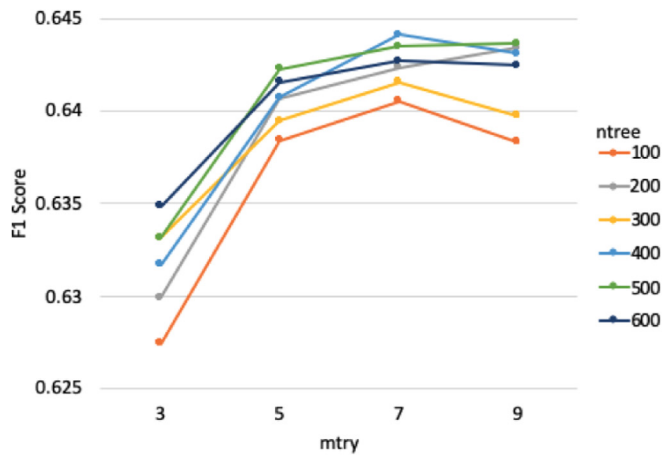
For many nonparametric machine-learning models, such as RF, we need to tune hyperparameters for it, where a hyperparameter is a parameter with a pre-determined value before training the machine-learning model. Hyperparameter tuning is to choose a set of optimal hyperparameters for a machine-learning model. The RF model mainly depends on two hyperparameters, i.e., the number of decision trees to grow (denoted by *ntree* here) and the number of variables randomly sampled as candidates at each split of the decision tree (denoted by *mtry* here). Note that we randomly split the entire dataset into two disjoint subsets, 90% for training and 10% for testing. The final random forest model will be evaluated using the separate testing set and the results will be presented in Subsection 3.4.

In this study, we apply grid search, a widely-adopted method for hyperparameter tuning. It searches exhaustively via a specified subset of hyperparameters. Here, we first define the search space, i.e., *ntree* = 100, 200, 300, 400, 500, 600 and *mtry* = 3, 5, 7, 9, so there are 24 different combinations of *ntree* and *mtry*. Then, these 24 candidate RF models are compared via repeated 10-fold cross validation (repeated for 10 times). Cross validation is a common approach to evaluating the predictive performance of different machine-learning models. A widely-applied cross-validation method is called 10-fold cross validation (which is used in our paper). It follows: 1) Randomly split the entire training set into 10 disjoint equal-sized subsets; 2) choose one subset for validation, the rest for training; 3) train all the machine-learning models on one training set; 4) test all the trained models on the validation set and compute the performance metric (e.g., predictive accuracy, F1 Score, and area under the curve [AUC]); 5) repeat Step 2) to 4) for 10 times, with each of the 10 subsets used exactly once for validation; and 6) the 10 validation results for each model are averaged to produce a mean estimate. If the aforementioned procedure is repeated for multiple times (in each repetition, the folds are split in a different way), it is called repeated 10-fold cross validation. The mean estimates from all repetitions are then averaged to obtain a final estimate. Repeated 10-fold cross validation aims to produce a more robust model performance estimate than a single 10-fold cross validation.

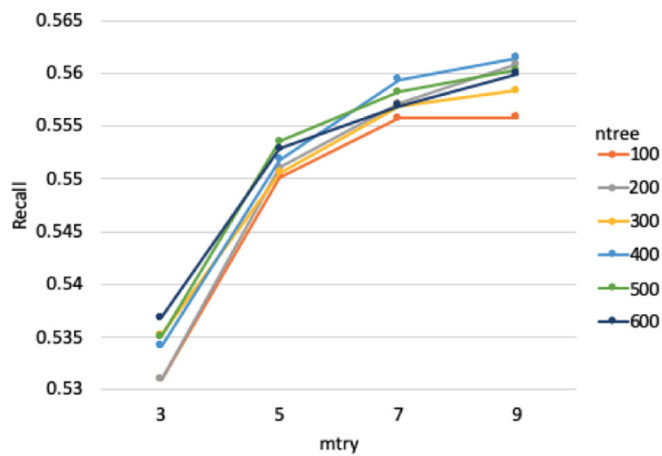
In this paper, F1 Score is used as the model performance metric to determine the best hyperparameters for RF, which is defined as

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

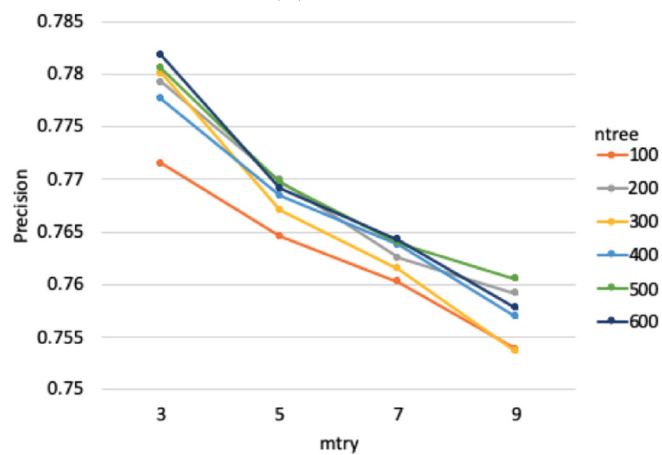
where Precision is the ratio of correctly predicted positive instances to the total predicted positive instances and Recall is the ratio of correctly predicted positive instances to all the instances in actual positive class. According to Eq. (6), F1 Score is the weighted average (more precisely, harmonic mean) of Precision and Recall. Higher F1 Score indicates better model performance. Even though F1 Score is not as intuitive as accuracy, it provides more useful information regarding the predictive



