Insurance: Mathematics and Economics 101 (2021) 140–162

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

www.elsevier.com/locate/ime

A random forest based approach for predicting spreads in the primary catastrophe bond market

Despoina Makariou *, Pauline Barrieu, Yining Chen

Department of Statistics, London School of Economics and Political Science, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received February 2020
Received in revised form June 2021
Accepted 17 July 2021
Available online 30 July 2021

JEL classification:
G1
G220

Keywords:
Catastrophe bond pricing
Interactions
Machine learning in insurance
Minimal depth importance
Permutation importance
Primary market spread prediction
Random forest
Stability

ABSTRACT

We introduce a random forest approach to enable spreads' prediction in the primary catastrophe bond market. In a purely predictive framework, we assess the importance of catastrophe spread predictors using permutation and minimal depth methods. The whole population of non-life catastrophe bonds issued from December 2009 to May 2018 is used. We find that random forest has at least as good prediction performance as our benchmark-linear regression in the temporal context, and better prediction performance in the non-temporal one. Random forest also performs better than the benchmark when multiple predictors are excluded in accordance with the importance rankings or at random, which indicates that random forest extracts information from existing predictors more effectively and captures interactions better without the need to specify them. The results of random forest, in terms of prediction accuracy and the minimal depth importance are stable. There is only a small divergence between the drivers of catastrophe bond spread in the predictive versus explanatory framework. We believe that the usage of random forest can speed up investment decisions in the catastrophe bond industry both for would-be issuers and investors.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Catastrophe bonds are **I**nsurance-**L**inked **S**ecurities (ILS), first developed in 1990s, in an effort to provide additional capacity to the reinsurance industry post mega-disasters. The pricing of these instruments is particularly challenging as most of these securities are traded over the counter. Over the last years, there have been several empirical papers trying to address this difficulty by studying the price of catastrophe bonds using real-market data, mainly in the explanatory framework, see Lane (2000), Lane and Mahul (2008), Lei et al. (2008), Bodoff and Gan (2009), Gatumel and Guegan (2008), Dieckmann (2010), Jaeger et al. (2010), Papachristou (2011), Galeotti et al. (2013), Braun (2016), Gürtler et al. (2016), Götze and Gürtler (2018), Trottier et al. (2018), and only very recently in the context of comparative studies for machine learning algorithms, Götze et al. (2020).

The main orientation of the explanatory based approach was to explain catastrophe bond price via means of identification of variables having a theoretically material and statistically significant link to it. This was mostly achieved through the use of explanatory statistical models. Certainly, the aforementioned works have shed light on the drivers of catastrophe bond prices in the presence of causal theory. However, there are certain limitations, namely, selection bias, predictor interactions, non-linearities, and a non-purely-predictive study goal. Starting from selection bias, the data samples used previously often excluded bonds of certain characteristics, unusual issuances were eliminated as outliers, and observations with missing entries were excluded from data sets, leading to a potential significant loss of information. See Bodoff and Gan (2009), Götze and Gürtler (2018), Galeotti et al. (2013), Braun (2016) and Lane and Mahul (2008). Meanwhile, in Papachristou (2011), concerns about interactions between independent variables were expressed but not investigated. Another limitation is the extensive use of linear regression without justification of its suitability in a catastrophe bond market setting. This was recognised in some cases, see Lane and Mahul (2008) and Papachristou (2011). Finally, as Major (2019) mentioned, in terms of study goal, past works did not aim directly at spread prediction, although there is a business need for it.

In this manuscript, we suggest a supervised machine learning method called random forest (Breiman 2001) to predict spreads in the primary catastrophe bond market. Some reasons about the

* Corresponding author.
E-mail addresses: D.Makariou@lse.ac.uk (D. Makariou), P.M.Barrieu@lse.ac.uk (P. Barrieu), Y.Chen101@lse.ac.uk (Y. Chen).

https://doi.org/10.1016/j.insmatheco.2021.07.003
0167-6687/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

model specification are discussed below. The model choice is partially based on the fact that random forest is widely considered as one of the most successful machine learning methods to date, see Berk (2008), and Biau and Scornet (2016) among others. It should be noted that random forest success in providing highly accurate predictions is mainly achieved by resolving the trade-off between over-fitting and prediction accuracy, as discussed in various works such as these of Breiman (2001), Díaz-Uriarte and De Andres (2006), Oh et al. (2003). Moreover, the recent novel research of Götze et al. (2020) compared different machine learning methods in a catastrophe bond market setting, which provides evidence that random forest outperforms neural networks, and linear regression which is combined with variable selection via Lasso and Ridge penalisations. In addition, we believe that the random forest method has a number of particular aspects which could help overcome some of the limitations presented previously in the explanatory framework in the literature. Firstly, random forest is a flexible method in a sense that makes no assumptions about the underlying data generative process. This is an important advantage that could help us effectively tackle the issue of non-linearities in the catastrophe bond market. Secondly, because the building blocks of the method are regression trees, random forest is reasonably robust to outliers. This is very useful given that catastrophe bonds can be extremely heterogeneous and losing information is particularly "costly" in this opaque market segment. Thirdly, once again, due to the tree structure of the method, variables are considered in such a way that allows to capture interactions without the need to specify them (Breiman et al. 1984). Fourthly, internal measures of variables' importance can be derived solely in a prediction context, and selection of the most important variables is feasible. Finally, there is only a small number of hyperparameters to tune, and the need for data pre-processing is minimal because many steps are integrated in the method itself, ensuring time efficiency from a business perspective.

Here, we apply the random forest method to predict spreads in the full spectrum of primary non-life catastrophe bond market. We aim to generate accurate spreads' predictions of new catastrophe bond observations on both temporal and non-temporal bases. Comparisons are made with highly competitive benchmark models. In absence of causal theory, we assess how spread predictors rank in terms of importance using two different methods, namely, permutation importance and minimal depth, where the latter is random forest specific. To our best knowledge, this work is among the first to apply the minimal depth method as described in Ishwaran et al. (2010) in a financial application. In addition, we explore whether the variables found by now to be good at explaining catastrophe bond spreads in the explanatory framework are similar to those good at prediction in absence of causal theory. As mentioned in Shmueli (2010), one should not expect these two to be exactly the same and indeed we find some small level of divergence. From an empirical perspective, we aim at random forest prediction accuracy and variables' importance results to be stable thus this aspect is also evaluated subject to multiple iterations of random subsampling. Besides, we assess the degree at which the prediction accuracy of random forest versus benchmark model is sensitive to simultaneous missingness of more than one predictor. By doing so we also check the degree to which the random forest captures predictors' interactions without specifying them, as well as its ability to extract information from existing variables to recover the loss of predictive power in the absence of other important predictors.

With regards to the benchmark model, first we reproduce and then improve the model of Braun (2016) to account for non-rated catastrophe bond issuances. Braun (2016) is chosen as it indicates the best out of sample performance to date in the relevant explanatory literature. Next, we build a new simple linear regression model based on the same set of variables we use for the random forest generation. For the first time, we include the risk modelling company and coverage type in the analysis as potential catastrophe bond spread drivers making a contribution in the explanatory framework. A potential reason for lack of prior works taking into account these variables is most probably due to the difficulty of finding information about them as they pin-point to very detailed aspects of a transaction - a view that Braun (2012) already expressed regarding the risk modelling company. The newly built regression model outperforms the improved version of Braun (2016) and thus is used as our benchmark in this manuscript.

The rest of the paper is organised as follows. In Section 2, we briefly introduce machine learning concepts. We explain our research methodology in Section 3 and present our catastrophe bond data set details in Section 4. Benchmark models are discussed in Section 5. We then demonstrate the random forest generation based on our catastrophe bond data in Section 6, followed by the evaluation of the random forest's performance in Section 7, and the importance analysis of catastrophe bond spread predictors in Section 8. Furthermore, in Section 9, we provide an example of how the random forest could be used in practice to assist issuers' and investors' decision making when they examine a new catastrophe bond issuance. Finally, concluding remarks follow in Section 10.

2. Machine learning preliminaries

In this section, we introduce some machine learning concepts that will be useful for the comprehension of methods used later on in our study. The explanations to be given are limited to a regression problem because catastrophe bond spread is a quantitative response variable.[1]

2.1. Supervised learning

Machine learning includes a set of approaches dealing with the problem of finding or otherwise learning a function from data (James et al. 2013). Supervised learning is a machine learning task where a function, otherwise called a hypothesis, is learned from a data set - often referred to as training set. The latter consists of a number of input-output pairs where for every single input in the training set the correct output is known. An algorithm is going through all data points in the training set identifying patterns and finding how to map an input to an output. Because the desired answer for the output is known, the algorithm modifies this mapping based on how different algorithm generated outputs are compared to the original ones in the training set (Friedman et al. 2001). Ultimately, the aim is that by the time the learning process finishes, this difference will be small enough for the algorithm to be able to map any set of new inputs the algorithm will come across in the future in a reasonable manner.

2.2. Ensemble learning

Sometimes instead of learning one mapping, it is useful to have a collection of mappings which merge their predictions to create an ensemble (Russell and Norvig 2016). Individual approximation functions in the ensemble are usually called base learners and predictions combination can happen in various ways with most usual ones being voting or averaging. Such techniques have

[1] We clarify that in machine learning literature, the term "regression problem" often refers to prediction using a continuous response variable, see James et al. (2013). We distinguish this from the term linear regression that is used throughout this work to describe either the linear regression models in the literature or our benchmark model.

been investigated quite early on, see for example Breiman (1996c), Clemen (1989), Perrone (1993) and Wolpert (1992). The main benefit of ensembles is that if each single hypothesis is characterised by high degree of accuracy and diversity then the ensemble is going to produce more accurate predictions than any of the individual hypotheses on its own, see Zhou (2012). Here, accuracy means that a hypothesis results in a lower error rate as opposed to one that would be derived from random guessing on new input values, while diversity means that each hypothesis in the ensemble makes different errors on new data points (Dietterich 2000a). Ensembles are usually built by utilising methods to derive various data sets out of the original data set for each base learner. One of the most famous methods to construct an ensemble is briefly discussed below.

2.3. Bagging

Bagging, an acronym for **b**ootstrap **agg**regat**ing** presented by Breiman (1996a), is a powerful ensemble learning method. As the name indicates, the ensemble uses the bootstrap, see Efron (1992), as resampling technique to take multiple data samples from which multiple base learners will be then generated. At the same time, aggregation, which is simple averaging for regression, is the way to combine the predictions of these individual base learners. There are various merits in using bagging for building ensembles. First, using a bootstrap sample to build each base learner means that a part of the original data (normally two third by default) are not used in its construction. Then, these unseen data points can constitute an unbiased test data to quantify how well each base learner generalises (Breiman 2001). Secondly, the method is useful when data is noisy (Opitz and Maclin 1999). Thirdly, and probably most importantly, by aggregating base learners which individually suffer from high variance, e.g. decision trees (Breiman et al. 1984), the ensemble as a whole achieves a variance reduction; see Breiman (1996a), Bauer and Kohavi (1999), Breiman (1996c), Breiman (1996b) and Dietterich (2000b). A pitfall of the method though is that whilst bagging reduces the ensemble variance, there are diminishing returns in variance reductions as the computational cost increases. This is because all bootstrap samples are drawn from the same original data set, meaning that base learners will inevitably be correlated. This latter point is where the idea of random forest is based on and it will be further discussed in Section 3.

3. Research methodology

Having provided necessary background information about certain machine learning concepts, the purpose of this section is twofold. We start by stating our catastrophe bond spread prediction problem introducing notations that will be used later in our study. We then continue by presenting our research methodology.

3.1. Problem statement with notations

Broadly, we use an ensemble algorithmic method to perform a supervised learning task for the primary catastrophe bond market. For now, let $\mathbf{x}$ generally denote[2] the input which reflects characteristics of catastrophe bonds available in the offering circular at the time of issuance and ILS market conditions. At the same time, let symbol y denote catastrophe bond spreads at the time of issuance. A function f of the form $y = f(\mathbf{x})$ relates catastrophe bond characteristics, conditions in the economic environment and possibly random effects to their spreads, however f is unknown. Based on past primary catastrophe bond data including information both for $\mathbf{x} = (x_1, x_2, \ldots, x_P)$ where $p = 1, 2, \ldots, P$ and y, we first want to find a function that approximates f so that we can predict spreads given new catastrophe bond input.

In particular, experience about past catastrophe bond issuances is captured by collecting $n = 1, \ldots, N$ distinct input-output pairs. The input is a vector of predictors, also called features, covariates or independent variables, $\mathbf{x}_n = (x_{1n}, x_{2n}, \ldots, x_{Pn})$ indexed by dimension $p = 1, 2, \ldots, P$ and it is a element of $\mathbb{R}^p$. The output, also called response or dependent variable, is a real-valued scalar denoted by y_n indexed by example number $n = 1, \ldots, N$. By assembling these N pairs, we collectively form a catastrophe bond data set $D = \{(\mathbf{x}_n, y_n), n = 1, 2, \ldots, N\}$ based on which the ensemble algorithmic method will search the space H of all feasible functions, in a process called learning, and find a function, denoted by h_{en}, that is able to predict the response y' given a new input $\mathbf{x}'$ as accurately as possible. Because, we use an ensemble method, h_{en} is in reality a collection of functions approximating f. We are also interested in assessing the importance of each input of $\mathbf{x}$ in predicting the spread. Finally, all results will be evaluated on the grounds of them being stable subject to random subsampling of the whole data set.

3.2. Random forest

The ensemble method that we use is called random forest. It is developed by Breiman (2001) and is used to solve prediction problems. Below we present the rationale behind the method, random forest construction process, main hyperparameters, and lastly how random forest is used to make predictions.

3.2.1. Underlying logic

As James et al. (2013) mentioned, the underlying logic of random forest is to "divide and conquer": split the predictor space into multiple samples, then construct a randomised tree hypothesis on each subspace and end with averaging these hypotheses together. Generally, random forest can be seen as a successor of bagging when the base learners are decision trees. This is because random forest addresses the main pitfall of bagging; the issue of diminishing variance reductions discussed earlier in Section 2.3. This is achieved by injecting an additional element of randomness during decision trees construction for them to be less correlated to one another. At the same time, since the base learners are decision trees there are not many assumptions about the form of the target function resulting in low bias.

3.2.2. Random forest construction process

The process of constructing a random forest involves various steps which are summarised in Fig. 1 and discussed straight after.

The first step in the random forest generation process is bootstrap sampling. In particular, from a data set, like D, we take $1, \ldots, K$ samples with replacement each of them having the same size as the original data set. The second stage is regression trees development. From each bootstrap sample, K regression trees are grown using recursive partitioning as done in **C**lassification **A**nd **R**egression **T**rees (CART) (Breiman et al. 1984) but with a smart twist which further randomises the procedure. At each level of the recursive partitioning process, the best predictor to conduct the splitting is considered based on a fresh, each time, random subsample of the full set of predictors denoted as $\mathrm{m_{try}}$. The best split is chosen by examining all possible predictors in this sub-sample and all possible cut-points as of their ability to minimise the residual sum of squares for the resulting tree. A tree stops growing when a minimum number of observations in a given node is reached but generally speaking trees comprising the random forest are fully grown and not pruned. By constructing these K trees we

[2] Our convention is that bold lowercase letters reflect random vectors.

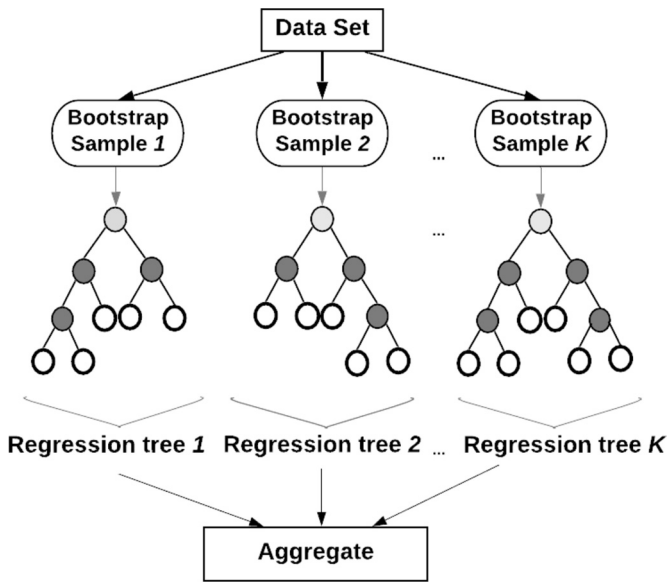


Fig. 1. Random forest construction scheme. For each regression tree, light grey circles indicate the root node, dark grey circles intermediate nodes and white colour circles terminal nodes.

effectively get K estimators of function f namely $h_1, h_2, \ldots, h_K$. The average of these individual estimators $h_{en} = \frac{1}{K}\sum_{k=1}^{K} h_k(\mathbf{x}_n)$ is the random forest.

3.2.3. Hyperparameters

From the above process description, it is evident that there are three parameters whose value needs to be fixed prior to random forest development; namely the number of trees grown, node size, and number of variables randomly selected at each split. Each of them respectively control the size of the forest, the individual tree size and an aspect of the within tree randomness. There are certain default values that have been suggested following empirical experiments on various data sets but one can use an optimising tuning strategy with respect to prediction performance to select the most suitable values specifically for the data set under study (Probst et al. 2018).

3.2.4. Making predictions

After the random forest is built, it can be used to provide predictions of the response variable. To make predictions though, it is necessary to feed the method inputs that have never been seen before during the construction process. As we have briefly mentioned in Section 2.3, due to bootstrap sampling, we can refrain from keeping aside in advance a portion of the original data set for testing purposes. The reason for this is that each tree uses more or less two thirds of the observations, from now on called in-bag observations, whilst the remaining one third of the observations are never used to build a specific tree, from now on called out of bag (OOB) observations. For each tree, the out of bag observations act as a separate test set. To predict the response variable value for the nth observation, one should drop its corresponding input down every single tree in which this observation was out of bag. This means that by doing so one will end up having in hand on average $K/3$ predictions for any $n = 1, \ldots, N$ observation. Then, in order to derive a single response prediction for the nth observation, the average of these predictions is taken. The same procedure is repeated for all other observations. Whether these predictions are good enough or not needs to be evaluated based on certain metrics as shown next.

3.3. Performance evaluation criteria for random forest

To assess the performance of any machine learning algorithm, one needs to set in advance the criterion upon which judgement will be made. In this paper, we employ two criteria for the performance evaluation of our random forest; prediction accuracy and stability. They are discussed below.

3.3.1. Prediction accuracy

Prediction accuracy is one of the most used performance indicators in machine learning algorithms aiming at prediction. This is no different for random forest algorithm as originally presented in Breiman (2001). In the current study, prediction accuracy is assessed based on two different perspectives: a temporal and a non-temporal one. We believe that such a distinction highlights different prediction needs and could add value in a practical context.

By employing a non-temporal approach, one can assess random forest predictions robustness when at the time of the spread prediction, the general catastrophe bond market conditions have been relatively stable over a time period prior the prediction, thus the time element could potentially be ignored. A non-temporal perspective would also be meaningful when simply the character of the prediction is not time relevant. With regards to the latter, an instance would be when there is ambiguity around the accuracy of spread information a company holds for a transaction or in cases that the spread information for a given transaction is unavailable resulting in a company having to face the issue of an incomplete data base. In both cases, the element of time may appear to be less important compared to the need of having a bigger and more diverse training set in the analysis. On the other hand, a temporal approach considers the robustness of the random forest in accurately predicting spreads over time. Effectively, such a point of view allows us to account for regime shifts and examine the degree at which an industry participant would be able to predict a new catastrophe bond spread no matter its features and risk profile. For this purpose, we need to split the data into separate train and test sets for various time periods and then assess the random forest and benchmark model prediction accuracy performance.

In the non-temporal context, prediction accuracy is primarily measured by means of the proportion of the total variability explained by the random forest, here denoted as R^2_{OOB}. Following Grömping (2009), the latter metric is defined as $R^2_{OOB} = 1 - \frac{SE_{OOB}}{TSS}$ where SE_{OOB} stands for the total out of bag squared errors and TSS for the total sum of squares. In addition, we denote the out of bag mean squared error as $MSE_{OOB} = SE_{OOB}/N$. With respect to MSE_{OOB}, it shows the variability in the response variable that is not forecasted by the random forest. It is calculated as $MSE_{OOB} = \{\sum_{n=1}^{N}(y_n - \overline{\hat{y}}_{n_{OOB}})^2\}/N$ where $\overline{\hat{y}}_{n_{OOB}}$ is the mean prediction for the nth observation where $n = 1, \ldots, N$ for all trees for which the nth data point was out of bag. In effect, MSE_{OOB} is a sound approximation of the test error for the random forest because every single data point is predicted based solely on the trees that were not constructed using this observation. Actually, when the number of trees K is very large then the MSE_{OOB} is roughly equivalent to leave one out cross validation James et al. (2013). With regards to TSS, as in linear regression, it reflects the degree at which the response variable, here the catastrophe bond spread, deviates from its mean value. It is defined as $TSS = \sum_{n=1}^{N}(y_n - \overline{y})^2$ where y_n is the response variable value for the nth observation where $n = 1, \ldots, N$ and $\overline{y}$ the mean value of the response variable. In this study, R^2_{OOB} is going to be expressed in percentage terms. The higher the R^2_{OOB}, the better the prediction accuracy of the random forest is. Whilst for random forest, it is somehow natural to use the R^2_{OOB}, see Breiman (2001), we deem useful to also

present the prediction accuracy results derived by two "more standard" statistical approaches, i.e. 10 fold cross validation and leave one out cross validation even though we expect that results may be fairly similar. In the temporal context, prediction performance will be assessed on the basis of out of sample R^2 denoted as R^2_{OOS}. The metrics presented here both in the non-temporal, and temporal context will be also used for the benchmark model to allow for a fair comparison.

3.3.2. Stability

The term stability here refers to how repeatable random forest results are when different samples taken from the same data generative process are used for its construction, see Turney (1995) and Philipp et al. (2018) for the rationale behind this approach. The rationale for investigating stability is rooted from the fact that consistent results are deemed more reliable, see Stodden (2015), Turney (1995), Yu (2013) and Philipp et al. (2018) for a discussion.

Various ways of measuring the stability of algorithmic results have been presented in Turney (1995), Lange et al. (2004), Ntoutsi et al. (2008), Lim and Yu (2016) and Philipp et al. (2018). In this study, we are inspired by the works of Turney (1995) and Philipp et al. (2018) with regards to stability and its empirical measurement. In particular, the idea is that by obtaining two sets of data from the same phenomenon sampled from the same underlying distribution the algorithm needs to produce fairly similar results from both data sets for it to be considered stable. One way to achieve this is to randomly partition the whole data set into two separate data sets multiple times. An important decision though is how to take the samples. Here, we propose taking the samples using the split-half technique as described in Philipp et al. (2018) meaning that the whole catastrophe data set will be split into two disjoint data sets of roughly equal size. This sampling method ensures that a similarity between the results is not attributed to the same observations being in both samples as this could result in similar results without meaning that the algorithm is actually stable. By choosing a small learning overlap it is possible to examine the degree of a result generalisation for independent draws from the catastrophe bond data generative process.

In particular, following Turney (1995) and Philipp et al. (2018), we obtain two sets of data from the same phenomenon and same underlying distribution with as little learning overlap as possible, then construct two random forests from each one and check whether prediction accuracy is fairly similar. To be more specific, we take a random 50% of the observations without replacement from the initial catastrophe bond data set, namely Sample A. The rest of the original data set observations, not included in Sample A, forms Sample B. Then, two separate random forests are grown out of Sample A and Sample B to assess the stability of random forest prediction accuracy to changes in the initial data set. We repeat this process 100 times. Optimal values for the number of variables randomly selected to be considered at each split are sought in both cases.

3.4. Evaluation of predictors' importance

The random forest algorithm allows for assessing how important each predictor is with respect to its ability to predict the response, a concept that is briefly called as variables importance. Its assessment is executed empirically (Grömping 2009) and see Chen and Ishwaran (2012) for a comprehensive review of various methods that can be used to achieve this. Here, the focus lies on two widely used approaches namely permutation importance, and minimal depth importance.

3.4.1. Permutation importance

The central idea of permutation importance, also known as "Breiman-Cutler importance" (Breiman 2001), is to measure the decrease in the prediction accuracy of the random forest resulting from randomly permuting the values of a predictor. The method provides a ranking for predictors' importance as end result and it is tied to a prediction performance measure. In particular, the permutation importance for x_p predictor is derived as follows. For each of the K trees: firstly, record the prediction error $\mathrm{MSE}_{\mathrm{OOB_k}}$; secondly, noise up, i.e. permute, the predictor x_p in the out of bag sample for the kth tree; thirdly, drop this permuted out of bag sample down the kth tree to get a new $\mathrm{MSE}^{x_{p\,perm}}_{\mathrm{OOB_k}}$ after the permutation and calculate the difference between these two prediction errors (before and after the permutation). In the end, average these differences over all trees. The mathematical expression of the above description is $I_{x_p} = \sum_{k=1}^{K}[\frac{1}{K}(\mathrm{MSE}^{x_{p\,perm}}_{\mathrm{OOB_k}} - \mathrm{MSE}_{\mathrm{OOB_k}})]$ where I_{x_p} is the importance of variable x_p, K the number of trees in the forest, $\mathrm{MSE}^{x_{p\,perm}}_{\mathrm{OOB_k}}$ the estimation error with predictor x_p being permuted for the kth tree, and $\mathrm{MSE}_{\mathrm{OOB_k}}$ the forecasting error with none of the predictors being permuted for the kth tree. The larger the I_{x_p} the stronger the ability of x_p to predict the response. Generally speaking a positive permutation importance is associated with decrease in prediction accuracy after permutation whilst negative permutation importance is interpreted as no decline in accuracy.

3.4.2. Importance based on minimal depth

The other approach for measuring predictors importance is based on measure named minimal depth, presented in Ishwaran et al. (2010) with the latter being motivated by earlier works of Strobl et al. (2007) and Ishwaran (2007). The minimal depth shows how remote a node split with a specific predictor is with respect to the root node of a tree. Thus, here the position of a predictor in the kth tree determines its importance for this tree. The latter means that unlike permutation importance, the importance of each predictor is not tied on a prediction performance measure. Also, in addition to ranking variables, the method also performs variable selection - a very useful feature for elimination of less important predictors.

Specifically, Ishwaran et al. (2010) have formulated the concept of minimal depth based on the notion of maximal sub-tree for feature x_p. The latter is defined as the largest sub-tree whose root node is split using x_p. In particular, the minimal depth of a predictor x_p, a non-negative random variable, is the distance between the kth tree root node and the most proximate maximal sub-tree for x_p, i.e. the first order statistic of the maximal subtree. It takes on values $\{0, \ldots, Q(k)\}$ where $Q(k)$ the depth of the kth tree reflects how distant is the root from the furthermost leaf node, i.e. the maximal depth (Ishwaran et al. 2011). A small minimal depth value for predictor x_p means that x_p has high predictive power whilst a large minimal depth value the opposite. The root node is assigned with minimal depth 0 and the successive nodes are sequenced based on how close they are to the root. The minimal depth for each predictor is averaged over all trees in the forest. Ishwaran et al. (2010) showed that the distribution of the minimal depth can be derived in a closed form and a threshold for picking meaningful variables can be computed, i.e. the mean of the minimal depth distribution. In particular, variables whose forest aggregated minimal depth surpasses the mean minimal depth ceiling are considered irrelevant and thus could be excluded from the model. However, since Ishwaran et al. (2010) suggests that variable selection using the minimal depth threshold is more meaningful for problems with high dimensionality, this aspect is not considered relevant in the current study.

3.4.3. Other evaluation factors

After calculating the importance of predictors using the methods described above, we consider useful to examine the results based on two additional criteria. Firstly, we want to ensure that primarily the importance rankings and secondarily the selected variables are repeatable. Because both permutation and minimal depth importance are linked to the random forest constructed, the stability of predictors' importance results is evaluated in line with the random forest stability evaluation for the catastrophe bond data set, as mentioned in Section 3.3.2. Secondly, we check whether the predictors' importance results reflect investors' knowledge from an empirical perspective. In a business context, it would be uncomfortable for an investor to see good catastrophe bond predictions but with importance rankings of the predictors outside their empirical knowledge, even though this type of agreement is not necessary from a statistical viewpoint.

4. Catastrophe bond data

In this section, we present how the catastrophe bond data used in this study have been collected and processed whilst details are given with respect to the choice of variables and their role in our study.

4.1. Collection

The core of catastrophe bond pricing cross sectional data has been collected from a leading market participant enabling us to work with a data set that is substantially larger than those used in the literature. The websites of ARTEMIS, Lane Financial LLC and Swiss Re Sigma Research have been also extensively used to cross validate data entries that were unclear or non-available in the main data body. Historical values of the Synthetic Rate on Line index have been given by Lane Financial LLC. To the best of our knowledge, our data set refers to all non-life catastrophe bonds issued in the primary market from December 2009 to May 2018, a total of 934 transactions. This time period is particularly interesting since it coincides with the restart of the catastrophe bond sector after almost two years of low activity following the collapse of Lehman Brothers, which played a counterparty role in several bonds and therefore ignited concerns and reflection around the structuring of transaction as to ensure security of collateral, see Hills (2009). The information gathered was related to investors' return, loss potential of the securitized risk, i.e. expected loss and attachment probability, various design characteristics of the risk transfer, i.e. issuance size, coverage period, coverage type, trigger, region, peril, credit score, risk modelling company, price cyclicality in ILS market, and BB corporate bond spreads level.

4.2. Preparation

Since we consolidated data from various sources there were pieces of information referring to the same concept but measured in different units across different data providers. Such scaling issues have been appropriately addressed to maintain consistency. With regards to the spread at issuance, it was derived from the coupon by subtracting the element of the money market rate. In the case of zero coupon catastrophe bonds the spread was derived from the implied coupon by subtracting the element of the money market rate.