(a) F1 Score



(b) Recall



(c) Precision

Fig. 3. Hyperparameter tuning results for random forest.

capability of the machine-learning models, especially when the class imbalance problem exists.

Fig. 3 displays the hyperparameter tuning results in terms of F1 Score, Recall and Precision. More specifically, for F1 Score (see Fig. 3a), among the 24 RF candidate models, the best-performing RF model has $n_{tree} = 400$ and $m_{try} = 7$, whose F1 Score is equal to 0.644. For Recall (see Fig. 3b), the best-performing RF model has $n_{tree} = 400$ and m_{try}

$= 9$. For Precision (see Fig. 3c), the best-performing one has $n_{tree} = 600$ and $m_{try} = 3$. Different metrics show different aspects of the model performance. One could argue that Precision is a more important metric than Recall for fire evacuation, as the stakes are higher for false positive (i.e., incorrectly predicting that an occupant is changing to RS [1] when she is not): low Precision relates to high false positive, potentially leading to underestimating the loss of life. However, we also want to point out that if a model has low Recall, then we may encounter the situation that the majority of the true evacuees cannot be correctly predicted, which will make the interpretations (e.g., what key factors drive people to evacuate and their nonlinear relationships and interactions) provided by variable importance and PDPs less trustworthy. Therefore, a model with high Recall and high Precision will be optimal, and F1 Score seems to be a better metric that takes both of them into account. In this paper, we decide to choose the best model mainly based on F1 Score (i.e., the model with $n_{tree} = 400$ and $m_{try} = 7$). This model has one of the best Recall values among the 24 candidate models and its Precision is only around 2% worse than the model with the best Precision. Therefore, we fit the final RF model with $n_{tree} = 400$ and $m_{try} = 7$ using the entire training set (i.e., 90% of the entire dataset).

3.3. Model interpretation

3.3.1. Variable importance

The variable importance plot of the final RF model is shown in Fig. 4. It indicates the impact of each variable on model's predictive performance. As discussed in Subsection 2.5.1, the variable importance presented here is computed using the first approach (permuting the OOB samples).

Fig. 4 shows that where the decision-maker is sitting has the most important influence on her choice to respond to the emergency (Row: 16.34%; Seat: 15.46%). This is consistent with the findings shown in Galea et al. [51]. The signal of the alarm system comes third in terms of variable importance (13.12%). The fourth important variable is the number of occupants visible to the decision-maker who are in RS (11.64%). This indicates that also social influences have a key role in the decision-making progress as illustrated previously by Nilsson and Johansson [28]. The time elapsed after the alarm and the number of occupants visible to the decision-maker who are close to her and are in NS come fifth and sixth, with similar variable importance (10.58% and 10.32%). The seventh to ninth important variables are GroupSize (8.69%), DMPerGroup_RS (7.62%) and VisDMClose_RS (6.23%).

3.3.2. Partial dependence plots

The partial dependence plots (PDPs) for all the input variables are illustrated in Fig. 5. They represent the relationships between the independent variables and the choice probability of RS (i.e. the probability of responding to the emergency).

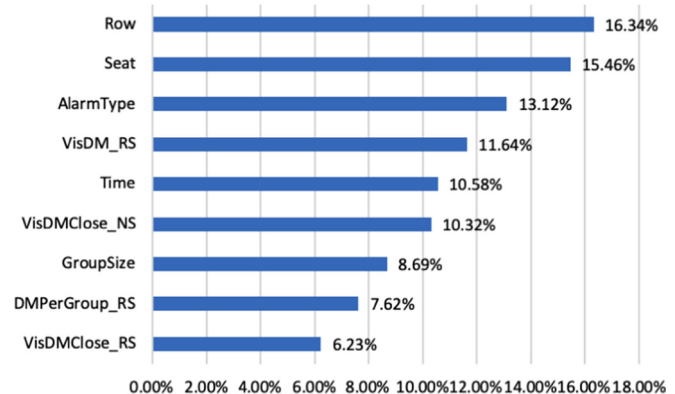


Fig. 4. Variable importance plot for random forest: Variable importance is scaled to sum up to 100%.