Through validating the data across various sources, we ensured that there are no missing values in the study, a pitfall in many previous works. On this note, it needs to be acknowledged that an exception in the above non-missing values claim is very few catastrophe bonds for which there was no information regarding the risk modelling firm because these transactions were privately placed even though our data set contains other private placement deals for which we did not have missing values. For these few deals for which vendor information was missing, we created a separate category level to capture this specific reason for missingness, i.e. private placement. Including this level is considered important via means that the developed algorithmic method will be able to predict spreads for these circumstances also. Further information on this category level can be found in Appendix A.

4.3. Discussion about the choice of variables

The variables included in the data set can be seen in Table 1, presented along with the definition, type, and their role in this study. In Appendix A, one can find basic statistical information and histograms for all variables along with a discussion to enhance the understanding of catastrophe bond data intricacies. With regards to the role of each variable in our research, the spread was chosen as dependent variable as it is an industry wide accepted lens through which one can see catastrophe bond pricing. The spread is of utmost interest to the investors as it indicates how much they could earn on the top of the risk free rate if they decided to employ their capital in this alternative risk transfer segment.

Since the goal of this study lies on the prediction of spread, a major consideration is that the independent variables need to be available at the time of the prediction. This is indeed the case here, as the predictors constitute information included in the placement material offered to investors prior to a new catastrophe bond issuance. Also, in the case of predictor RoL, investors are also aware of the general ILS market conditions and possibly we could assume that the Financial Lane LLC Synthetic Rate on Line index values are readily available at an investment company level. Similar rationale applies for the BB spread regarding its availability at the point of the prediction. The reason why we have incorporated RoL and BB spread in the study is because the prior literature shows that such macroeconomic variables have a relevant influence on catastrophe bond spreads, see Braun (2016) and Gürtler et al. (2016) for example.

We note that there are previous works (see Galeotti et al. (2013), Braun (2016), Gürtler et al. (2016), and Trottier et al. 2018, among others) refraining from using the attachment probability (AP) as a predictor for the spread forecast, even though the reason was not mentioned explicitly. A potential explanation for why the EL was preferred over the AP in these works is because the EL is a coherent risk measure, meaning that if we were to examine catastrophe bonds in a portfolio context, then the EL contributes proportionately to the portfolio EL. This is not the case when a risk measure such as Value at Risk (VaR) is developed using AP as basis because there are instances where the subadditivity condition of coherent risk measures, i.e. $VaR_{AP}(X) + VaR_{AP}(Y) \geq VaR_{AP}(X+Y)$, is not satisfied; see Galeotti et al. (2013). Having said that, whether or not AP might be appropriate for assessing catastrophe bonds at the portfolio level is not examined in the current study. It should also be mentioned that for our purpose, it may be helpful to include the variable AP, thanks to the fact that the correlation between EL and AP appears heterogeneous in this dataset. For example, whilst it is somehow expected that EL and AP are highly correlated, if we were to focus on transactions with large spreads, then the correlation between EL and AP is around 70%, indicating that AP contains information which is not captured by EL for these cases. Moreover, since our study aims at prediction, the addition of an extra variable is not an issue for the random forest. In addition, including AP does not materially affect the performance of LR model in this example either.

With regards to the variable loc_peril, we use a location - peril code categorisation closely in line with the data provided to reflect industry practice. In Appendix A, we provide details regarding all

Table 1
Catastrophe bond data set glossary.

Variable	Description	Type	Role
spread	The amount of interest earned on the top of the risk free rate.	continuous	response
AP	(**A**ttachment **P**robability). The probability of incurred losses surpassing the attachment point. For catastrophe bonds with parametric triggers, AP is translated as the probability that measured parameters will surpass the agreed trigger point.	continuous	predictor
BB spread	U.S. High Yield BB Option-Adjusted Spread for the examined time period computed as the difference between a yield index for the BB rating category and the Treasury spot curve, as in Braun (2016). It reflects the BB rated corporate bond spread, with the BB rating being chosen because, because out of the rated catastrophe bonds, the vast majority of them exhibit this rating. It can be considered as a macroeconomic variable.	continuous	predictor
coverage	Contract term indicating whether protection is offered for a string of loss events or a single loss event.	categorical	predictor
EL	(**E**xpected **L**oss). The annual expected loss within the layer in question divided by the layer size.	continuous	predictor
rating	A dummy variable indicating the credit rating quality of the bond (granular rating), or whether it has not been rated at all.	categorical	predictor
iss_year	The year of issuance of a given catastrophe bond to capture cyclical effects.	continuous	predictor
loc_peril	A location-peril combination.	categorical	predictor
RoL	(**R**ate **o**n **L**ine). Quarterly values of Lane Financial LLC Synthetic Rate on Line Index for the examined time period capturing the level of rates in the ILS, and ILW markets. It can be considered as a macroeconomic variable, see Fig. 8.	continuous	predictor
size	Catastrophe bond nominal amount.	continuous	predictor
term	Years passed from issuance to maturity date.	continuous	predictor
trigger	Mechanism through which a loss payment is activated.	categorical	predictor
vendor	Catastrophe risk modelling software firm.	categorical	predictor

location peril combinations we have considered. Finally, the reason why we have incorporated the issuance year in the predictors set is to account for any other unknown drivers of spread related to a particular issuance year. As an example, one possible instance of such a driver would be the release of an updated model by a risk model vendor which would significantly influence underwriting as it happened in 2011 when RMS released its software Version 11.

To the best of our knowledge, one of the novelties in our study is that we explore the association between coverage type and catastrophe bond spreads. This is in line with current sector discussions as expressed in ILS[3] speciality articles, such as Risk (2019) and Muir-Wood (2017). There, the need to incorporate the coverage type in catastrophe bond pricing was highlighted following the extensive capital freezes investors experienced after California wildfires in 2018. Briefly touching upon this topic, wildfires, a not well understood peril, has been mostly transferred to investors with a provision that losses are covered on an aggregate basis. By design, aggregate deals tend to obtain losses easier, even from small events, compared to their per occurrence counterparts, as a string of loss events triggers the bond. The incapacity of the models to account for this to date led to big losses from aggregate deals and pressure for spreads to incorporate this transaction aspect. This signifies the importance of considering this variable. A further addition into the variables kit for studying the spread is the incorporation of information regarding the modelling company employed to calculate the frequency and severity of the securitised catastrophe risks. The software used for this purpose is firm specific thus it is interesting to explore whether by knowing this information part of the spread can be predicted.

A final note for the variables of this study regards credit ratings. Following Braun (2016), we initially thought to consider whether an issuance was allocated an investment grade by an independent credit rating agency and add an additional categorical value to account for transactions which were not rated as in our data set the majority of catastrophe bonds were issued without a credit rating attached to them. However, as we explain in Section 5, our benchmark model performed better when used granular rating for the rated transactions with the extra categorical value for the non-rated deals - thus our analysis follows this set up. It is worth noting that the absence of credit rating in new issuances is not solely an observation in the current data set. In ILS professional circles, the popularity of non-rated catastrophe bonds is justified from a catastrophe bond market evolution perspective; investors feel more comfortable and trust the risk modelling companies for the calculation of loss and the analysis of the risk return profile more. As a result, credit ratings are somehow no longer seen as essential as they used to be in the past and this is reflected in the increasing issuance pace of non-rated bonds, see ARTEMIS (2019). In the following sections, we choose our benchmark model out of two alternative ones and then apply the research methodology of Section 3 to the catastrophe bond data set that we have just discussed here.

5. Benchmark models

Before we report the random forest generation and prediction accuracy results, we discuss the benchmark models we considered. Even if random forest is not a new approach, it would be helpful to use a benchmark model for its performance assessment and evaluation as its rationale somehow differs from the methods used in most of the previous studies. In search for a benchmark, we looked into the models of Galeotti et al. (2013), Gürtler et al. (2016), Braun (2016), and Trottier et al. (2018), as they are non-fragment[4] and exhibit high out of sample performance. Given

[3] ILS is an abbreviation for Insurance Linked Securities or Insurance Linked Securitisation depending on the context in which it is used.

[4] By non-fragment, we mean that multiple peril - territory coding has been considered.

that the majority of catastrophe bond transactions in our data set are non-rated transactions, we decided to slightly alter the model of Braun (2016) to account for non-rated transactions in addition to those having attached an investment or non-investment grade credit quality tag. Such an alteration allows us to use all 934 observations in our data set. However, the results of the original Braun (2016) model can be found in Appendix C.[5]

As an alternative benchmark model, we also built a new linear regression model (from now on denoted as LR) using the set of variables we consider for the random forest generation, as presented in the previous section.[6] The improved Braun (2016) model is then compared to our LR model by means of in sample overall R^2 and out of sample R^2 resulting from 10 fold cross validation, leave one out cross validation, and bootstrap.[7] The results are presented in Tables 2 and 3.

It appears that LR model outperforms the improved model of Braun (2016) both in terms of in sample and out of sample performance. The overall in sample R^2 for LR is around 85% compared to 80% when using the improved Braun (2016) model. LR also gives consistently better R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} results compared to the improved model of Braun (2016). Consequently, the random forest model will be compared to the more competitive LR regression model when we examine its prediction accuracy.

6. Random forest generation

In order to build[8] the random forest using our catastrophe bond data set, we first needed to decide the hyperparameters' values that we will use, i.e. number of trees, number of variables randomly selected at each split and node size. Breiman (2001) has suggested certain default values that seem to work well after multiple empirical experiments; still we have incorporated certain tuning strategies for the most important hyperparameters. Our approach in choosing these values is explained below.

6.1. Number of trees

The number of trees in the random forest controls its size. Generally, it is good to have a large number of trees as their resulting decisions will be complementing each other more, having a positive impact on random forest prediction accuracy. At the same time, a large number of trees is a safe option in case the optimal value of hyperparameter m_{try} is small so that each variable has enough of a chance to be included in the forest prediction process. However, except for the computational cost which is associated with growing large random forests, it was found by Breiman (2001) that there are diminishing returns in the prediction accuracy increase by adding a bigger number of trees. Taking these reflections into account, we start the random forest development process by growing 2000 trees and in Fig. 2 one can see how the MSE_{OOB} converges for various values of random forest size up to this level (the plot is produced on a logarithmic scale for the ease of readability). From a first sight, it does not take a large number of trees for MSE_{OOB} to stabilise. Before even reaching 100 trees, MSE_{OOB} drops from around 35000 to less than 5500. By the time we reach to 200 trees, it seems that the MSE_{OOB} is almost stabilised. Finally, we find that 500 trees, i.e. the default value that Breiman (2001) suggests, is adequate for our problem as it corresponds to virtually the same R^2_{OOB} as when using 2000 trees and has a much smaller computational cost. Therefore, we choose 500 as the number of trees in the random forest.

6.2. Node size

The hyperparameter node size, i.e. the minimum number of data points in the terminal nodes of each tree, controls the size of the tree in the random forest and effectively determines when the recursive partitioning should stop. A large node size results in shallower trees because the splitting process stops earlier. This has the advantage of lower computation times, but it effectively means that the tree will not learn some patterns resulting in lower prediction accuracy. A small node size translates to a higher computational cost but more thorough learning of patterns and consequently a more accurate base learner. The recommended value for node size given by Breiman (2001) is 5 for regression problems. This default value was also suggested and used by many other authors, as Wang et al. (2018), Grömping (2009), and Berk (2008) and therefore we also employ it as node size value here. The random forest needs to consist of trees which are fully or almost fully grown, see Breiman (2001), thus there is not much added value in exploring this aspect further as 5 meets this requirement and there is a general consensus for its appropriateness.

6.3. Number of variables selected at each split

The number of candidate predictors getting randomly considered at each split, m_{try}, is the most important hyperparameter. This is because it mostly affects the performance of the random forest and the predictors' importance measures, see Berk (2008). The significance of m_{try} lies on the fact that it influences at the same time both the prediction accuracy of each individual tree but also the diversity of the trees in the forest. To get the most out of the random forest, one wants each tree to have good prediction performance but at the same time trees not to be correlated to one another. However, these two goals are conflicting. An individual tree will be the most accurate when m_{try} has a high value but this would result in high correlation for the ensemble. In particular, an extreme case of $\text{m}_{\text{try}} = P$ would force the process to account to simple bagging (James et al. 2013). Generally, a small m_{try} is preferable as, for a sufficiently large number of trees, each predictor will have higher chance to get selected and thus contribute to the forest construction. All in all, the trade-off between individual learner accuracy and diversity needs to be managed by finding an optimal value which secures balance for the data set we study.

In Breiman (2001), the default value of $\text{m}_{\text{try}} = P/3$ (rounded down) is suggested for regression problems. This means that in our problem where $P = 12$, the algorithm would consider 3 predictors at each potential split. We have investigated the relevance of this empirical rule using a tuning strategy called grid search followed by 5-fold cross validation. The goal was to ensure that the most appropriate m_{try} is chosen. The process started by specifying the range of all possible values that m_{try} can take, namely the grid. In the current study, this is between 1 and 12, i.e. as many as the number of predictors. Then, 12 different versions of the random forest algorithm were built one for each possible value of m_{try}.

[5] Both the in sample, and out of sample results of the improved Braun (2016) model are very similar to those of the original Braun (2016) model, even though the improved model performs slightly better.

[6] For an alternative, yet worse performing, linear regression model where, instead of the variable rating as described in Table 1, we include the variable Investment grade (IG) as per the improved model of Braun (2016), see Appendix B.

[7] The reason why we present the bootstrap results here is because it is used as measure of prediction performance in the following sections.

[8] The statistical software used is R, version 3.5.1. The statistical packages employed to perform computations are the following. `randomForest` (Liaw and Wiener, 2002) for developing the random forest as well as calculating permutation importance values, `randomForestSCR` (Ishwaran and Kogalur, 2019) for calculating minimal depth importance measures, and `caret` (Kuhn, 2008) for tuning the main hyperparameter using grid search methodology. It should be mentioned that whenever packages `randomForestSCR` and Kuhn (2008) were used, algorithm arguments used agreed to those used in package `randomForest` to avoid inconsistencies.

Table 2
In sample fit of the improved linear regression model of Braun (2016) versus the new linear regression model LR. Further information on the category levels of the LR variables can be found in Appendix A.