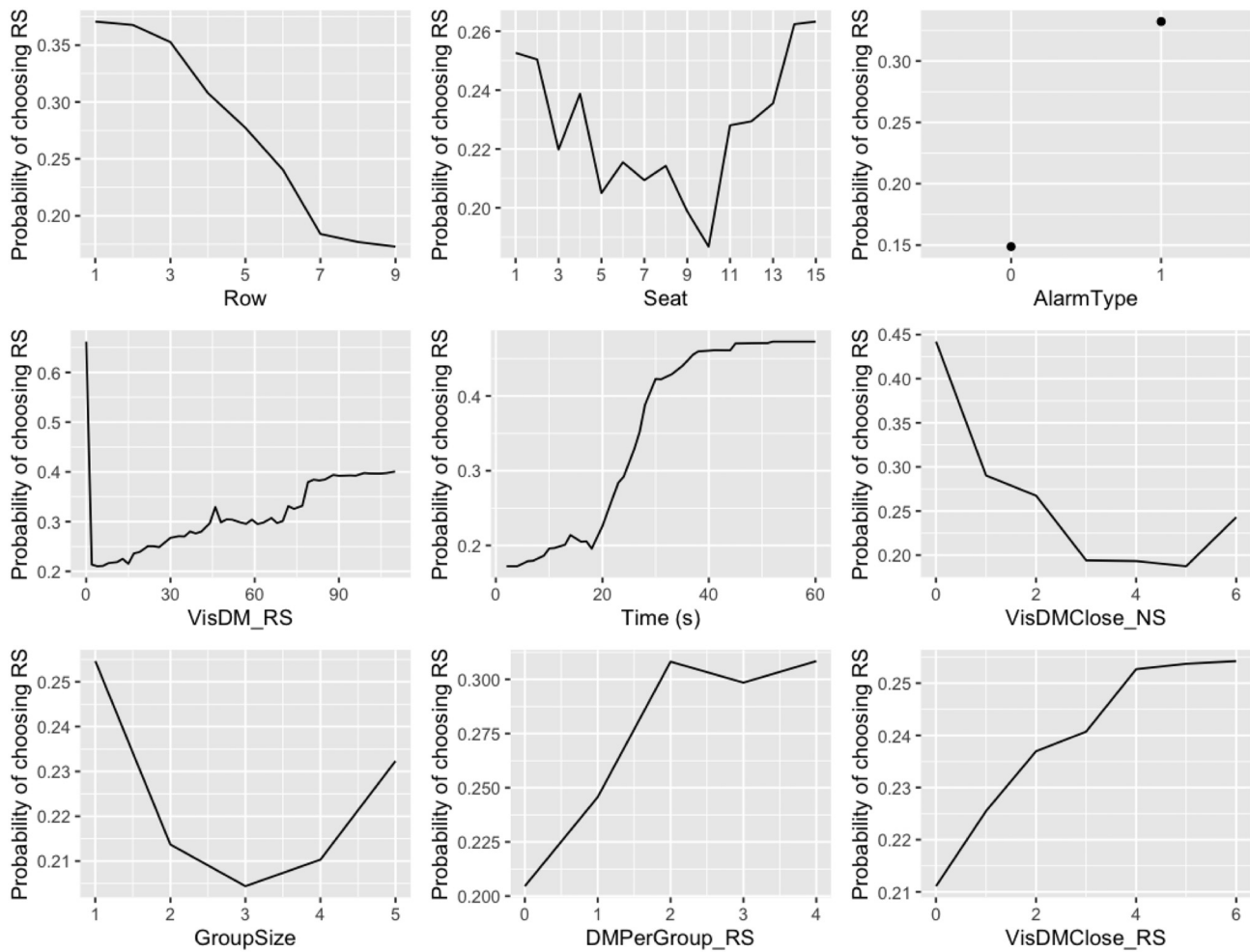


Fig. 5. Partial dependence plots for all independent variables.

The PDP for Row indicates that people who sit in the back are less likely to evacuate. However, one may argue that the people in the back might be as likely to respond as the people in the front, and they “passively” see more other people responding due to their row location. In other words, there might exist interactions between Row and VisDm_RS. This aspect will be discussed in detail later in this Subsection. The PDP for Seat shows that people who sit in the middle are less likely to respond to the emergency compared to people who sit close to the aisle. Galea et al. [51] argue that “This may be due to the expectation that those located towards the middle of the row will be unable to move before those closer to the perpendicular exit aisle have moved.”

The PDP for AlarmType illustrates that the pre-evacuation response for the alarm bell is much quicker than the pre-recorded message. This was because everyone waited for the voice alarm to finish. In the case of the alarm bell, someone responded quickly and the rest followed their behaviours in line with the social influence theory. In terms of the PDP for Time, it illustrates that the time from the start of the alarm has impact on this choice with a S-shaped trend. This indicates that the time has a low impact at the beginning of the emergency. The time starts having a strong impact after 20 s from the start of the emergency as it increases the choice probability of nearly 0.3. For GroupSize, the relationship takes a U-shape, with the choice probability reaching the minimum at the group size of three people. When the decision-maker is by herself (i.e., group size of one person), the probability of choosing RS reaches the maximum. This result may also be linked to social influence. When a person was alone (i.e. personal group equal to one), she does not need to check with others whether they respond or not. If there

are many in a personal group, someone is likely to respond more quickly, which influences everyone else in the group. The impact of group size on the pre-evacuation delay was also observed by Galea et al. [51]. In their study of a theatre evacuation, they observed four evacuees belonging to the same social group taking some time in an apparent discussion deciding what to do and to assist each other.

From the PDPs of VisDMClose_NS and VisDMClose_RS, we can observe that with more occupants visible to the decision-maker who are close to her and are in NS, she is less likely to respond; in contrast, with more occupants visible to the decision-maker who are close to her and are in RS, she is more likely to respond. This indicates how the decision is affected by the social surrounding of the decision-maker. This observation is consistent with the results provided by Nilsson and Johansson [28]. The PDP of DMPerGroup_RS shows that with more occupants belonging to the decision-maker’s personal group who are in RS, she is more likely to respond to the emergency.

However, the PDP of VisDM_RS shows a seemingly unrealistic outcome. More specifically, the relationship between the choice probability and VisDM_RS is reasonable after 0, but the value at 0 is unexpectedly large (0.662). This is counter-intuitive, as one may hypothesize that when the decision-maker sees no one is responding to the emergency, she is more likely to wait instead of entering RS. The unreasonable relationship revealed by PDP leads us to explore the patterns of the observed data with VisDM_RS = 0. In Table 3, we present the descriptive statistics for two subsets and the entire training set, where Subset 1 represents the collection of observations with VisDM_RS = 0 and Subset 2 represents the collection of observations with VisDM_RS = 2. As highlighted in Table 3, the percentage of

Table 3
Descriptive statistics for different subsets with comparison of training set.

Variable	Category	Subset 1 ^a	Subset 2 ^b %	Training set	Subset 1		Subset 2		Training set	
					Mean	SD	Mean	SD	Mean	SD
Response	NS (0)	30.00	87.78	79.86						
	RS (1)	70.00	12.22	20.14						
Row					2.28	1.40	5.27	2.45	6.01	2.15
Seat					8.15	4.02	7.63	4.18	7.91	4.20
AlarmType	0	75.00	67.20	60.96						
	1	25.00	32.80	39.04						
VisDM_RS					0.00	0.00	2.00	0.00	20.52	21.32
Time					17.03	5.75	13.76	6.51	18.25	6.65
VisDMClose_NS					2.85	1.44	3.70	1.66	3.06	1.69
GroupSize					3.10	0.93	2.72	1.01	2.66	0.98
DMPerGroup_RS					0.43	0.59	0.29	0.68	0.49	0.78
VisDMClose_RS					0.00	0.00	0.47	0.76	1.44	1.54

^a Subset 1: A subset of the training set with VisDM_RS = 0.

^b Subset 2: A subset of the training set with VisDM_RS = 2.

choosing RS for Subset 1 (70.00%) is significantly higher than that for Subset 2 (12.22%) and that for the training set (20.14%). In addition, observe that the mean of Row for Subset 1 (2.28) is significantly lower than that for Subset 2 (5.27) and that for the training set (6.01). Other independent variables of Subset 1 all have the mean values within 1 standard deviation of the training set. Therefore, it is very likely that Subset 1 is biased in terms of row numbers: most decision-makers who belong to Subset 1 (i.e., seeing no one is responding to the emergency) are sitting in the front. We further hypothesize that there may exist interactions between Row and VisDM_RS.

Therefore, we also plot a two-dimensional PDP to illustrate the potential interactions between variables and to explore how they affect the predictions. Fig. 6 reveals the relationship between RS probability and the interaction of VisDM_RS and Row, both of which take on values within the convex hull of their training values. The results clearly indicate that decision-makers who sit in the back and see few occupants entering RS are least likely to respond. Such a trend is visible on the upper left corner of Fig. 6. Another important observation is that decision-makers who lie in the diagonal of Fig. 6 are most likely to respond first, regardless of their row numbers. Furthermore, we find that with seeing the same number of people responding, people in the back are less likely to respond than people in the front. It shows that decision-makers may be more sensitive to the proportion of occupants visible to them who are responding than the number of occupants visible to them who are responding.