Improved LR model of Braun (2016)	Estimate	Std. error	t value	Pr(> |t|)
(Intercept)	−801.78	41.53	−19.30	0.000 ***
Swiss Re	16.56	13.31	1.24	0.210
RoL index	6.90	0.40	17.52	0.000 ***
BB spread	68.55	8.35	8.21	0.000 ***
Investment grade no (*baseline*)				
Investment grade yes	−180.01	92.74	−1.94	0.050 *
Investment grade nr	62.68	14.74	4.25	0.000 ***
Peak territory	196.42	17.36	11.31	0.000 ***
Expected Loss	1.13	0.02	44.20	0.000 ***
R^2	80.04%			
Adjusted R^2	79.89%			
Res. Std. Error	183.70 (df = 926)			
F Statistic	530.60 (df = 7; 926)			
LR	**Estimate**	**Std. error**	**t value**	**Pr(> |t|)**
(Intercept)	55940.00	9396.00	5.95	0.000 ***
RoL	6.25	0.41	15.10	0.000 ***
BB spread	46.24	8.39	5.51	0.000 ***
term	−28.83	8.57	−3.36	0.001 ***
size	0.00	0.00	2.521	0.012 *
trigger industry loss index (*baseline*)				
trigger indemnity	−21.10	14.49	−1.46	0.146
trigger model	−69.30	38.66	−1.79	0.073
trigger multiple	−44.55	42.70	−1.04	0.297
trigger parametric index	−28.88	40.60	−0.71	0.477
trigger parametric	−167.60	36.55	−4.59	0.000 ***
coverage aggregate (*baseline*)				
coverage both	52.27	82.10	0.64	0.525
coverage occurrence	−49.45	13.56	−3.65	0.000 ***
vendor AIR (*baseline*)				
vendor AON	98.50	87.51	1.13	0.261
vendor EQECAT	−0.82	32.80	−0.03	0.980
vendor pp	9.08	71.18	0.13	0.899
vendor RMS	27.97	18.73	1.49	0.136
AP	−13.99	5.90	2.37	0.018 *
EL	1.26	0.09	14.91	0.000 ***
iss_year	−27.95	4.65	−6.01	0.000 ***
APAC_Quake (*baseline*)				
loc_peril APAC_Typh	−63.60	43.11	−1.48	0.141
loc_peril Europe_APAC_Multi_Peril	−105.00	125.50	−0.84	0.403
loc_peril Europe_Quake	8.94	55.94	0.16	0.873
loc_peril Europe_Wind	−148.20	39.36	−3.77	0.000 ***
loc_peril NA_APAC_Multi_Peril	55.08	47.07	1.17	0.242
loc_peril NA_Europe_APAC_Multi_Peril	139.00	41.14	3.38	0.001 ***
loc_peril NA_Europe_Multi_Peril	118.10	39.79	2.97	0.003 **
loc_peril NA_Multi_Peril	158.90	27.33	5.82	0.000 ***
loc_peril NA_Quake	−23.64	34.40	−0.69	0.492
loc_peril NA_Wind	86.64	29.74	2.91	0.004 **
loc_peril SA_Quake	139.90	103.80	1.35	0.178
rating B (*baseline*)				
rating BB	−146.80	18.05	−8.13	0.000 ***
rating BBB	−346.70	83.25	−4.16	0.000 ***
rating CCC	−45.11	94.25	−0.48	0.632
rating nr	−2.09	18.90	−0.11	0.912
R^2	85.07%			
Adjusted R^2	84.52%			
Res. Std. Error	161.10 (df = 900)			
F Statistic	155.40 (df = 33; 900)			
Note for signif. codes:	*p < 0.1; **p < 0.05; ***p < 0.01			
Observations number:	934			

Table 3
Out of sample performance measured in terms of R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} for the improved linear model of Braun (2016) versus the new linear regression model LR.

Model	R^2_{OOB}	R^2_{10CV}	R^2_{LOOCV}
Improved Braun (2016)	79.71%	80.81%	79.40%
LR	83.30%	84.42%	83.84%

The prediction accuracy of each random forest version, measured by means of R^2_{OOB}, was evaluated through a 5-fold cross validation.

The results, shown in Fig. 3, reveal that there is considerable improvement in random forest performance when the m_{try} value is increased from 1 to 2, and then 3 to 4. No real advantage in terms of prediction accuracy seems to be yielded from further increasing the m_{try} value above 4 which also happens to be the default value $m_{try} = 12/3$ as per the suggestion of Breiman (2001). Moreover, since variable importance measures are to be calculated later on, we deem preferable to choose the smaller value of $m_{try} = 4$, by discipline, as this would lead to less correlated trees giving the

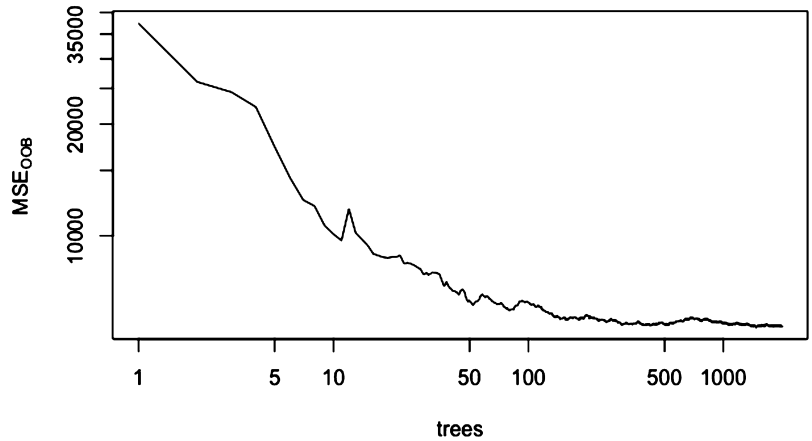


Fig. 2. Out of bag mean squared error convergence with respect to random forest size. The line corresponds to the mean squared error based on out of bag samples (MSE_{OOB}) versus the number of trees in random forest. The plot is produced on a logarithmic scale for the ease of readability.

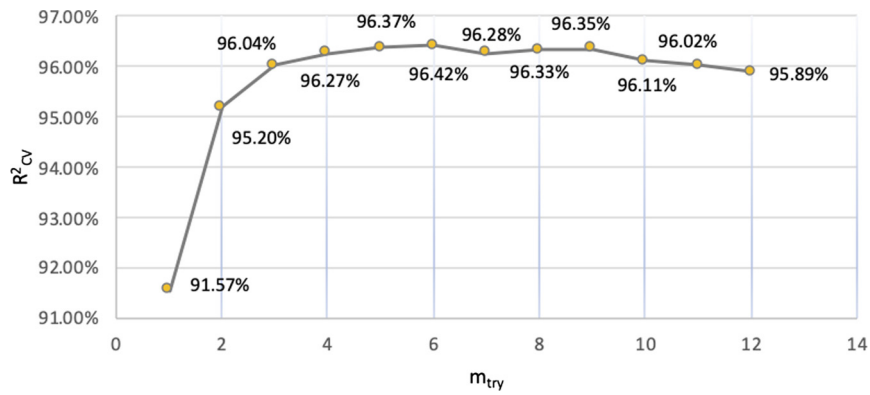


Fig. 3. Tuning of main random forest hyperparameter through grid search followed by 5-fold cross validation. Out of bag based R^2 (R^2_{OOB}) for random forest versus number of candidate predictors getting randomly considered at each split (m_{try}) during forest generation.

Table 4
Description of final random forest in terms of sample size, predictors number, and hyperparameters values.

Final random forest description	
sample size	934
number of predictors	12
random forest type	regression
number of trees	500
no. of variables tried at each split (m_{try})	4
node size	5

opportunity to see the influence of weaker predictors to catastrophe bond spreads prediction. Also, a smaller m_{try} value would lead to a simpler model which would be less costly in terms of computational time. Having decided on the hyperparameter values, the final random forest was generated and a summary description is provided in Table 4. The next section investigates how well the random forest performed in our catastrophe bond setting.

7. Random forest performance evaluation

In this section, we evaluate how well our random forest performs with regards to its prediction accuracy and stability.

7.1. Random forest prediction accuracy

As mentioned in Section 3.3.1, the ability of the random forest to predict catastrophe bond spreads on new inputs is investigated from both a non-temporal and a temporal point of view. In the former case, the prediction accuracy metrics we consider are R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} whilst in the latter case R^2_{OOS} is used to assess the out of sample performance. The prediction accuracy results of the random forest versus the benchmark model are presented and discussed below for each of the two perspectives.

7.1.1. Non-temporal prediction accuracy

We start by clarifying what we regarded as new inputs followed by how the catastrophe bond spread predictions were made for the computation of R^2_{OOB} as it may appear to be a less standard approach (especially for the benchmark model) compared to 10 fold, and leave one out cross validation.

Starting from the random forest, as new inputs for a given tree, we have accounted its out of bag observations. Due to the property of sampling with replacement, only around two thirds of $N = 934$ data points were used to build each of the 500 unpruned and almost fully grown (node size = 5) regression trees. For a given tree, the remaining one third of $N = 934$ data points were never used during the building process and as a result they formed a reliable test set for it. Secondly, a prediction for the spread at issuance for the $n = 1$ observation, $\hat{y}_1$, was produced by dropping its corresponding input down every single tree in which the $n = 1$ observation was out of bag. This resulted on average to around one third of 500 catastrophe bond spread predictions for the $n = 1$ observation. Then, a single spread prediction for the $n = 1$ observation was made by taking the average value of these predictions. After having predicted the catastrophe bond spread value for the observation $n = 1$, the same process has been repeated for the $n = 933$ observations left. Finally, in order to evaluate the prediction accu-

Table 5
Prediction accuracy performance measured in terms of R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} versus in sample performance measured in terms of R^2 for random forest (RF), and linear regression (LR).

Model	R^2_{OOB}	R^2_{10CV}	R^2_{LOOCV}	R^2
RF	96.57%	96.49%	96.59%	99.25%
LR	83.30%	84.43%	83.84%	84.52%

racy of our random forest, the metrics discussed in Section 3 were calculated. In particular, we have computed the mean squared error based on the out of bag data as $\text{SE}_{\text{OOB}} = \sum_{n=1}^{934}(y_n - \overline{\hat{y}}_{n_{\text{OOB}}})^2$, the total sum of squares as $\text{TSS} = \sum_{n=1}^{934}(y_n - \overline{y})^2$ and, the variability explained by our random forest as $\text{R}^2_{\text{OOB}} = 1 - \frac{\text{SE}_{\text{OOB}}}{\text{TSS}}$.

With respect to the benchmark model, the calculation of R^2_{OOB} was done as it was described in the case of random forest; we used 500 bootstrap samples to refit the model and for each observation, we only considered predictions from bootstrap samples not including that observation. The results for the prediction accuracy metrics both for the random forest,[9] and the benchmark model are presented in Table 5. Note that in Table 5, we have also included the in sample R^2 for both models as reference.

It stands out that our random forest explains more than 96% of the total variability in the non-temporal context no matter whether using the bootstrap or one of the other two cross validation methods. At the same time, the non-temporal predictive performance of our benchmark model, i.e. linear regression, is lower - the highest total variability it explains across all metrics is 84.43%. Once again the prediction accuracy results for the benchmark model are very similar across different resampling methods. As a result, from now on we will be focusing on the out of bag related metrics in the non-temporal context.

With regards to the in sample R^2, it seems that random forest may lead to some degree of overfitting which is expected as the individual regression trees are fully grown. The latter signifies the fact that measuring performance in terms of R^2 for random forest in this instance might not be as appropriate as in the case of linear regression.[10]

An important aspect is to evaluate whether the 96.57% random forest non-temporal prediction accuracy is high enough given the nature of the problem under study. On a broader perspective, making predictions in a financial market setting is not an easy task. Inefficiencies, multiple market participants and, the influence of psychology on their behaviour are only few of the factors making the prediction task complex. Consequently, one might claim that achieving an R^2_{OOB} of more than 96% here corresponds to a very satisfactory level of prediction accuracy. Of course this also holds true for the around 84% benchmark model prediction performance but since there is a considerable difference in the reported R^2_{OOB}, we would conclude that using random forest in a non-temporal context may be preferable.

7.1.2. Temporal prediction accuracy

In a temporal context, we focus on the forecasting ability of the random forest versus the linear regression model over time. In this case, the training data is not picked randomly thus we are able to assess robustness towards potential regime shifts. Regime shifts in the catastrophe bond market can include regulatory changes, issuances with unusual features, demand forces etc but their identification for the time period we study is beyond the goals of this study. Our aim here is simply to examine the degree by which the trained models (random forest, and linear regression) can accurately predict catastrophe bond spreads even in presence of such changes. The temporal prediction performance challenge between random forest and linear regression is designed by using the train-test data set split approach in eight cycles of operation so that we can have a more complete picture of how models performance compare as the catastrophe bond market evolves. We start by using as train data set, the data from December 2009 to December 2010. We fit the random forest and linear regression models to this train data set and we make spread predictions using data from 2011. The second cycle of operation includes adding bonds from 2011 into the train data set and using bonds from 2012 as test set. The aforementioned process is repeated up until the train data set reflects the period up to 2017 and the test data set includes the catastrophe bond issuances in 2018. The prediction accuracy results measured in terms of out of sample R^2 (R^2_{OOS}) for each cycle of operation are presented in Table 6. It should be mentioned that assessing the prediction performance on a temporal context has a particular limitation. That is, new observations in a given test set cannot (directly) include categorical variable levels which did not appear in the respective train set. Thus, in order to avoid deleting deals having new levels in some of the categorical predictors in any given test year, we have imputed these values based on the most commonly observed categories in the corresponding training set accordingly.

Table 6
Prediction accuracy of random forest (RF) versus linear regression (LR) measured in terms of R^2_{OOS} for various train-test sets by issuance year.

Train set	Test set	RF R^2_{OOS}	LR R^2_{OOS}
2009-2010	2011	64.42%	$< 0.00\%$
2009-2011	2012	58.04%	71.36%
2009-2012	2013	45.64%	16.84%
2009-2013	2014	88.74%	72.90%
2009-2014	2015	55.12%	70.47%
2009-2015	2016	84.23%	89.55%
2009-2016	2017	91.24%	91.06%
2009-2017	2018	88.59%	91.69%
Average R^2_{OOS} across all train sets		72.00%	62.98%

We note that there seem to be some noticeable regime shifts especially in the first few cycles of operation. However, we cannot be definite about which model, the random forest or linear regression, handles regime changes best as in some years random forest does better than linear regression and vice versa. By looking at the variability of the R^2_{OOS} across all cycles for both models, it appears that the R^2_{OOS} range for random forest is between 45% and 91% whilst the respective range for the linear regression model is around between below 0% and 91%. It should be noted that in the first year, LR exhibits a very low temporal predictive power but we believe that this may be the result of the imputation in a small data set sample and perhaps the fact that some categorical variables contain quite granular information; see Table 13 in Appendix A. It is worth mentioning that the worst performance for both models is observed for the 2013 test set. By looking into how the regression model is parameterized for the test sample in 2013, it appears that the poor performance of the regression model on this test sample is largely due to the fact that catastrophe bonds in 2013 indicated record high EL values, i.e. the largest of them almost doubled the maximum EL value that was observed prior to 2013, for which models based on earlier observations might not be entirely suitable for the purpose of prediction. A potential reason

[9] In a robustness check, see Appendix D, we provide the prediction accuracy of the random forest when the categorical variables in the catastrophe bond data set are pre-processed with categorical dummies as in the case of linear regression. As we see the two random forest versions, i.e. with and without dummies, lead to very similar results.

[10] It should be noted that this problem is not specific to random forest but more general and lies in the use of in-sample performance measures in case of over-fitted models and therefore can also be seen in a linear regression context.

Table 7
A typical realisation regarding random forest prediction accuracy stability results.

Random forest summary	Sample A	Sample B
sample size	467	467
number of predictors	12	12
random forest type	regression	regression
number of trees	500	500
no. of variables tried at each split	5	6
node size	5	5
MSE_{OOB}	13855.34	14744.59
R^2_{OOB}	92.14%	90.71%

Table 8
Random forest stability measured in terms of minimum, mean, and maximum absolute difference of R^2_{OOB} between Sample A and Sample B across 100 iterations.