In summary, this analysis highlights that the response of a decision-

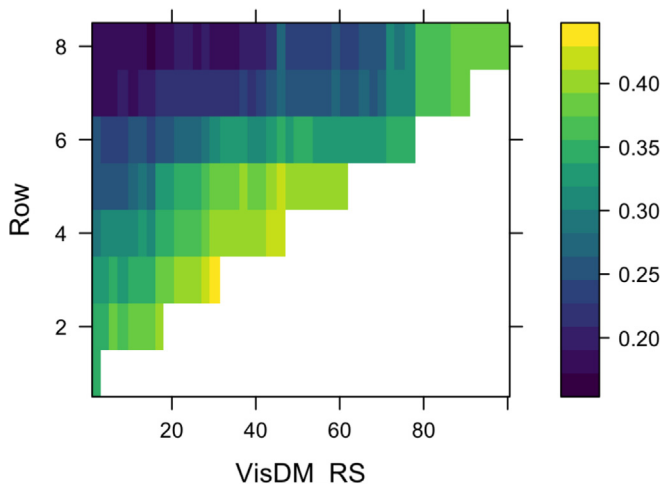


Fig. 6. Two-dimensional PDP of RS probability and the interaction of VisDM_RS and Row.

maker is affected by a combination of social factors (i.e. response of other occupants) and environmental factors (i.e. position of the decision-maker and type of alarm). As such, this analysis merges the findings provided by independent studies showing that both social influences [28] and decision-makers position [51] affect their responses.

3.4. Model comparison

In this subsection, we compare the machine-learning model (i.e., RF) with the traditional approach (i.e., logistic regression) in terms of prediction. Logistic regression is widely used to model nominal dependent variables, where the log-odds of the outcomes are modelled as a linear combination of the independent variables. Logistic regression and its extensions, such as ordered logit and mixed logit, have been applied in pre-evacuation behaviour modelling only in a couple of studies carried out by Lovreglio et al. [6, 11]. We assess the model performance with various metrics by adopting a 10-fold cross validation on the training set and conducting a separate evaluation on the testing set. The R package *stats* [53] is used fit the logistic regression model.

The model comparison results are presented in Table 4. In addition to Precision, Recall, and F1 Score, we also compute Accuracy, which is defined as the proportion of correct predictions in a validation/testing set. The 10-fold cross validation results show that RF is significantly outperforming the logistic regression model, in terms of all the four metrics. In particular, even though the mean predictive accuracy of RF (0.876) is higher than that of logistic regression model (0.830) by 5.5%, the RF's F1 Score (0.644) is 1.4 times of the logistic regression model's F1 Score (0.455). This indicates that the RF model can better deal with the class imbalance problem and thus demonstrate better predictive performance. We also evaluate the two models on the testing set that was completely excluded from the training process. The outputs are consistent with the previous findings that RF is presenting much better model performance compared to the logistic regression model.

Table 4
Model comparison results.

Metric	10-Fold cross validation on training set				Testing set	
	Logistic regression		RF		Logistic regression	RF
	Mean	SD	Mean	SD		
Accuracy	0.830	0.014	0.876	0.017	0.824	0.870
Precision	0.642	0.086	0.765	0.059	0.612	0.761
Recall	0.353	0.042	0.558	0.035	0.291	0.495
F1 Score	0.455	0.052	0.644	0.036	0.395	0.600

3.5. Model implementation

Similarly to existing logit solutions [6, 11, 54], the proposed RF model can be easily implemented in any existing agent-based models to predict the decisions to respond for each simulated evacuee regardless of the assumption used to model the movement [7]. Such an implementation can be done using two approaches: event-based or time-based. In the first case, the decision to respond is simulated for each agent for each event changing the state of the system. The second case assumes that a decision is simulated at each time-step fixed by the users. Both approaches have advantages and disadvantages in terms of the implementation simplicity, the computational cost and the impact of settings defined by the users [11, 55]. As such, future implementations of the proposed RF model need careful accounting for these aspects.

3.6. Limitations

One of the limitations of this work is related to the data used to estimate the pre-evacuation response. The data set does not provide information regarding the participants' demographics and characteristics, such as previous fire experience. As such, the proposed modelling solution does not allow the investigation of the impact of these factors on the decision-making. Future investigations will require the combination of CCTV footage with follow-up non-anonymous questionnaire as suggested in [56].

Another limitation of this work is that the proposed model does not allow an evacuee to return to NS while they are in RS (see the second assumption in Section 2.1). Such an assumption derives from the general time-line evacuation framework for which the model is proposed [1, 57]. The time-line model does not allow simulating evacuees stopping RS and returning to NS. However, such a behaviour might occur in some actual emergencies since limited ambiguous information might lead evacuees to return to their pre-alarm activities. The data used in this work does not allow investigating this specific behaviour.

4. Conclusion

Through this work, we have modelled and interpreted pre-evacuation decision-making using random forest and machine-learning interpretation tools. The results showcased that random forest can not only generate better predictive performance than logistic regression, but also provides rich behavioural interpretations by automatically capturing interactions among independent variables and nonlinearities between the independent variables and the outcome. With these interpretations, key stakeholders such as policy makers and emergency managers can extract useful insights to develop more effective evacuation plans.

This paper shows the potential of using machine-learning methods to investigate and understand complex and nonlinear behaviour in fire evacuations. We illustrate how the pre-evacuation decision-making is affected by a combination of social and physical factors: such as the behaviour of other people, seat of the decision-maker and group size. The results illustrate that most of the relationships between factors and the probability to respond is not linear. Some of these (i.e. decision-makers' seat and group size) are not monotonic, showing a U-shape trend. We also find that decision-makers seem to be more sensitive to the proportion of responding occupants than the number of them, so people in the back are showing slower response rate compared to people in the front.

In future work, we plan to use the interpretations (such as nonlinearities and interactions) gained from the machine-learning model to better specify the traditional random utility models (such as traditional logistic models and mixed logit models). Furthermore, the machine-learning-based pre-evacuation decision modelling can be integrated with the agent-based evacuation model, aiming at achieving more accurate and realistic emergency behaviour simulations.

Declaration of competing interest

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.

Acknowledgements

This research was partly supported by the Florida Sea Grant Program Development Fund and the Natural Hazards Center Quick Response Research Program. The Quick Response program is based on work supported by the National Science Foundation (Award #1635593). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or the Natural Hazards Center. The authors would like to thank Zhengze Zhou for providing helpful comments on interpretable machine learning. We also want to thank the three anonymous reviewers for their valuable suggestions.

Appendix A

The model comparison and model interpretation programs were implemented in R and open-sourced on GitHub (<https://github.com/EvacuationBehavior/Pre-Evacuation-Decision-Making>).

References

- [1] S. Gwynne, K. Boyce, Engineering data, *The Handbook of Fire Protection Engineering*, Springer, 2016, pp. 2429–2551, https://doi.org/10.1007/978-1-4939-2565-0_64.
- [2] G. Ma, Z. Wu, BIM-based building fire emergency management: combining building users' behavior decisions, *Autom. Constr.* 109 (2020) 102975, <https://doi.org/10.1016/j.autcon.2019.102975>.
- [3] R. Lovreglio, E. Kuligowski, S. Gwynne, K. Boyce, A pre-evacuation database for use in egress simulations, *Fire Saf. J.* 105 (2019) 107–128, <https://doi.org/10.1016/j.firesaf.2018.12.009>.
- [4] M. D'Orazio, S. Longhi, P. Olivetti, G. Bernardini, Design and experimental evaluation of an interactive system for pre-movement time reduction in case of fire, *Autom. Constr.* 52 (2015) 16–28, <https://doi.org/10.1016/j.autcon.2015.02.015>.
- [5] S. Gwynne, L. Hulse, M. Kinsey, Guidance for the model developer on representing human behavior in egress models, *Fire Technol.* 52 (3) (2016) 775–800, <https://doi.org/10.1007/s10694-015-0501-2>.
- [6] R. Lovreglio, E. Ronchi, D. Nilsson, A model of the decision-making process during pre-evacuation, *Fire Saf. J.* 78 (2015) 168–179, <https://doi.org/10.1016/j.firesaf.2015.07.001>.
- [7] E.D. Kuligowski, R.D. Peacock, B.L. Hoskins, *A Review of Building Evacuation Models*, US Department of Commerce, National Institute of Standards and Technology, 2005, <https://www.nist.gov/publications/review-building-evacuation-models-2nd-edition>.
- [8] G. Proulx, G. Hadjisophocleous, Occupant response model: a sub-model for the NRCC risk-cost assessment model, *Fire Saf. Sci.* 4 (1994) 841–852, <https://doi.org/10.3801/IAFSS.FSS.4-841>.
- [9] P.A. Reneke, *Evacuation decision model*, US Department of Commerce, National Institute of Standards and Technology, 2013, <https://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7914.pdf>.
- [10] E. Retana, M. Spearpoint, Assessment of the New Zealand verification method pre-travel scenarios using the Evacuation Decision Model, *6th Human Behaviour in Fire Symposium*, 2015 https://www.researchgate.net/profile/Michael_Spearpoint/publication/285579319_Assessment_of_the_New_Zealand_verification_method_pre-travel_scenarios_using_the_Evacuation_Decision_Model/links/5a016dc3aca2725286e309da/Assessment-of-the-New-Zealand-verification-method-pre-travel-scenarios-using-the-Evacuation-Decision-Model.pdf.
- [11] R. Lovreglio, E. Ronchi, D. Nilsson, An evacuation decision model based on perceived risk, social influence and behavioural uncertainty, *Simul. Model. Pract. Theory* 66 (2016) 226–242, <https://doi.org/10.1016/j.simpat.2016.03.006>.
- [12] A. Lhéritier, M. Bocamazo, T. Delahaye, R. Acuna-Agost, Airline itinerary choice modeling using machine learning, *J. Choice Model.* (2018), <https://doi.org/10.1016/j.jocm.2018.02.002>.
- [13] X. Zhao, X. Yan, A. Yu, P. Van Hentenryck, Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models, *Travel Behav. Soc.* 20 (2020) 22–35, <https://doi.org/10.1016/j.tbs.2020.02.003>.
- [14] C. Ding, X. Cao, C. Liu, How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds, *J. Transp. Geogr.* 77 (2019) 70–78, <https://doi.org/10.1016/j.jtrangeo.2019.04.011>.
- [15] S. Bandini, S. Manzoni, G. Mauri, S. Redaelli, L. Vanneschi, GP generation of pedestrian behavioral rules in an evacuation model based on SCA, *International Conference on Cellular Automata*, Springer, 2008, pp. 409–416, https://doi.org/10.1007/978-3-540-78888-8_27.