R^2_{OOB} Min Abs. Dif.	R^2_{OOB} Mean Abs. Dif.	R^2_{OOB} Max Abs. Dif.
0.01%	2.19%	6.92%

is the largest version change in the history of RMS model, implemented towards the end of 2012, which affected all 2013 renewals in having the potential to increase insured loss results even above 100% in some cases. Overall, it appears that the random forest is relatively more robust than linear regression in this respect. More comparisons between RF and LR when taking into account missingness of more than one predictor at a time follow in Section 8.5.

7.2. Random forest stability

We now examine the stability of random forest prediction accuracy results over the entire time period of interest. This is measured empirically from a practitioner's point of view as presented in Section 3.3.2 and in Table 7, we present a typical realisation of 1 out of 100 iterations with respect to the repeatability of prediction accuracy results.

As we observe in Table 8, across all 100 iterations, the recorded mean absolute difference of R^2_{OOB} between Sample A and Sample B for the catastrophe bond data set is 2.19% with the minimum and maximum absolute differences being 0.01% and 6.92% respectively.[11] Given that our problem sits in the intersection of financial and insurance market spheres where many behavioural aspects can affect prices, we consider the reported difference for the catastrophe bond data set being small. In essence, it is unlikely that an ILS fund would reject the use of the method solely for such a level of dissimilarity. In fact, the repeatability of prediction results here means that we can fairly safely say that our initial random forest prediction accuracy result, i.e. of an R^2_{OOB} of 96.57% presented in Table 5, is reliable, in the non-temporal context.

This finding is beneficial for the usage of the method in the industry. With new catastrophe bonds being issued, the random forest would need to be validated at some point in time as any other model in an insurance related firm. Surely, in a business context, there is no point in investing time and capital to introduce a new model if the latter provides accurate predictions strictly for one particular data set. Having gone through the examination of prediction accuracy results stability, we proceed with determining the importance of each independent variable in the study.

[11] As a robustness check, we have also repeated the random forest stability evaluation using two popular Open Source data sets, namely Boston Housing and Abalone, which are also used in the original paper of Breiman (2001) for the empirical assessment of the random forest method and are available at the UCI repository. The results are close with those derived for the catastrophe bond data set.

8. Predictor importance analysis

The importance of predictors is assessed using the methodologies of permutation and minimal depth importance presented in Section 3. It should be highlighted that the goal here is to find how powerful each independent variable is in predicting catastrophe bond spreads at issuance. No kind of relationship between spread at issuance and the predictors is to be established - the focus lies solely on their prediction ability. We then compare the stability of predictors' importance results for both methods. Then based on the ranking of the most stable predictors importance method, we examine the sensitivity of the random forest versus the benchmark to simultaneous missingness of multiple predictors in an effort to reveal and understand variables interactions. Next, by considering once again the most stable importance method, we examine the degree of similarity in predictors importance results in the predictive versus explanatory modelling frameworks. Finally, we discuss whether the rankings make empirical sense from investors' viewpoint.

8.1. Permutation importance

The importance of each independent variable in predicting catastrophe bond spreads has been here assessed on the basis of a percentage increase in MSE_{OOB} when a predictor is randomly permuted from the out of bag data whilst others remain untouched. First, the MSE_{OOB} for each of the 500 trees comprising the random forest, was recorded. The same process was repeated after randomly shuffling the values of a particular x_p across all observations. Then, the change between these two mean squared errors, before and after x_p permutation, has been calculated and averaged across the 500 trees after being normalised by the standard deviations of the differences. In this way, the importance score for x_p has been derived. Finally, based on these scores, an importance ranking has been produced. The ranking of catastrophe bond predictors based on their permutation importance score is shown in Fig. 4. Variables higher on the vertical axis are more important in predicting catastrophe spread at issuance with respect to this measurement.

One of the first observations is that all scores have positive value, indicating that each of the independent variables presented here does contribute towards prediction of catastrophe bond spreads. The predictors EL and RoL followed closely by term appear as the most important predictors of spread at issuance. In particular, when EL is shuffled, the out of bag mean squared error increases by around 41% whilst the respective percentages for RoL and term are slightly lower between 33% and 34%. Next, had any of the predictors; loc_peril and AP been randomly permuted, the prediction performance of the random forest would have been deteriorated between 31% and 32%. By shuffling the predictor iss_year, we see an almost 28% decrease in random forest prediction accuracy whilst the respective percentages for BB spread and size are in the range between 27% and 28%. Rating contributes to the reduction in the prediction accuracy of the random forest by around 19% and the least important predictors are coverage and vendor resulting in an approximately 16% and 13% prediction accuracy decrease respectively.

8.2. Minimal depth importance

The focus is now shifted from using a specific prediction performance measure to assess variables importance to a criterion based on the way that the forest was constructed, namely, the minimal depth. A tour over the constructed random forest was made to find

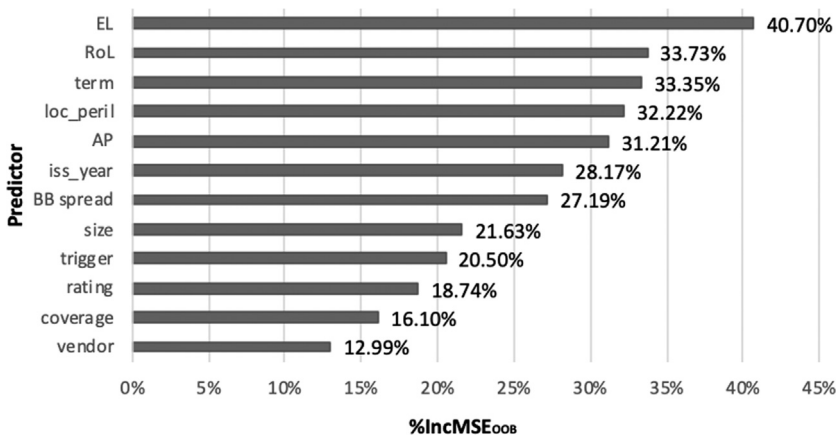


Fig. 4. Permutation importance based ranking of predictors. Predictors being permuted versus percentage increase in MSE_{OOB} as a result of the permutation.

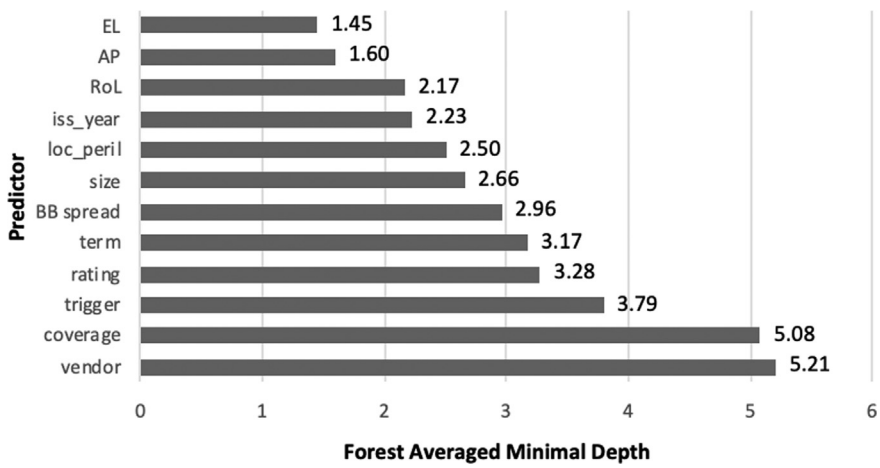


Fig. 5. Minimal depth importance based ranking of predictors. Predictors and their forest averaged minimal depth.

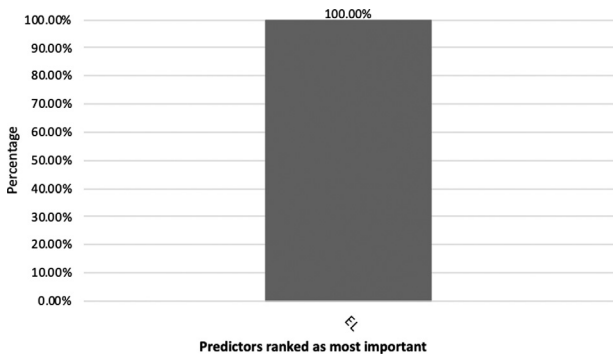
the maximal subtree[12] within each of the $K = 500$ trees for a particular x_p predictor. From there, the minimal depth for x_p within each tree was identified following the rationale explained in Section 3. Then, the forest level minimal depth for x_p was derived by averaging the minimal depth for x_p within each tree among all 500 trees. Fig. 5 illustrates the ranking of the covariates with respect to their average minimal depth; higher values of minimal depth correspond to less predictive variables.

Predictors EL and AP, with random forest average minimal depths of 1.45 and 1.60 respectively, have the largest impact in predicting catastrophe bond spreads. In particular, such small values of minimal depth demonstrate that these two variables were mostly used to split either the root node or any of its child nodes at least in most of the trees in the forest. Straight after in rankings comes the variable RoL followed closely by iss_year which on average were chosen to split a node for the very first time at a depth equal to 2.17 and 2.23 respectively. At similar level of importance stand the loc_peril and size with a level of depth still closer to 2.00 rather than 3.00 implying that they also have a considerable forecasting power. It appears that predictors BB spread, term, rating, and trigger were on average chosen to split the third node in the regression trees comprising the random forest. The aforementioned predictors appear as not being as powerful because they split nodes which naturally have less data points due to their proximity to the terminal nodes. Then the remaining variables, coverage and vendor, have minimal depth measurements of 5.08 and 5.21 respectively. These values are the highest among all predictors, revealing that coverage and vendor have the most limited forecasting ability out of all predictors.

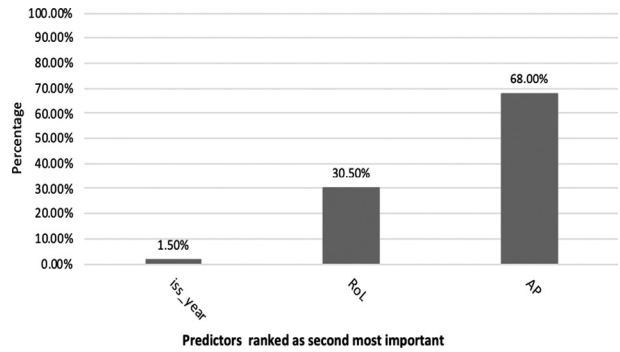
8.3. Divergence between permutation and minimal depth importance results

Permutation and minimal depth importance procedures presented for ranking or selecting catastrophe bond spread predictors above are not directly comparable. This is because, as it has been seen, each of them follows a different approach in defining and quantifying the importance in prediction. However, empirically we would expect that there should be some consensus between the two methods. What we see is that whilst there is indeed a degree of agreement for the very top and bottom of the rankings, there is some divergence at the upper middle ranks. This realisation makes us think which of the two variable importance approaches leads to the most trustworthy results for our catastrophe bond spread prediction problem. Indeed, empirically, an answer to this question would be to examine which ranking makes more sense from a practitioner's perspective. However, we believe that it is also preferable to bring our attention back to the concept of stability, but this time for the catastrophe bonds features importance. If one of the two methods is unstable, then we can shift our focus to the other one that is more robust and then discuss whether the ranking it provides makes sense from an investor's perspective. In the following, we present the stability checks for the importance results derived by both permutation and minimal depth importance.

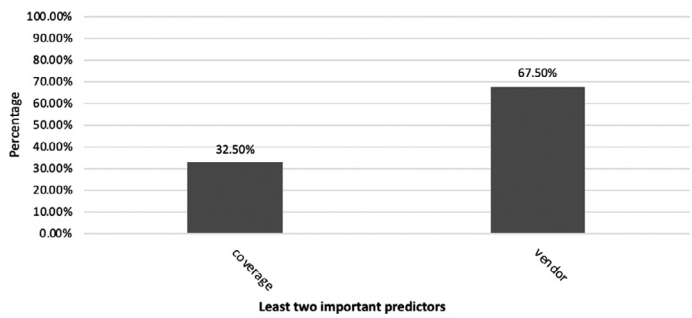
[12] See Section 3.4.2 for an explanation of what constitutes a maximal subtree.



(a) Predictor ranked as most important



(b) Predictors ranked as second most important



(c) Least two important predictors

Fig. 6. Bar plots showing the percentage frequency where a given predictor was ranked as top, second from top and in the last two positions for the most stable variable importance method in terms of ranking, i.e. minimal depth.

Table 9
Stability of ranking of predictors by different importance ranking method.

Ranking position	Agreement % (Permutation)	Agreement % (Mimimal depth)
Top	98%	100%
Second from top	36%	46%
Second from bottom	27%	69%
Bottom	22%	69%
Last two	10%	100%

8.4. Stability checks for predictors importance results

Here a predictor ranking method will be considered reliable if its importance ranking for catastrophe bond spread predictors is fairly robust to certain type of changes in the data set, such as random splitting. If a change in the catastrophe bond data set from which the random forest is constructed lead to a big change at the top and at the bottom of predictors importance rankings, then that particular importance ranking method will be considered unstable and thus probably unreliable.

Towards this direction, since both permutation importance and minimal depth importance are procedures derived internally after the construction of the random forest, the stability of permutation and minimal depth importance has been mainly examined based on the 100 random forests pairs grown out of 100 Sample A and Sample B pairs which have been previously used when the stability of the random forest was investigated in Section 7.2. In Table 9, we report by variable importance method, the percentage of times where there was an agreement between Sample A and Sample B in the predictor chosen at the top, second, and bottom positions of the ranking for all data sets. As bottom positions of the rankings we consider the last two positions jointly. This is because we understand that the further we go down the ranking, the more susceptible variables may jump from the position to its neighbours across different iterations. It is evident that minimal depth importance method provides more stable ranking results for both top, and bottom positions compared to permutation importance method. The biggest differences between the two methods are recorded for the second from bottom, bottom, and last two ranking positions combined where the discrepancy in the agreement percentage reaches 42%, 47% and 90% respectively. As previously highlighted in Chen and Ishwaran (2012), the complex randomisation element of permutation importance procedure makes it difficult to assess the underlying cause for it being relatively more unstable. However, it should be mentioned that this is not the first work when this measure showed an irregular conduct. As an example from bioinformatics, Calle and Urrea (2010) showed that permutation importance rankings were unstable to small perturbations of a gene data set related to the prognosis bladder cancer. All in all, it should be acknowledged that the appropriateness of a feature importance method is mostly data set specific and at least for the catastrophe bond set in hand it seems that permutation importance is not as reliable.[13] Based on the above, any discussion from now on which is relevant to predictors importance will be based on results of minimal depth importance as presented in Section 8.2.

Moving forward, it is interesting to examine stability within the minimal depth importance output with respect to which variable is chosen at a given position of the minimal depth importance rankings. In order to do so, we considered the number of counts out of 200 sub-samples taken in 100 iterations (or 400 samples taken in 100 iterations when we consider the last two ranking positions

[13] We have also examined the robustness of the predictors importance stability results using the Boston Housing and Abalone Open Source data sets, as we did in the case of random forest stability evaluation. The results align with those derived for the catastrophe bond data set, i.e. the minimal depth method, at least for the top ranking positions, appears to be more reliable compared to the permutation importance one.

Table 10
Sensitivity analysis for random forest (RF) versus linear regression (LR) to missing predictors using R^2_{OOB} as performance measure. The performance is examined by removing predictors sequentially based on the minimal depth importance ranking. Here we also report the R^2_{OOB} of RF and LR without any missing predictors to facilitate comparison.

Missing predictors	RF R^2_{OOB}	LR R^2_{OOB}
no missing predictors	96.57%	83.30%
EL	95.74%	79.98%
EL, AP	87.69%	50.68%
EL, AP, RoL	87.86%	47.30%
EL, AP, RoL, iss_year	84.83%	43.65%
EL, AP, RoL, iss_year, loc_peril	84.56%	33.33%
EL, AP, RoL, iss_year, loc_peril, size	68.81%	30.76%
EL, AP, RoL, iss_year, loc_peril, size, BB spread	30.86%	21.18%
EL, AP, RoL, iss_year, loc_peril, size, BB spread, term	18.89%	17.69%
EL, AP, RoL, iss_year, loc_peril, size, BB spread, term, rating	12.23%	10.40%
EL, AP, RoL, iss_year, loc_peril, size, BB spread, term, rating, trigger	7.60%	7.69%
EL, AP, RoL, iss_year, loc_peril, size, BB spread, term, rating, trigger, coverage	5.67%	6.03%

Table 11
Sensitivity analysis for random forest (RF) versus linear regression (LR) to missing predictors using R^2_{OOB} as performance measure. The performance is examined by randomly removing M predictors from the original data set for $M = 1, \ldots, 11$. For each M, this experiment is repeated 100 times, and the average R^2_{OOB} is reported. Here we also report the R^2_{OOB} of RF and LR without any missing predictors to facilitate comparison.

Number of missing predictors at random - M	RF R^2_{OOB}	LR R^2_{OOB}
no missing predictors	96.57%	83.30%
1	96.21%	82.47%
2	95.87%	80.58%
3	95.19%	78.19%
4	92.84%	73.56%
5	91.01%	70.78%
6	89.28%	67.34%
7	73.76%	59.74%
8	67.49%	51.74%
9	60.08%	43.45%
10	45.02%	29.22%
11	28.09%	16.97%

jointly), where a given predictor was ranked as top, second from top, or in last two positions in terms of importance by variable importance method. The results are shown in Fig. 6 in terms of percentage frequency. We see that minimal depth method is also fairly stable with regards to its predictors' choices for the examined ranking positions. That said, in the top position the predictor EL was chosen 100% of the times and only a small variation is visible for the second from top and last two ranking positions. In the next section, we provide some further analysis on how well the random forest handles missingness of important predictors as opposed to LR model.

8.5. Further analysis - on handling missingness of important variables

We now assess the sensitivity of prediction accuracy of random forest in the absence of important predictors, and contrast the outcomes with those from the benchmark model. Doing so also allows us to understand and characterise interactions between predictors. Here we consider removing more than one predictor each time and then report the resulting prediction accuracy of both random forest and the benchmark model. The removal of predictors is made firstly, sequentially based on the minimal depth ranking presented in Section 8.2, from the most important one to the least important one, and secondly, by (uniformly) randomly dropping M predictors from the original data set for $M = 1, \ldots, 11$. For each M, the second experiment is repeated 100 times, with the average R^2_{OOB} computed for both RF and LR. The sensitivity results are presented in Table 10 and Table 11.

When predictors are removed sequentially according to the minimal depth ranking, it appears that random forest prediction accuracy results seem to be considerably more robust compared to the ones derived from LR when the most important predictors, as identified in the minimal depth analysis, such as EL and AP, are jointly missing. For example, when the most important predictors EL and AP are excluded from the analysis, the RF prediction accuracy drops by around 8% as opposed to 29% in the case of LR compared to the respective prediction performances when only EL, i.e. the most important predictor, is missing. This may be an indication that there are potentially interactions, as well as non-linearities, between the predictors, which random forest appears to be capturing whereas the linear regression model struggles. Another observation is that we see a significant drop in random forest prediction accuracy when size and even more so BB spread are included in the missing predictors set. In particular, when size is removed, RF prediction accuracy deteriorates by 16%, i.e. the biggest drop up to this point since the beginning of the minimal depth based sequential removal of predictors. When BB spread is excluded, RF prediction accuracy declines by an additional 38% which is the highest drop in RF prediction accuracy across the whole experiment. A potential interpretation is that there is a certain degree of information redundancy among all the predictors. Here the predictors size and BB spread contain a large amount of useful information of all its predecessors found to be of higher importance in catastrophe bond spread prediction, which can be effectively extracted by random forest.

Similar observations are made when randomly removing M predictors from the original data set for $M = 1, \ldots, 11$ repeated 100 times and taking the average of R^2_{OOB} for the RF and LR respectively. RF still shows a better predictive performance than LR having an average R^2_{OOB} of around 90% even by randomly dropping half of the variables, again forcing the impression that RF is more flexible than LR and is likely better at capturing interactions and dealing with possible missingness of the predictors. It should be noted that thanks to the random dropping mechanism, the results in Table 11 appear smoother than these in Table 10 where we exclude the most important variables first - a strategy which acts more like assessing the worst case scenario. In summary, random forest is better at borrowing strength from existing predictors to (partially) recover the predictive power lost due to the absence of important predictors.

8.6. Predictive versus explanatory importance

Now we discuss whether the importance results in our predictive framework agree with those presented in explanatory models of past works but also the LR model in the current study.

As mentioned in Shmueli (2010), variables which are considered important in explaining the response are tied to theoretical hypotheses which are set at the beginning of the study, and on

the notion of statistical significance. These aspects are immaterial in a purely predictive modelling framework as the one we present by using random forests. Exploring the level of this divergence is meaningful, as it can add value in understanding the full spectrum of catastrophe bond spread drivers for both prediction, and explanation. It should be mentioned that this is an exercise that shall be made with extra caution as, to our best knowledge, every study in the explanatory catastrophe bond pricing literature to date and our predictive study has utilized different data sets and made different assumptions (apart from the LR model). However, given the fact that satisfactory level of agreement has been recorded in the past for certain variables in the explanatory framework, even under these constraints, it merits a short discussion.

The starting point is independent variables where harmony with respect to predictive and explanatory importance between this and previous studies has been observed. In particular, in Section 8.2, it is seen that EL is the most major contributor in predicting spreads in the primary catastrophe bond market. This result comes in agreement with our LR model presented in Table 3, and the majority of the previous explanatory oriented literature, see Lane (2000), Lane and Mahul (2008), Bodoff and Gan (2009), Dieckmann (2010), Braun (2016), Galeotti et al. (2013), and Jaeger et al. (2010). In Lei et al. (2008), the conditional expected loss is considered instead of expected loss, despite the fact that the former is not found to be statistically significant, while other variables related to the loss distribution are.

At the same time, in this study we observe that the probability of losses outstripping the attachment point has almost equal forecasting power as the expected loss. Moreover, we see that the predictor AP is statistically significant in LR too. Lane (2000) also supports that the catastrophe bond premium is derived through an interplay between frequency and severity of catastrophe bond expected losses. On the top of this, Lei et al. (2008) and Jaeger et al. (2010) agree with the view that the attachment probability is of high significance in explaining catastrophe bond spreads. Moving forward, the importance of variables reflecting the cyclicality of the market is high both in a predictive, and explanatory context, see LR, Lane and Mahul (2008), and Braun (2016). At the same time, peril-territory combination which is found particular importance for its ability to forecast spreads here and in the explanatory framework. In particular, alike results are obtained by LR, Gatumel and Guegan (2008), Jaeger et al. (2010), and Götze and Gürtler (2018). Similarly, trigger is predictive in the current research whilst Dieckmann (2010), Götze and Gürtler (2018) and Papachristou (2011) also commented about the explanatory significance of this variable in their models. Finally, the predictor rating which is found to be predictive in our study (although not of top importance), is seen as major determinant of spread in our LR model, and also in Lei et al. (2008), and Götze and Gürtler (2018); even though Götze and Gürtler (2018) have examined rating from a different perspective to the one we employ, i.e. the variable related to rating does not refer to the credit quality of the bond but to that of the cedent instead.

With respect to the predictor term, no general consensus on its statistical significance has been reached in the literature up until now, although here it appears to be relevant for both prediction and explanatory purposes as LR reveals. For example, Papachristou (2011) and Braun (2016) exclude the variable term from their analysis whilst on the other hand Dieckmann (2010), Galeotti et al. (2013), and Gürtler et al. (2016) highlight its importance. At the same time, the predictor size is minded as less influential or not significant at all by the models of Papachristou (2011), Lei et al. (2008), Braun (2016), and LR (zero coefficient even if the variable is significant) but it is considered sufficiently important for prediction purposes in our study. This divergence may once again stem from the way weak predictors are treated in a typical linear regression model versus random forests. As it is mentioned by Berk (2008), in a traditional regression framework a variable having a very small association with the response is most often excluded from the model being regarded as noise. Nevertheless, a big number of small associations when considered not on an individual basis but on an aggregate level can have a substantial impact on fitted values. That is not to say that linear regression is not capable of capturing interactions, however to do so any interactions need to be explicitly specified - a complicated task when the number of predictors in the study starts increasing. On the contrary, random forests, as a tree based method, is naturally able to capture associations between predictors without the need to specify them. Indeed, Papachristou (2011) also acknowledges that in the context of his study, the fact that the term is not considered as important enough to be included in the suggested model may be due to the challenge of capturing complex effects between covariates. Coming back to the discrepancy between explanatory and predictive power for predictor size, we recall that in Section 8.5 the interacting behaviour of this variable is also observed in the predictive framework.

Finally, our study indicates that the variables vendor and coverage are predictive despite of their appearance at the bottom of the ranking. Since this is the first time that these variables are studied, we can only compare them with LR in the explanatory framework. In particular, vendor does not appear as a statistically significant variable whilst coverage is. Overall, we can conclude that explanatory (based on LR and past literature) and predictive power appears to coexist for all catastrophe bond spread drivers considered in our study apart from size and vendor.

8.7. Discussion of predictors' importance results from an industry perspective

Looking broadly at the minimal depth ranking presented in Fig. 5, we observe that the predictors may fall into three groups: those of utmost (the top two), medium (the next seven) and low prediction strength (the last two). We acknowledge that the bounds of where medium and lowest importance variables groups start may be subjective. The distinction here is made looking at the ranking from the perspective of a practitioner. The reason why we want to avoid focusing on individual importance scores is that explaining results in such a detailed way would neither be appropriate nor meaningful for a prediction oriented study. This section is not about interpreting results but seeing whether the results capture somehow investors' perception and knowledge of the market.

Having explained our rationale, the group of top importance predictors comprises from the two fundamental ingredients in any risk quantification process, that is the product of severity and frequency of losses, i.e. EL, and AP. This is something that would most probably not surprise insurance professionals, risk managers or even investors if the variable importance results were to be presented to them. Especially with respect to investors, it is well comprehended that the return to be earned by investing into a catastrophe bond deal needs to surpass the expected value of catastrophe bond payouts. Thus, from an empirical viewpoint, investors would expect that by knowing the expected loss and probability of them losing the first dollar, at least a part of the spread value can be predicted.

The second group refers to some cyclical market elements and catastrophe bond features which could influence investors' interest in a deal. The high importance of cyclical aspects in the prediction of a new issuance spread is somehow natural since a hard or soft market directly sets some bounds on the top of which a deal's specific loss profile and characteristics would be assessed. One reason why certain catastrophe bond features could influence

an investor's appetite considering a deal, is the effect that these features could have on investors' portfolio returns. In particular, investors would most probably agree with the predictor loc_peril having a high position in the ranks, as this type of information acts as the window shop for them entering the transaction. The rarity of the peril combined with the coverage territory indirectly informs investors about the diversification effect that the particular security can bring into their portfolio; a significant incentive for them to invest in this asset class. We acknowledge that this may not be true for new or rare perils, for which the existing catastrophe models are not yet trusted, however even in this case the peril-territory combination is informative in this sense. Another reason why the predictors of the second group could trigger investment interest is because some of these features are typical in traditional bond types traded in the financial markets and investors are already accustomed to this type of information such as issuance size, BB spread level, time between issuance and maturity date, credit rating related information, and trigger of payment. Consequently, one can say that the location of these variables in the ranking supports the way an average investor would think even for a typical non-insurance linked investment.

Finally, the last group of predictors in the importance ranking comprises from variables having strong technical weight in the securitization process and being insurance sector specific. The first predictor in this group, i.e. coverage type, refers to a contract term found in insurance contract whilst the second one, i.e. vendor, to the software company used to calculate the expected loss and various loss probabilities. Whilst this may not be immaterial information, there is not direct equivalent of such features in the financial markets. Thus, the average investor not specialising in insurance linked securities would not really dig deep into analysing vendor model updates, and historical loss catalogues, or even the wording of the transaction when thinking of returns prediction. Especially for vendor, it is a matter of fact that there is a global oligopoly in firms offering catastrophe risk modelling solutions in the insurance industry. Although the software developed by each of these companies is based on different assumptions, their scientific grounds are not disputed in the marketplace. This can be mostly attributed to the fact that these companies have been founded years before the birth of the first catastrophe bond and also that they have a long track record of being used in the traditional insurance and reinsurance markets. Thus, there is a contract of trust between them and the market participants as all vendors are perceived to be of equivalent reputational standing. Having said that, it does not mean that investors are sure about the reliability of the expected loss computation. It is just that most likely they would not believe that one vendor will have a much more valid estimate of loss compared to another. Similarly, coverage type really matters from an investor's perspective when seen in conjunction with the trigger or the combination of peril and geography. For example, catastrophe bonds with indemnity triggers or not well understood risks when combined with aggregate coverage terms can be risky in trapping investors' capital, as it was seen after 2018 Californian wildfires (Risk 2019). Taking into account all the above, the minimal depth predictors' importance ranking seem to reasonably reflect investors' current understanding of the market.

9. Example of random forest application in the industry

In this section, we present some possible examples of how the random forest could add value to ILS industry participants' daily operations. In particular, we discuss how the random forest could assist a would-be catastrophe bond issuer or investor in making faster and more informed decisions. In other words, we attempt to showcase examples of random forest applicability both from the "buy" and "sell" sides of the catastrophe bond market.

Starting from the sell side, a would-be catastrophe bond issuer along with their investment advisors, prior to finalising the terms of a new catastrophe bond issuance, would use the random forest to predict the likely spread at which investors would accept the offering. Getting to know this information is important as it allows for exploration of terms which would make the deal appear more attractive to an investor. In case this would not be feasible, the would-be issuer would realise faster that it may be preferable to explore alternative risk financing options.