- 1007/978-3-540-79992-4_53.
- [16] S. Sharma, H. Singh, A. Prakash, Multi-agent modeling and simulation of human behavior in aircraft evacuations, *IEEE Trans. Aerosp. Electron. Syst.* 44 (4) (2008) 1477–1488, <https://doi.org/10.1109/TAES.2008.4667723>.
- [17] C. Şahin, J. Rokne, R. Alhaji, Human behavior modeling for simulating evacuation of buildings during emergencies, *Physica A* (2019) 121432, <https://doi.org/10.1016/j.physa.2019.121432>.
- [18] M.E. Yuksel, Agent-based evacuation modeling with multiple exits using NeuroEvolution of augmenting topologies, *Adv. Eng. Inform.* 35 (2018) 30–55, <https://doi.org/10.1016/j.aei.2017.11.003>.
- [19] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, R. Shibasaki, Modeling and probabilistic reasoning of population evacuation during large-scale disaster, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2013, pp. 1231–1239, <https://doi.org/10.1145/2487575.2488189>.
- [20] K. Wang, X. Shi, A.P.X. Goh, S. Qian, A machine learning based study on pedestrian movement dynamics under emergency evacuation, *Fire Saf. J.* 106 (2019) 163–176, <https://doi.org/10.1016/j.firesaf.2019.04.008>.
- [21] Y.-J. Zheng, H.-F. Ling, J.-Y. Xue, S.-Y. Chen, Population classification in fire evacuation: a multiobjective particle swarm optimization approach, *IEEE Trans. Evol. Comput.* 18 (1) (2013) 70–81, <https://doi.org/10.1109/TEVC.2013.2281396>.
- [22] S. Lo, M. Liu, P. Zhang, R.K. Yuen, An artificial neural-network based predictive model for pre-evacuation human response in domestic building fire, *Fire Technol.* 45 (4) (2009) 431–449, <https://doi.org/10.1007/s10694-008-0064-6>.
- [23] M. Liu, S.M. Lo, The quantitative investigation on people's pre-evacuation behavior under fire, *Autom. Constr.* 20 (5) (2011) 620–628, doi.org/10.1016/j.autcon.2010.12.004.
- [24] L. Breiman, Random forests, *Mach. Learn.* 45 (1) (2001) 5–32, <https://doi.org/10.1023/A:1010933404324>.
- [25] A. Liaw, M. Wiener, et al., *Classification and regression by RandomForest*, *R News* 2 (3) (2002) 18–22.
- [26] J. Hagenauer, M. Helbich, A comparative study of machine learning classifiers for modeling travel mode choice, *Expert Syst. Appl.* 78 (2017) 273–282, <https://doi.org/10.1016/j.eswa.2017.01.057>.
- [27] K. Bayer, K. Rejnö, *Utrymningslarm-Optimering genom fullskaleförsök*, Lund University - Report N: 5053, 1999, <https://lup.lub.lu.se/student-papers/search/publication/1767237>.
- [28] D. Nilsson, A. Johansson, Social influence during the initial phase of a fire evacuation-analysis of evacuation experiments in a cinema theatre, *Fire Saf. J.* 44 (1) (2009) 71–79, <https://doi.org/10.1016/j.firesaf.2008.03.008>.
- [29] R.F. Fahy, G. Proulx, L. Aiman, Panic or not in fire: clarifying the misconception, *Fire Mater.* 36 (5-6) (2012) 328–338, <https://doi.org/10.1002/fam.1083>.
- [30] E. Kuligowski, Predicting human behavior during fires, *Fire Technol.* 49 (1) (2013) 101–120, <https://doi.org/10.1007/s10694-011-0245-6>.
- [31] C. Xie, J. Lu, E. Parkany, Work travel mode choice modeling with data mining: decision trees and neural networks, *Transp. Res. Rec. J. Transp. Res. Board* (1854) (2003) 50–61, <https://doi.org/10.3141/1854-06>.
- [32] Y. Liu, H.H. Zhang, Y. Wu, Hard or soft classification? Large-margin unified machines, *J. Am. Stat. Assoc.* 106 (493) (2011) 166–177, <https://doi.org/10.1198/jasa.2011.tm10319>.
- [33] G. Biau, Analysis of a random forests model, *J. Mach. Learn. Res.* 13 (Apr) (2012) 1063–1095 <http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf>.
- [34] R. Caruana, N. Karampatziakis, A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, *Proceedings of the 25th International Conference on Machine Learning*, ACM, 2008, pp. 96–103, <https://doi.org/10.1145/1390156.1390169>.
- [35] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Application of machine learning to construction injury prediction, *Autom. Constr.* 69 (2016) 102–114, <https://doi.org/10.1016/j.autcon.2016.05.016>.
- [36] A. Lebedev, E. Westman, G. Van Westen, M. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, A. Simmons, Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, *NeuroImage Clin.* 6 (2014) 115–125, <https://doi.org/10.1016/j.nicl.2014.08.023>.
- [37] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study, *Decis. Support. Syst.* 50 (3) (2011) 602–613, <https://doi.org/10.1016/j.dss.2010.08.008>.
- [38] S. Mangalathu, H. Sun, C.C. Nweke, Z. Yi, H.V. Burton, Classifying earthquake damage to buildings using machine learning, *Earthq. Spectra* (2019) 0000-0000, <https://doi.org/10.1177/8755293019878137>.
- [39] L. Breiman, *Classification and Regression Trees*, Routledge, 2017 ISBN 9780412048418.
- [40] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, (2018) ISBN 9780244768522.
- [41] L. Breiman, Bagging predictors, *Mach. Learn.* 24 (2) (1996) 123–140, <https://doi.org/10.1007/BF00058655>.
- [42] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer series in statistics New York, NY, USA: 1 (2001) ISBN 978-0387848570.
- [43] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016 ISBN 978-0-12-804291-5.
- [44] A. Liaw, M. Wiener, Classification and Regression by RandomForest, *R News* 2 (3) (2002) 18–22 <http://CRAN.R-project.org/doc/Rnews/>.
- [45] X. Zhao, X. Yan, P. Van Hentenryck, Modeling Heterogeneity in Mode-Switching Behavior Under a Mobility-on-Demand Transit System: An Interpretable Machine Learning Approach, arXiv preprint, 2019. <http://arXiv:1902.02904>.
- [46] J.H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Stat.* (2001) 1189–1232 <https://projecteuclid.org/download/pdf1/euclid.aos/1013203451>.
- [47] A. Fisher, C. Rudin, F. Dominici, Model class reliance: variable importance measures for any machine learning model class, from the “Rashomon” perspective, arXiv preprint, 2018. <http://arXiv:1801.01489>.
- [48] Q. Zhao, T. Hastie, Causal interpretations of black-box models, *J. Bus. Econ. Stat.* (just-accepted) (2019) 1–19, <https://doi.org/10.1080/07350015.2019.1624293>.
- [49] D. Muchlinski, D. Siroky, J. He, M. Kocher, Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data, *Polit. Anal.* 24 (1) (2016) 87–103, <https://doi.org/10.1093/pan/mpv024>.
- [50] D.B. McRoberts, S.M. Quiring, S.D. Guikema, Improving hurricane power outage prediction models through the inclusion of local environmental factors, *Risk Anal.* 38 (12) (2018) 2722–2737, <https://doi.org/10.1111/risa.12728>.
- [51] E. Galea, S. Deere, C. Hopkin, H. Xie, Evacuation response behaviour of occupants in a large theatre during a live performance, *Fire Mater.* 41 (5) (2017) 467–492, <https://doi.org/10.1002/fam.2424>.
- [52] S. Sheather, *A Modern Approach to Regression with R*, Springer Science & Business Media, 2009 ISBN 978-0-387-09608-7.
- [53] W.N. Venables, B.D. Ripley, *Modern Applied Statistics with S*, fourth, Springer, New York, 2002 <http://www.stats.ox.ac.uk/pub/MASS4> ISBN 978-0-387-21706-2, ISBN 0-387-95457-0.
- [54] R. Lovreglio, A. Fonzone, L. dell’Olio, A mixed logit model for predicting exit choice during building evacuations, *Transp. Res. A Policy Pract.* 92 (2016) 59–75, <https://doi.org/10.1016/j.tra.2016.06.018>.
- [55] B. Zhang, W.K. Chan, S.V. Ukkusuri, Agent-based discrete-event hybrid space modeling approach for transportation evacuation simulation, *Proceedings of the 2011 Winter Simulation Conference*, IEEE, 2011, pp. 199–209, <https://doi.org/10.1109/WSC.2011.6147751>.
- [56] R. Lovreglio, D. Borri, E. Ronchi, A. Fonzone, L. Dell’Olio, The need of latent variables for modelling decision-making in evacuation simulations, *IX International Workshop on Planning and Evaluation, Bari* (2015), <https://doi.org/10.13140/RG.2.1.4521.7684>.
- [57] E.D. Kuligowski, Human behavior in fire, *The Handbook of Fire Protection Engineering*, Springer, 2016, pp. 2070–2114, https://doi.org/10.1007/978-1-4939-2565-0_58.