From the buy side point of view, the random forest could also be beneficial to investors. In particular, just before a new catastrophe bond is issued, potential investors are provided with an offering circular. This document includes information about the deal which is to be launched and an invite for them to attend a road show, post which the issuance pricing will be settled. The information disclosed in this package refers to risk details, various design characteristics of the issuance and a price guidance. Investors want to make sure that the suggested spread compensates them enough for the true element of risk that they would undertake had they entered the transaction. However, a detailed analysis of this aspect can be time consuming as various departments and sometimes even external risk modelling firms get involved in the process. Whilst this process is undoubtedly important, investors would like to have a first flavour for a new deal's potential faster. Then, let's imagine how useful a straightforward prediction tool like random forest would be, where investors could plug in details provided in the circular of the new issuance the moment they receive it to get a quick spread prediction for the new transaction they investigate on the spot. This prediction would then be compared with the spread guidance offered and give investors an initial idea on whether the bond is overpriced, under-priced or "fairly" priced based on past catastrophe bond experience. This would direct investors to identify bargains faster and ask more relevant questions about the deal whilst on the road show. Then if the deal would be of interest, they could send all information needed to their modelling teams to perform the usual tasks of remodelling the underlying risk exposure and calculate the marginal impact that this new investment would bring into their portfolio. Overall, random forest is a solution that can speed up the investment decisions and help ILS investment firms not to use their valuable human resources for irrelevant catastrophe bond deals. As mentioned in Section 3.3.1, random forest could also be used to populate incomplete catastrophe bond deals databases when there is uncertainty or missingness of spread values for past transactions. We believe that its suitability for this purpose is very likely given the fact that we have some evidence about its high non-temporal prediction accuracy (see Section 7.1.1), and its "robustness" when information for more than one predictor is missing simultaneously (see Section 8.5) - a relatively usual phenomenon in an opaque market setting.

Besides, one note that needs to be made is that when assessing the discrepancy between the predicted spread value provided by the random forest (which for the buy side is the price guidance, and for the sell side it is the price for which the issuer would think that investors would accept the deal), one might first want to look back at what happened in the past, i.e. the historical discrepancy between the predicted and actual values recorded in the prediction phase post the random forest training. This may shed some light on the level at which a mispriced deal according to random forest is due to the portion of variability that the random forest could not explain or merely due to the fact that the new catastrophe bond has characteristics that have never been recorded in the past. The latter problem, could be mitigated if the random forest would be re-trained at frequent intervals, as part of the model validations taking place at least annually in a business context, enriching the training data set with more deals.

Finally, although many other parameters could be taken into account for random forest to be incorporated into internal business processes, here we give an idea of how the prediction power of random forest can liaise with issuers and investors' personal judgement to make faster and more informed decisions. It should be highlighted that recent developments in the catastrophe risk market also support the use of machine learning techniques. Prime examples are the new cyber risk model of AIR vendor, see AIR (2018), and a new platform for analysing deals and facilitating transparency in the catastrophe bond market, see Jones (2019).

10. Concluding remarks and future research

Until recently, the data-driven catastrophe bond pricing literature was mainly focused on building statistical models with an aim to test causal theory. The centre of interest lied on identification of variables which have a theoretically material and statistically significant link to catastrophe bond price, i.e. hypotheses of relationship between price and each independent variable were made. Then a statistical model, mostly linear regression, was applied to observed data to compute the size of this effect and the statistical significance of each independent variable in relation to the causal hypotheses set at the beginning. For model evaluation, in sample R^2 has been the classical way to assess model success, even though few more recent studies, such as Galeotti et al. (2013), Gürtler et al. (2016), and Braun (2016) have also considered out of sample model performance, and in some cases robustness checks for stability over different time periods. Model selection happened on the basis of keeping statistically significant factors and sometimes those non-significant ones having large coefficients to match the function connecting catastrophe bond spread and factors to the true underlying catastrophe bond data generation process.

The approach presented in the current research study was fundamentally different. A machine learning method called random forest was applied to a rich primary market catastrophe bond data set with a goal to predict catastrophe bond spreads at issuance given information in the offering circular and knowledge about current market conditions available at the time of prediction. Here, we did not focus on the underlying data generation process instead we learned the association between catastrophe bond spreads and predictors from the data directly using the random forest. The performance of our method was assessed on how accurately it predicted spreads based on unseen catastrophe bond observations on both temporal and non-temporal bases as well as the sensitivity of this prediction accuracy when possibly interacting predictors are missing. Variable importance measures referred to predictive ability and not the power to explain how the spreads are generated in this universe. There was also interest in securing repeatable prediction accuracy and predictors' importance results because of the multiple levels of randomness incorporated in random forests thus relevant checks were performed. The degree of divergence between predictive and explanatory importance was also of interest.

It was found that random forest has at least as good prediction performance as linear regression in the temporal context, and better prediction performance in the non-temporal one. Random forest performed better than linear regression when multiple predictors were missing from the model, as it has the ability to capture and extract interactions between existing variables. By assessing variables' importance on a non-explanatory basis, we found that all examined predictors have a say in the prediction of spread even if this is in varying degrees. The prediction accuracy, and predictors' importance results of random forest were stable. Taking prior explanatory literature and LR model into account, it appeared that predictive and explanatory power coexist for all catastrophe bond spread drivers considered in our study apart from size and vendor. There is potential for random forest to be used in the catastrophe bond industry to fast track investment decisions from both the buying and selling sides.

Based on the above findings there are certain aspects that would be interesting to research in the future. Although by using random forest as presented here, an investor, for instance, can see whether a new issuance of any type has a competitive price guidance or not, they do not get informed about the suitability of a new deal given their current portfolio composition. Addressing this need is a significant and important topic for future research. Another subject for future study is to extend our data set prior to 2009 to focus on the years of the financial crisis, and also additionally examine whether the drivers of private placements differ compared to those of non-private catastrophe bond deals. Finally, for the explanatory framework, another direction is for the variables size and BB spread to be further investigated as they stand out due to their potential interactions with other variables when other important variables are missing in the context of random forest.

In conclusion, our research provides some evidence that utilising both predictive and explanatory modelling can enhance the understanding of catastrophe bond market segment, increase its transparency and contribute to its development.

Declaration of competing interest

The authors declare that they have no conflict of interest.

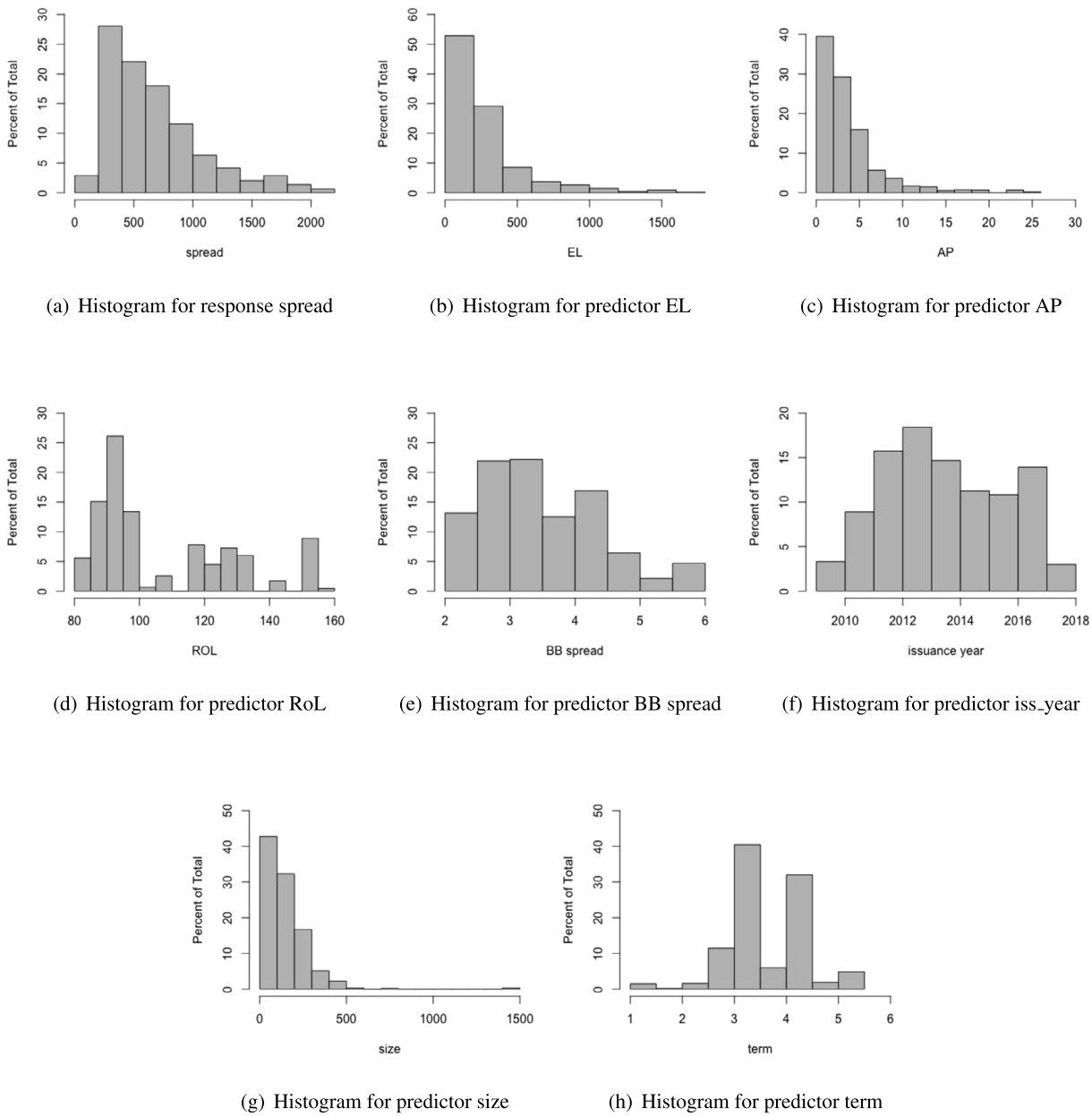
Acknowledgements

We thank the anonymous reviewers for their careful reading, whose insightful comments and suggestions helped significantly improve the content and presentation of this manuscript.

Appendix A. Summary statistics for the catastrophe bond data set

We now provide further information about the catastrophe bond data set used in this research paper. Summary statistics are presented for all variables, both continuous and categorical ones. Starting from the continuous variables, we present histograms in Fig. 7 and measures of central tendency and spread of the observations in our data set in Table 12. In Fig. 7, we see that all continuous variables have a right skewed distribution except variables RoL, BB spread, iss_year, and term. In particular, we see that the majority of catastrophe bond issuances in our data set corresponded to a RoL value of less than 100 indicating a soft market. Moreover, most of catastrophe bonds were issued in the year 2012-2013. It appears that term distribution has two peaks reflecting that most catastrophe bond issuances have a 3 to 5 year time horizon. Looking at Table 12, we notice that the range between minimum and maximum values for all continuous variables as well as the interquartile range are rather broad indicating that data points are well spread out. Such a data structure is anticipated in a catastrophe bond market setting. In essence, each issuance is a bespoke product developed to meet a very specific risk transfer need and consequently the population of catastrophe bond deals is heterogeneous.

Moving forward to categorical variables in Table 13, we present for each of them the number of level and number of observations under each level, with the latter quantity also being expressed as a percentage of the total number of observations. All variables levels are those used by the industry unless otherwise stated. Some comments regarding each categorical variable follow. With regards to coverage type, we find that the majority of catastrophe bonds during the studying period were issued to provide compensation in situations where a single large-scale loss event would activate the



(a) Histogram for response spread (b) Histogram for predictor EL (c) Histogram for predictor AP

(d) Histogram for predictor RoL (e) Histogram for predictor BB spread (f) Histogram for predictor iss_year

(g) Histogram for predictor size (h) Histogram for predictor term

Fig. 7. Histograms for the continuous variables. Percentage of total observations versus different ranges of a given numerical variable.

Table 12
Continuous variables summary statistics. The unit in which each continuous variable is measured is provided in brackets.

Continuous variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
spread (in basis points)	50.00	375.00	590.0	687.70	871.50	2200.00
EL (in basis points)	1.00	111.00	188.50	274.60	333.80	1735.00
AP (%)	0.02	1.36	2.51	3.72	4.68	25.04
RoL (in basis points)	83.57	91.73	96.06	106.97	124.57	159.73
BB spread (%)	2.12	2.61	3.41	3.46	4.14	5.98
iss_year (as numeric value)	2009.00	2012.00	2014.00	2014.00	2016.00	2018.00
size (in million US dollars)	3.00	75.00	130.00	164.70	200.00	1500.00
term (in years)	1.00	3.02	3.18	3.49	4.02	5.12

trigger, i.e. per occurrence coverage, as opposed to this happening due to a collection of insured loss events i.e. aggregate coverage. In very few instances in the data set, such as tranches A and B of Riverfront Re Ltd Series 2017-1 for example, per occurrence and annual aggregate coverage co-existed.

With respect to loc_peril, we shall start by providing some explanations in terms of abbreviations. The first part in each loc_peril level name indicates (a) geographical region(s). In particular, APAC stands for perils specific to Asia Pacific region, NA for perils relevant to North America, SA for prominent perils in South America,

Table 13
Summary statistics for all the categorical variables. Levels of each categorical variable are presented by number of observations and percentage of total observations. Abbreviations are explained in the text.

Categorical variable	Levels	No. of observations	Percentage (%)
coverage	aggregate	303	32.4
	occurrence	627	67.1
	both	4	0.5
loc_peril	APAC_Quake	51	5.46
	APAC_Typh	22	2.36
	Europe_APAC_Multi_Peril	2	0.21
	Europe_Quake	12	1.28
	Europe_Wind	54	5.78
	NA_APAC_Multi_Peril	26	2.78
	NA_Europe_APAC_Multi_Peril	36	3.85
	NA_Europe_Multi_Peril	39	4.18
	NA_Multi_Peril	425	45.50
	NA_Quake	80	8.57
	NA_Wind	184	19.70
	SA_Quake	3	0.32
rating	B	141	15.09
	BB	286	30.62
	BBB	4	0.43
	CCC	4	0.43
	nr (**n**ot **r**ated)	499	53.43
trigger	indemnity	511	54.7
	parametric	29	3.1
	industry loss index	325	34.8
	parametric index	23	2.5
	model	22	2.4
	multiple	24	2.6
vendor	AIR	741	79.3
	AON	4	0.4
	EQECAT	42	4.5
	RMS	141	15.1
	pp	6	0.6

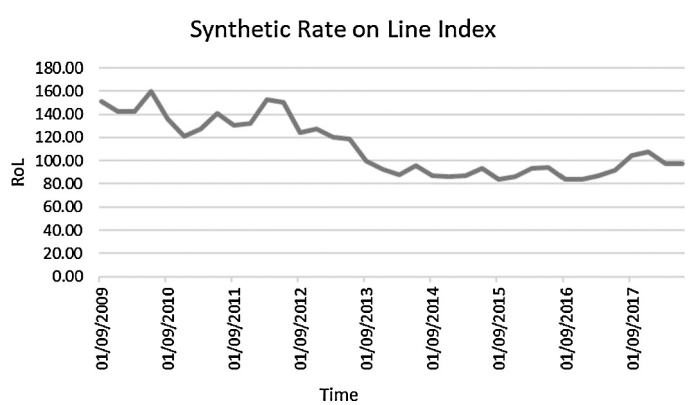


Fig. 8. Historical development of the Lane Financial LLC Synthetic Rate on Line Index (measured in percentage terms). Values above 100 indicate a hard market.

and Europe for perils in the aforementioned region. What follows the geographical region code, for instance APAC, is the peril type covered in the aforementioned location. There, except for those that are self-explanatory, Typh stands for typhoon and Multi_Peril includes various individual perils in the earlier indicated regions. For instance, one out of the NA_Europe_APAC_Multi_Peril tagged transaction provide cover against US named storms, Canadian earthquake, European earthquake, Australian wind and Australian earthquake. We see that almost half of the catastrophe bond deals in the data set had a mixture of perils in NA geographical territory which is quite expected since the perils in the area are generally considered to be more well understood and there is a longer heritage of issuances there. For example, bonds covering wind in North America are very popular even if the assumption of losses in the area is more likely due to the effect of hurricane seasons. Nevertheless, the high frequency of events had allowed risk modelling companies to understand the risk better, and build more trustworthy models with investors feeling more secure to buy exposures in this region. Looking into the credit quality allocation of the bonds issued, it is evident that more than 99% of catastrophe bonds in the data set either were characterised as non-investment grade securities or they did not receive a rating by any independent credit quality agency - the latter point has already been discussed more thoroughly in Section 4.3.

With regards to triggers, indemnity ones were the most popular among the bonds included in the study followed by industry

indices. This clearly shows a preference from cedents' perspective to get compensated for the exact level of losses that they anticipate to experience or at least to be compensated in line to industry losses. Deals which are triggered when pre-determined event parameters are satisfied or surpassed accounted only for 5.6% of the total market in the period under study. Examples of parametric index deals in the current data set is Atlas VI Capital Ltd. Series 2010-1 and Bosphorus Ltd. Series 2015-1 whilst IBRD CAR 118-119 is an example of pure parametric trigger deal issued by the International Bank for Reconstruction and Development for Mexico's natural disaster fund named FONDEN. The least used triggers were those combining different trigger types such as Fortius Re II Ltd. Series 2017-1 and those based on the modelled losses of the cedent's exposure portfolio calculated based on event parameters gathered from specified agencies, such as Akibare II Ltd. single tranche.

With respect to the risk modelling company used to calculate the expected loss of investors' exposure to underlying peril, we see that AIR Worldwide is the most widely used followed by RMS. Together, they account for the 94.4% of all non-life securitisations in the data sample followed by EQECAT, AON and pp accounting for the rest 5.6%. It is worth to note that pp abbreviation is not a risk modelling firm but it stands for **p**rivate **p**lacement. Examples are the single tranches of Merna Re Ltd. Series 2016-1, 2017-1, 2018-1 which were privately purchased by specialized ILS funds. Finally, the internal model of AON was used for very few deals where the aforementioned company had acted as the structuring and placement agent, such as in the case of Windmill I Re series 2013-1.

Appendix B. In sample and out of sample performance of LR model using the variable Investment Grade (IG) instead of the variable (granular) rating

See Tables 14 and 15.

Table 15
Out of sample performance measured in terms of R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} for the improved linear model of Braun (2016) versus the linear regression model with **I**nvestment **G**rade (IG) variable to indicate credit quality (LR with IG), and the benchmark linear regression model (LR) which includes the variable rating presented in Table 1 and Table 13 in Appendix A.

Model	R^2_{OOB}	R^2_{10CV}	R^2_{LOOCV}
Improved Braun (2016)	79.71%	80.81%	79.40%
LR with IG	82.22%	82.35%	82.72%
LR (with granular rating)	83.30%	84.42%	83.84%

Table 14
In sample fit of the linear regression model with **I**nvestment **G**rade (IG) variable to indicate credit quality (LR with IG) as opposed to variable rating presented in Table 1 and Table 13 in Appendix A.

	Estimate	Std. error	t value	Pr(> |t|)
(Intercept)	61540	9677	6.36	0.000 ***
RoL	6.05	0.43	14.15	0.000 ***
BB spread	50.34	8.66	5.81	0.000 ***
IG 0 (*baseline*)				
IG 1	−253	85.34	−2.96	0.003 **
IG nr (**n**ot **r**ated)	87.99	15.71	5.6	0.000 ***
term	−27.25	8.86	−3.07	0.002 **
size	0.00	0.00	3.06	0.002 **
trigger industry loss index (*baseline*)				
trigger indemnity	−0.58	14.77	−0.04	0.969
trigger model	−92.47	39.91	−2.32	0.021 *
trigger multiple	−44.9	40.13	−1.12	0.264
trigger parametric index	−23.9	42.02	−0.57	0.570
trigger parametric	−122.6	37.4	−3.28	0.001 **
coverage aggregate (*baseline*)				
coverage both	55.80	85.00	0.66	0.512
coverage occurrence	−59.95	13.95	−4.3	0.000 ***
vendor AIR (*baseline*)				
vendor AON	101.1	90.59	1.11	0.265
vendor EQECAT	−0.68	33.96	−0.02	0.98
vendor pp	21.33	73.65	0.29	0.772
vendor RMS	15.4	19.25	0.8	0.424
AP	−16.32	6.09	−2.68	0.008 **
EL	1.33	0.09	15.22	0.000 ***
iss_year	−30.79	4.79	−6.42	0.000 ***
APAC_Quake (*baseline*)				
loc_peril APAC_Typh	−77.74	44.6	−1.74	0.082
loc_peril Europe_APAC_Multi_Peril	−9.11	129.2	−0.07	0.944
loc_peril Europe_Quake	−13.43	57.83	−0.23	0.816
loc_peril Europe_Wind	−138.1	40.7	−3.4	0.001 ***
loc_peril NA_APAC_Multi_Peril	100.1	47.48	2.1	0.035 *
loc_peril NA_Europe_APAC_Multi_Peril	152.7	42.51	3.6	0.000 ***
loc_peril NA_Europe_Multi_Peril	149.9	40.96	3.66	0.000 ***
loc_peril NA_Multi_Peril	166.8	28.25	5.9	0.000 ***
loc_peril NA_Quake	−19.79	35.59	−0.55	0.57
loc_peril NA_Wind	97.83	30.72	3.18	0.002 **
loc_peril SA_Quake	133.1	107.4	1.24	0.216
R^2	83.96%			
Adjusted R^2	83.41%			
Res. Std. Error	166.8 (df = 902)			
F Statistic	152.3 (df = 31; 902)			
Note for signif. codes:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$			
Observations number:	934			

Appendix C. In sample and out of sample performance of Braun (2016) model using a subset of our catastrophe bond data

See Tables 16 and 17.

Table 16
In sample fit of candidate benchmark models specification. Here, the linear regression model of Braun (2016) was applied on our catastrophe bond data set. We notice that only 434 data points are considered as the binary Investment grade variable does not take into account non-rated transactions.

	Estimate	Std. error	t value	Pr(> |t|)
(Intercept)	−665.97	40.33	−16.51	0.000***
Swiss Re	−20.12	14.57	−1.38	0.168
RoL index	5.24	0.44	12.02	0.000***
BB spread	57.18	11.66	4.91	0.000***
Investment grade	−39.17	73.32	−0.53	0.593
Peak territory	224.95	19.22	11.70	0.000***
Expected Loss	1.64	0.07	23.85	0.000***
R^2	79.97%			
Adjusted R^2	79.69%			
Res. Std. Error	143 (df = 428)			
F Statistic	284.8 (df = 6; 428)			
Note for signif. codes:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$			
Observations number:	434			

Table 17
Out of sample performance measured in terms of R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} for the linear model of Braun (2016) versus the linear regression (LR) in this study.

Model	R^2_{OOB}	R^2_{10CV}	R^2_{LOOCV}
Braun (2016)	79.20%	80.40%	79.12%
LR	83.30%	84.43%	83.84%

Appendix D. Prediction accuracy performance of RF with categorical dummy variables

See Table 18.

Table 18
Prediction accuracy performance measured in terms of R^2_{OOB}, R^2_{10CV}, and R^2_{LOOCV} for random forest (RF) when converting all the categorical variables into dummies in the catastrophe bond data set.

Model	R^2_{OOB}	R^2_{10CV}	R^2_{LOOCV}	R^2
RF	96.57%	96.49%	96.59%	99.25%
RF_dummies	96.48%	96.16%	96.63%	99.18%

References

AIR, 2018. AIR develops advanced probabilistic model for global cyber risks. https://www.air-worldwide.com/Press-Releases/AIR-Develops-Advanced-Probabilistic-Model-for-Global-Cyber-Risks/. (Accessed 19 May 2019).

ARTEMIS, 2019. Decline in ILS ratings shows the asset class isn't so alternative: Kbra. http://www.artemis.bm/news/decline-in-ils-ratings-shows-the-asset-class-isnt-so-alternative-kbra/. (Accessed 19 March 2019).

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36 (1–2), 105–139.

Berk, R.A., 2008. Statistical Learning from a Regression Perspective, 2 edn. Springer.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25 (2), 197–227.

Bodoff, N.M., Gan, Y., 2009. An analysis of the market price of cat bonds. In: Casualty Actuarial Society E-Forum, Spring 2009.

Braun, A., 2012. Determinants of the cat bond spread at issuance. Zeitschrift für die gesamte Versicherungswissenschaft 101 (5), 721–736.

Braun, A., 2016. Pricing in the primary market for cat bonds: new empirical evidence. The Journal of Risk and Insurance 83 (4), 811–847.

Breiman, L., 1996a. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 1996b. Heuristics of instability and stabilization in model selection. The Annals of Statistics 24 (6), 2350–2383.

Breiman, L., 1996c. Stacked regressions. Machine Learning 24 (1), 49–64.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software.

Calle, M.L., Urrea, V., 2010. Letter to the editor: stability of random forest importance measures. Briefings in Bioinformatics 12 (1), 86–89.

Chen, X., Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics 99 (6), 323–329.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5 (4), 559–583.

Díaz-Uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7 (1), 3.

Dieckmann, S., 2010. By force of nature: explaining the yield spread on catastrophe bonds. http://dx.doi.org/10.2139/ssrn.1082879.

Dietterich, T.G., 2000a. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.

Dietterich, T.G., 2000b. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning 40 (2), 139–157.

Efron, B., 1992. Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics. Springer, pp. 569–593.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA.

Galeotti, M., Gürtler, M., Winkelvos, C., 2013. Accuracy of premium calculation models for cat bonds—an empirical analysis. The Journal of Risk and Insurance 80 (2), 401–421.

Gatumel, M., Guegan, D., 2008. Towards an Understanding Approach of the Insurance Linked Securities Market, Documents de travail du centre d'economie de la sorbonne. Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne.

Götze, T., Gürtler, M., 2018. Sponsor-and trigger-specific determinants of cat bond premia: a summary. Zeitschrift für die gesamte Versicherungswissenschaft, 1–16.

Götze, T., Gürtler, M., Witowski, E., 2020. Improving cat bond pricing models via machine learning. Journal of Asset Management 21 (5), 428–446.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. American Statistician 63 (4), 308–319.

Gürtler, M., Hibbeln, M., Winkelvos, C., 2016. The impact of the financial crisis and natural catastrophes on cat bonds. The Journal of Risk and Insurance 83 (3), 579–612.

Hills, S., 2009. Catastrophe bond market emerges from crisis. https://www.reuters.com/article/us-catbonds-revival-analysis-idUSTRE59F1MK20091016. (Accessed 10 July 2020).

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Electronic Journal of Statistics 1, 519–537.

Ishwaran, H., Kogalur, U., 2019. Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.8.0.

Ishwaran, H., Kogalur, U.B., Chen, X., Minn, A.J., 2011. Random survival forests for high-dimensional data. Statistical Analysis and Data Mining: The ASA Data Science Journal 4 (1), 115–132.

Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S., 2010. High-dimensional variable selection for survival data. Journal of the American Statistical Association 105 (489), 205–217.

Jaeger, L., Müller, S., Scherling, S., 2010. Insurance-linked securities: what drives their returns? The Journal of Alternative Investments 13 (2), 9–34.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.

Jones, L., 2019. Shah: data-driven cat bond can be replicated. Trading Risk 119. 9–9.

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28 (5), 1–26.

Lane, M., Mahul, O., 2008. Catastrophe Risk Pricing: An Empirical Analysis. The World Bank.

Lane, M.N., 2000. Pricing risk transfer transactions 1. STIN Bulletin: The Journal of the IAA 30 (2), 259–293.

Lange, T., Roth, V., Braun, M.L., Buhmann, J.M., 2004. Stability-based validation of clustering solutions. Neural Computation 16 (6), 1299–1323.

Lei, D.T., Wang, J.-H., Tzeng, L.Y., 2008. Explaining the spread premiums on catastrophe bonds. In: NTU International Conference on Finance. Taiwan.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Lim, C., Yu, B., 2016. Estimation stability with cross-validation (escv). Journal of Computational and Graphical Statistics 25 (2), 464–492.

Major, J.A., 2019. Methodological considerations in the statistical modeling of catastrophe bond prices. Risk Management and Insurance Review 22 (1), 39–56.

Muir-Wood, R., 2017. The case of the trapped collateral. https://www.rms.com/blog/2017/11/16/the-case-of-the-trapped-collateral/. (Accessed 19 March 2019).

Ntoutsi, I., Kalousis, A., Theodoridis, Y., 2008. A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees. In: Proceedings of the 2008 SIAM International Conference on Data Mining. SIAM, pp. 810–821.

Oh, J., Laubach, M., Luczak, A., 2003. Estimating neuronal variable importance with random forest. In: 2003 IEEE 29th Annual Proceedings of Bioengineering Conference. IEEE, pp. 33–34.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research 11, 169–198.
Papachristou, D., 2011. Statistical analysis of the spreads of catastrophe bonds at the time of issue. ASTIN Bulletin: The Journal of the IAA 41 (1), 251–277.
Perrone, M.P., 1993. Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization. PhD thesis. Physics Department, Brown University, Providence, RI.
Philipp, M., Rusch, T., Hornik, K., Strobl, C., 2018. Measuring the stability of results from supervised statistical learning. Journal of Computational and Graphical Statistics 27 (4), 685–700.
Probst, P., Bischl, B., Boulesteix, A.-L., 2018. Tunability: importance of hyperparameters of machine learning algorithms. arXiv preprint. arXiv:1802.09596.
Risk, T., 2019. Controlling a blazing risk. Trading Risk ILS Investor Guide, 14–15.
Russell, S.J., Norvig, P., 2016. Artificial Intelligence: A Modern Approach. Pearson Education, Limited.
Shmueli, G., 2010. To explain or to predict? Statistical Science 25 (3), 289–310.
Stodden, V., 2015. Reproducing statistical results. Annual Review of Statistics and Its Application 2, 1–19.
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8 (1), 25.
Trottier, D.-A., Charest, A.-S., et al., 2018. Cat bond spreads via hara utility and nonparametric tests. The Journal of Fixed Income 28 (1), 75–99.
Turney, P., 1995. Bias and the quantification of stability. Machine Learning 20 (1–2), 23–33.
Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S., Ahrentzen, S., 2018. Random forest based hourly building energy prediction. Energy and Buildings 171, 11–25.
Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5 (2), 241–259.
Yu, B., 2013. Stability. Bernoulli 19 (4), 1484–1500.
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC.