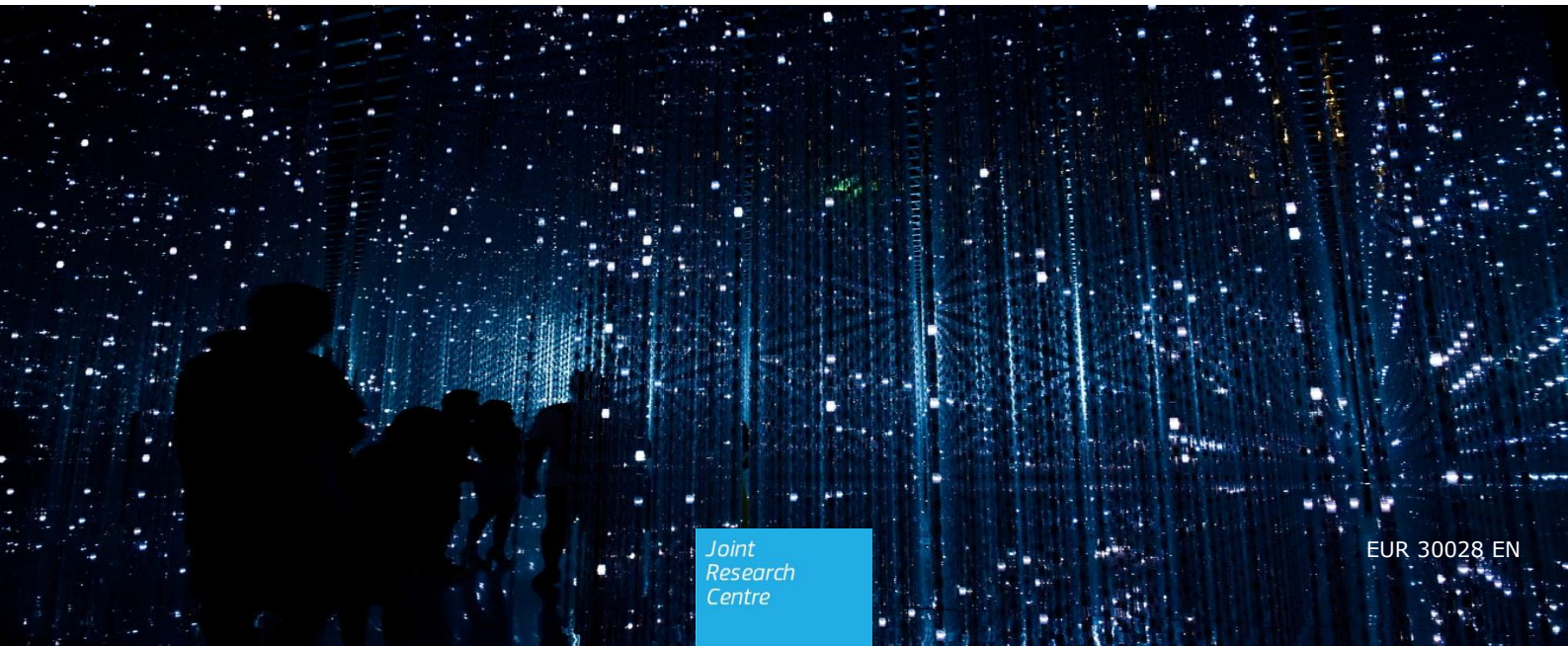
European Commission

JRC TECHNICAL REPORTS

The Global Conflict Risk Index: Artificial Intelligence for Conflict Prevention

HALKIA Matina, FERRI Stefano, DEEPEN Yannick,
PAPAZOGLOU Michail, VAN DAMME Marie-Sophie,
BAUMANN Kathrin Manuela

2019



EUR 30028 EN

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Contact information
Name: Matina Halkia
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 480, 21027 Ispra (VA), Italy
Email: matina.halkia@ec.europa.eu
Tel.: +39 0332786242

JRC Science Hub
https://ec.europa.eu/jrc

JRC118746

EUR 30028 EN

PDF ISBN 978-92-76-14306-2 ISSN 1831-9424 doi:10.2760/004232

Luxembourg: Publications Office of the European Union, 2019

How to cite this report: Halkia, S., Ferri, S., Deepen, Y., Papazoglou, M., Van Damme, M. and Baumann, K., The Global Conflict Risk Index: Artificial intelligence for conflict prevention, EUR 30028 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-14306-2 (online), doi:10.2760/004232 (online), JRC118746.

Contents

Acknowledgments

JRC would like to thank Marie SCHELLENS, who during her JRC traineeship initiated the work on random forest conflict risk modelling. Additionally, Marie-Laure CHARPIGNON, a JRC visiting scientist, contributed further to the implementation of the initial GCRI AI code. The GCRI AI models presented in this report relies on the modelling effort of Yannick DEEPEN.

JRC would also like to thank FPI.2, Marc FIEDRICH and Giovanni SQUADRITO, as well as EEAS/CSDPCR.PRISM, Rene VAN NES, Gosia SENDROWSKA and Pavla DANISOVA for their unwavering support to the development of the Global Conflict Risk Index. The GCRI project has been financed by the Instrument contributing for Stability and Peace (IcSP).

Authors

HALKIA Matina[1], FERRI Stefano[1], DEEPEN Yannick[1], PAPAZOGLOU Michail[2], VAN DAMME Marie-Sophie[3], BAUMANN Kathrin Manuela[1]

([1]) European Commission, Joint Research Centre (JRC), Ispra, Italy.

([2]) Unisystems S.A, Via Michelangelo. Buonarroti 39, 20145 Milano, Italy.

([3]) Piksel Ltd Italian Branch, Via Breda 176, Milano (MI), Italy.

Abstract

The Global Conflict Risk Index (GCRI), which was designed by the European Commission's Joint Research Centre (JRC), is the quantitative starting point of the EU's conflict Early Warning System. Taking into consideration the needs of policy-makers to prioritize actions towards conflict prevention, the GCRI expresses the statistical risk of violent conflict in a given country in the upcoming one to four years. It is based on open source data and grounded in the assumption that the occurrence of conflict is linked to structural conditions, which are used to compute the probability and intensity of conflicts.

While the initial GCRI model was estimated by means of linear and logistic regression models, this report presents a new GCRI model based on the Artificial Intelligence (AI) random forest (RF) approach. The models' hyperparameters are optimized using a ten-fold cross validation.

Overall, it is demonstrated that the random forest GCRI models are internally stable, not overfitting, and have a good predictive power. The precision and accuracy metrics are above 98%, both for the national power and subnational power conflict models.

The AI GCRI, as a supplementary modelling method for the European conflict prevention policy agenda, is scientifically robust as a baseline quantitative evaluation of armed conflict risk additional to the linear and logistic regression GCRI.

1 Introduction

During the last decade, the bulk of scholarly debate in conflict and peace research has increasingly shifted from the view that existing quantitative modelling techniques of conflict risk are not sufficient enough to forecast conflicts (Weidmann & Ward, 2010) to the view that prediction is feasible and policy-relevant within a 'limited spatial and temporal scope' (Cederman & Weidmann, 2017).

The dominant approach to forecasting the risk of a conflict is based on linear or logistic regression models, which are used to estimate the intensity or the probability of a violent conflict event (De Groeve, Hachemer & Vernaccini, 2014; Halkia et al., 2017, 2019).

So far, the Global Conflict Risk Index (GCRI) is the only model that calculates both the probability and intensity of the structural risk of a country experiencing armed conflict at the national or subnational level.

It uses open-source data from 1989 to 2018 for 191 countries worldwide as country-year observations, and it considers solely the structural conditions characterising a country. For this purpose, it takes 25 individual variables in six risk areas into account, i.e. social, economic, security, political, demographic, and geographical/environmental. All variables used in the models have extensively been used as explanatory or control variables in the literature on peace and conflict. The datasets used are all freely accessible on the internet and have been compiled by diverse international organizations, such as the World Bank and the United Nations, or academic institutions (Halkia et al., 2017, 2019).

With the latest improvements in the fields of artificial intelligence (AI) and machine learning (ML), the GCRI has been extended using AI approaches such as the random forest method. RF can be used for classification and regression tasks similar to the already implemented techniques in the GCRI. However, given that the random forest model is composed of several decision trees, it allows for more flexibility in the model, which may capture otherwise undetected relationships between the variables of interest, and it delivers predictions with a higher precision than the previously applied methods. Furthermore, the random forest can estimate the variables' importance to rank which has the highest predictive power.

In this report, we present and validate the AI version of the GCRI (AI GCRI) based on the random forest approach. First, we give a brief description of the random forest approach and discuss its added value though the following sections. Next, we introduce the application of the random forest approach in the AI GCRI, the specification and the validation of the model. Finally, we present and discuss the results of the random forest

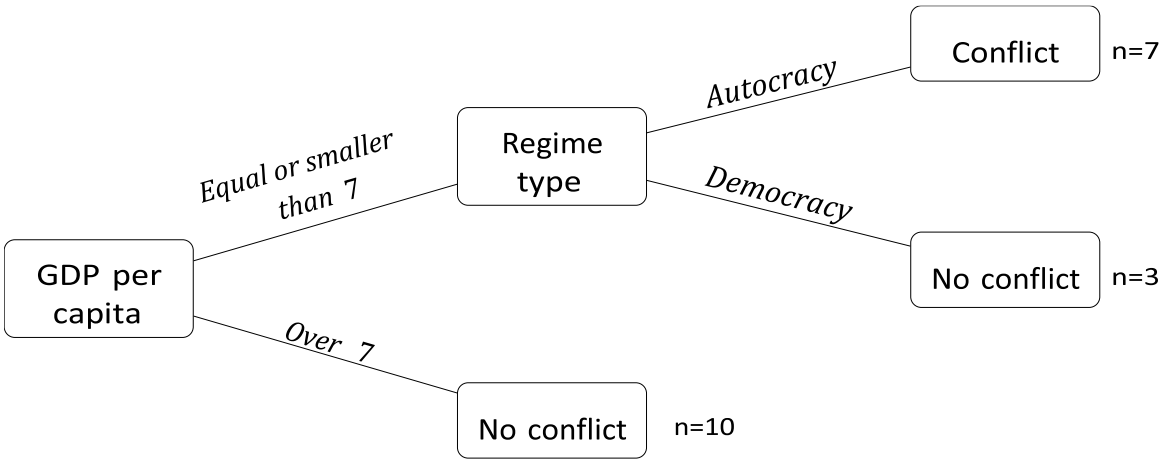
model, while we also test and present the results after using a potential new variable, which was introduced in the GCRI indicator basket recently, the internally displaced people (IDPs) (Halkia et al., 2018). Possible steps for further development of AI in the GCRI are also discussed.

2 Random Forest models

2.1 Introduction to random forest models

The random forest (RF) model is an ensemble method that is based on the use of decision trees. A decision tree can be best understood as a tree diagram with several splits on different nodes. To illustrate the function of a decision tree, a highly simplified conflict prediction model is provided in Figure 1. Assume that we have 20 observations (n=20), each with a value for regime type and the GDP per capita (rescaled from 0 to 10). At the initial node, the decision tree splits the data into two different groups based on the GDP per capita. If the GDP per capita is higher than 7, the model will predict that there will be no conflict. If the GDP per capita is smaller or equal to 7, the model creates a new split on the decision node for the regime type variable. The model hereby does not take into account the previous decisions and treats the current decision node as the new initial node.

Figure 1. Simple decision tree for conflict prediction based on the GDP per capita and regime type



Source: JRC, 2019.

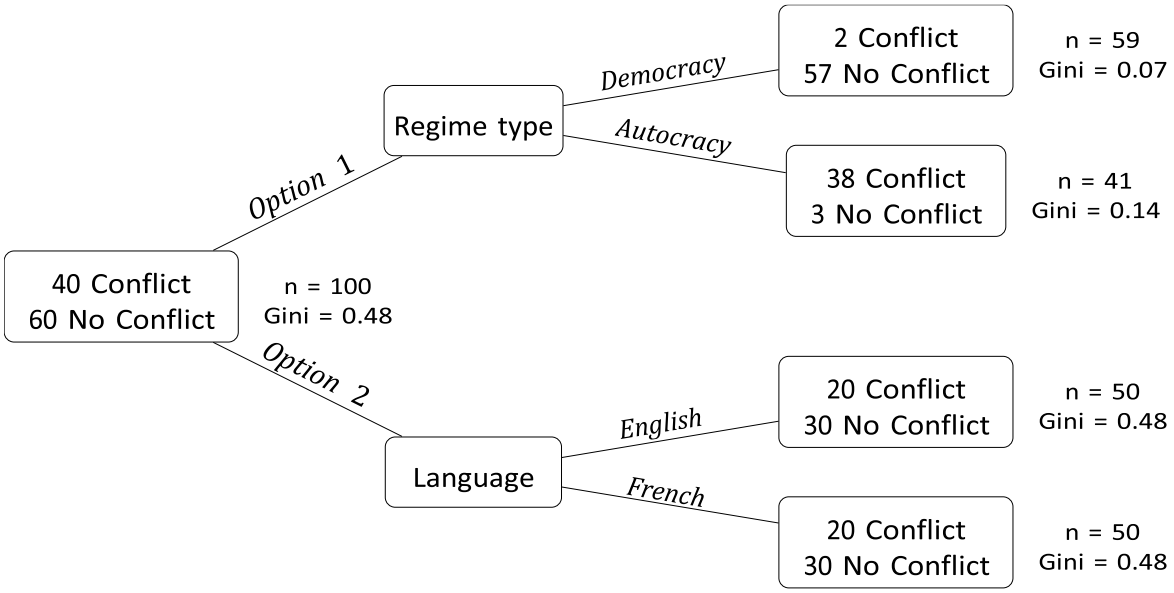
At each node, the tree splits the sample into two or more rather homogenous sets based on the variable that gives the best split. To do so, the tree calculates the possible splits for each variable and selects then the one which results in the most homogenous sub-nodes. The criteria to choose the best variable vary for different

algorithms. The most common used algorithm is the CART algorithm (Classification and Regression tree). The CART algorithm relies on the GINI index to decide on the best split. The Gini Index is defined as

$$Gini(B) = 1 - \sum_{i=1}^{1} p(b_i)^2$$

where B stands for the variable split in question, for instance regime type, and b_i for the probability of an observation to be classified as either 0 or 1 at the leaves following the variable split. To calculate the Gini index, we sum the weighted, squared probabilities and subtract the result from 1. Like this, the Gini Index can be used as a measure of purity for the split. To illustrate how this works, consider the following simplified, fictive example of a tree, where we try to split a conflict data set based on two different variables: the relevant variable regime type, and a rather unimportant variable such as the language spoken in the country. Further assume that we have a mix of 40 conflict observations and 60 observations without conflict at the initial node (see Figure 2). The first option would be to split up the sample based on the regime type to see if we can split the data even more homogenously. Based on the divide by the regime type, we have 2 conflict observations and 57 observations without conflict in democratic countries (Figure 2). If we insert this values in the formula, we get $1 - \left((\frac{2}{59})^2 + (\frac{57}{59})^2\right) \sim 0{,}07$. We do the same for the autocracies and get a value of ~0.14. To account for the different numbers of observations, we weight the values by the number of observations and add up the Gini values. In our example, this is $0.14 * 0.41 + 0.07 * 0.59 \sim 0.1$. Now, we compare this value with the Gini value of the original node with 40 times conflict and 60 times no conflict. For the original node, the Gini value is 0.48. Hence, the difference is 0.38, which is called the *information gain*. If we look at the second option, the languages, we can see that there is no difference in the Gini value, thus the information gain is 0.

Figure 2. Split of a decision tree for conflict prediction based on regime type and language



Source: JRC, 2019.

Based on this criteria, the tree would choose the regime type variable over the language variable. Without specifying any limitations, the tree will continue to split until there is no more information gain possible In other words, the tree will almost certainly overfit, thus it will create one final node, also called leaf for decision trees, per observation. For the validation, each observation of the test data goes through the tree and is classified as either conflict or not in our example.

2.2 How random forest works

Even though decision trees have a clear advantage in terms of construction and interpretation, there are several issues that come along when using decision trees. First, decision trees have been proven to be unstable learners since small changes in the data might change the entire tree (Breiman, 1996). Second, decision trees tend to overfit easily. This requires either a constraint in the parameters, such as a minimum number of observations per leaf, or to prune the tree once it is fully grown. Third, given the large availability of high dimensional data nowadays, a single tree will unlikely model all the information in the available data (Zhang & Wang, 2009). This is especially true in cases where there are only few observations (Zhang & Wang, 2009), such as in armed conflict.

One of the most commonly used techniques to stabilize unstable learners is the bootstrap aggregation, also known as bagging. Hereby, bootstrap samples[4] with a replacement are taken from the original learning sample. For each bootstrap sample, a tree is grown which then classifies a given observation. The final prediction is the majority class of the predicted classes by the trees (Breiman, 1996). The core idea here is to overcome the weakness of each decision tree by increasing the number of trees and relying on the wisdom of the mass instead a single decision. Therefore, this method belongs to the ensemble methods, which rely on the decisions of several different models to reach a final output.

The random forest model presents a special case of bootstrap aggregation. In the random forest, multiple decision trees are created, each based on a bootstrap sample with replacement from the original learning data. Different to the normal CART mechanism, which uses the full model to evaluate which variable is best to use for the split at a given node, the random forest relies on a random subspace selection method. For the random subspace selection, the random forest randomly selects a predefined number of independent variables to choose from for the split at each node. Each descendant node is hereby treated as a new initial node without knowledge of the prior selection. The criterion for the final decision of the split is the Gini criterion: the random forest will select the split with the highest information gain at each node.

In the end, each tree classifies an observation according to the different variables and splits used for each tree. Afterwards, the number of trees voting for a certain class is counted. The final predicted class for the observation in question is the class with the majority of votes among all trees.

More formally, the final vote is calculated based on the fraction of trees that voted for a certain class such that:

$$\hat{v}' = \frac{1}{B}\sum_{b=!}^{B} f_b(x')$$

where $\hat{v}'$ is the share of votes, B is the number of trees, and $f_b(x')$ returns as 1 if tree b predicts that x' belongs to a class (here conflict), and 0 otherwise (no conflict). Afterwards a cutoff (C) is chosen so that the predicted class (final vote) y' is defined as follows:

$$y' = \begin{Bmatrix} 1\ if\ \hat{v}' \geq C \\ 0\ if\ \hat{v}' < C \end{Bmatrix}$$

(4) The bootstrap sample is to the sample, as the sample is to the population.

The default cutoff is 0.5, thus majority voting. For example, if 51 out of 100 trees voted for 1 (conflict), the predicted class will be 1, since it is over 0.5.

In sum, the method can be described as follows:

— Sample for each of a defined number of trees a bootstrap sample with replacement from the original learning dataset.
    - At each initial node, randomly select a predefined number of variables from the data.
    - Find the best split for each of the variables and choose the one with the lowest impurity.
— Repeat the steps for all descendant nodes until the impurity can no longer be lowered.
— For the validation, each tree votes for a certain class. The final vote is the class that receive the majority of votes.

Given that we rely on different decision trees with a random variable selection, we have several models that may capture different (i.e. non-linear) relationships; this is not possible with a single model such as logistic regression, which relies on a linearity assumption. Second, this ML approach chooses randomly from a predefined number of variables for each split. Given that the CART mechanism uses the full model for the selection, it will tend to use those variables that have a strong impact on the dependent variable and leave out weaker variables. However, this increases the correlation between the trees, which in turn lowers the predictive power of the model.

To address this issue, the random selection of variables in the random forest ensures that different variables are chosen at each split, thereby lowering the correlation between the trees, thus increasing the explanatory power of the model. In the next chapter, we will describe the setup of the random forest in the GCRI in more detail.

3 Hyperparameter for random forest models

The random forest requires the specification of certain hyperparameters. Mainly, the random forest relies on a set of three parameters. First, the number of variables from which to randomly choose at each node, known as *mtry*. Second, the number of trees used, known as *ntree*. Third, the minimum number of observations in the final leafs, known as *nodesize*. The following subsections will present each parameter separately and illustrate how different settings of the parameters influence the output results.

3.1 The optimal number of variables (mtry)

The main goal of choosing *mtry* is to lower the correlation between the trees by limiting the number of possible variables from which to choose from. In other words, *mtry* sets the possible variable subsets which can be used in the model. It can ranges from 1 to 25 which is the total number of the variables we have as input. The limitation of *mtry* ensures more randomness in each tree, which lowers the correlation between the trees. Less correlation between the trees implies that the trees differ in their structure, allowing for the detection of patterns that might not have been discovered otherwise. Furthermore, the lowered correlation leads to an increased accuracy which in turn reduces the generalization error (Breiman, 1996). At the same time, each tree has less predictive power given the reduced covariates for each tree (Palmer et al., 2007). The reason is that if *mtry* is set too low, it is possible that none or few of the relevant predictors are chosen for a split, which leads to a lowered predictive power. However, if we choose a large value for *mtry*, the variables with smaller effect are unlikely to be selected at a given node (Janitza & Hornung, 2018). This contradicts the idea of a random forest with many different decision trees, since many would look alike. It is therefore challenging to find the optimal trade-off between accuracy and loss in predictive power.

Yet, the default setting of *mtry* for regressions is $\frac{p}{3}$ and for classification it is equal to $\sqrt{p}$, whereby *p* refers to the maximum number of independent (explanatory) variables. The full GCRI model consists of 25 variables. Accordingly, the default value for the classification is equal to $\sqrt{25} = 5$. For the regression, the default value is equal to $\frac{25}{3} \approx 8$.

3.2 The optimal number of trees (ntree)

The number of trees used in the random forest is important for the stabilization of the results. The core idea of the random forest is to overcome the weakness of each decision tree by increasing the number of trees and relying on the wisdom of the mass. The lower the number of trees, the greater the influence of each individual tree on the final prediction. Given that the decision trees are sensitive to changes in the data, the final predictions are rather unstable. Thus, the number of trees should be chosen to be sufficiently high to achieve stable results that are rather insensitive to small changes in the input data. It could be shown that results tend to converge once a certain threshold of trees is reached (Guan et al., 2013). Thus, it is recommendable to use a rather large number of trees since an increase in trees will stabilize, but not harm the results. The default value for the random forest is 500. Scholars proposed different thresholds over the past decade such as 100

(Pal, 2005; Guan et al., 2013), 1000 (Prasad, Iverson & Liaw, 2006; Sesnie et al., 2010; Rees et al., 2014; Colditz, 2015), 5000 (Díaz-Uriarte & Alvarez de Andrés, 2006; Stumpf & Kerle, 2011; Adelabu, Mutanga & Adam, 2014; Millard & Richardson, 2015; Nitze, Barrett & Cawkwell, 2015) and up to over 6500 (Adam & Mutanga, 2012).

However, there seems to be agreement that there are diminishing benefits once the number of trees is larger than the default value of 500. After 500 trees, the result will not change substantially, yet, the computational effort to calculate the results increases linearly to the number of trees chosen for the random forest (Scornet, 2017).

3.3 The optimal depth of trees (nodesize)

The *nodesize* defines the minimum number of observations that should be on the terminal leaf of a tree. The lower the *nodesize*, the larger the tree can grow, since the tree is likely to have more splits. The default value is 1, i.e. the minimum number of observations per final leaf is 1 (fully grown trees). In theory, this may introduce overfitting for the individual tree. However, if we set the number of trees high enough, the ensemble can compensate for the overfitting of individual trees. In other words, a smaller *nodesize* should be chosen alongside a larger number of trees. The lower the *nodesize*, the better it can detect more complicated relationships between the independent variables and the dependent variable. In practice, the default value for *nodesize* is rarely changed (De Santana et al., 2019).

4 The Artificial Intelligence version of the Global Conflict Risk Index

For the optimization of our model, we performed a ten-fold cross validation for each possible parameter combination. Given the way we predict conflict, we optimized the random forest (RF) regression model taking into consideration the GCRI steering committee opinion[5]. For the *ntree* parameter, we set thresholds of 100, 500, 1000, and 2000. For *mtry*, we let the value vary between 1 and 25. We set the *nodesize* to the default value of 5 as well as 1 and 20. We optimized the model by means of the RSME.

To account for the high imbalance within our data set, SMOTE oversampling (Chawla et al., 2002) was performed within the cross-validation procedure (CV), as an over-sampling before the CV would lead to biased results (Santos et al., 2018). We first split the data into 10 stratified partitions. For each iteration of the CV, we

([5]) see 5th GCRI workshop proceedings JRC 117492

oversample only the training data, leaving the test data unmodified. This way, we are able to generalize from our training data.

5 Results

The results for the RF regression analysis on the national level (NP) revealed that the best performance was achieved under the setting *mtry*=2, *ntree*=2000, *nodesize*=1 (see Table 1) with an error rate of 1.148. As a reference, the default model with *ntree*=500, *mtry*=8 and *nodesize*=5 has an error rate of 1.219.

Table 1. Models' optimal regression hyperparameters.

	mtry	**ntree**	**nodesize**	**RSME**
NP	2	2000	1	1.148
SN	3	1000	1	1.252

For the subnational conflict level (SN), the setting of *ntree*=1000, *mtry*=3 and *nodesize*=1 gave the smallest error with 1.252 (see Table 1). As a reference, the default model with *ntree*=500, *mtry*=8 and *nodesize*=5 has an error rate of 1.299.

Table 2. Models' statistical metrics using the optimal parameters.

	NP	**Default NP**	**SN**	**Default SN**
mtry	2	5	11	5
ntree	1000	500	500	500
nodesize	5	1	20	1
Accuracy	0.94406	0.94365	0.91608	0.91196
Kappa	0.74224	0.73772	0.72959	0.71344
Sensitivity	0.98992	0.9916	0.98887	0.98993
Specificity	0.66908	0.65619	0.66361	0.64158
Neg. Pred.	0.91789	0.92938	0.94596	0.94926
Precision	0.94722	0.94535	0.91075	0.90553
Recall	0.98992	0.9916	0.98887	0.98993
F1	0.96809	0.96791	0.94817	0.94582

The result for the parameter tuning can be found in Table 2. On the national level, the best performance was achieved for the combination of *mtry* = 2, *ntree* = 1000 and *nodesize* = 5. The precision score was 0.947. However, the results did not significantly vary for the different settings. The default setting reached a precision score of 0.945. The lowest performance still achieved a score of 0.942. In other words, the differences between the various parameter settings are marginal, indicating that the final prediction results are robust across the entire range of parameters.

A similar pattern can be observed for subnational conflicts. The best performance was achieved for the setting of *mtry* = 11, *ntree* = 500 and *nodesize* = 20. The precision value was 0.910. The worst performance still achieved a value of over 0.902. For the default setting (*mtry* = 5, *ntree* = 500 and *nodesize* = 1), the precision value was 0.905. Hence, also on the subnational conflict level, the actual impact of the tuning on the final prediction is marginal.

The results and the statistical metrics of the hyperparameters' application to the overall dataset are available in the following (Table 3).

Table 3. Model's statistical metrics using a cut-off point of 5.

	NP	**SN**
Accuracy	0.94796	0.91875
Kappa	0.74966	0.73144
Sensitivity	1	1
Specificity	0.63597	0.63694
Neg. Pred. Value	1	1
Precision	0.94276	0.90525
Recall	1	1
F1	0.97053	0.95027
AUC	0.81798	0.81847

Table 4. Confusion matrix for the NP a cut-off point of 5.

		Actual Conflict situation	
		Conflict (P)	No Conflict (N)
Predicted Conflict situation	Conflict (P)	**442 (TP)**	0 (FP)
	No Conflict (N)	253 (FN)	**4167 (TN)**

Table 5. Confusion matrix for the SN a cut-off point of 5.

		Actual Conflict situation	
		Conflict (P)	No Conflict (N)
Predicted Conflict situation	Conflict (P)	**693 (TP)**	0 (FP)
	No Conflict (N)	395 (FN)	**3774 (TN)**

Based on the results, we can further improve the models' performance using the Receiver Operating Characteristic (ROC) curve. The use of the ROC curve to calculate a threshold (cut-off point) is very common in the literature (Weidmann & Ward, 2010; Bean, Stafford & Brashares, 2012; Hegre et al., 2013). Selecting a threshold where the ROC curve starts bending would maximize sensitivity (minimizing omission rate), while minimizing the fall out rate (maximizing specificity) (Giancristofaro & Salmaso, 2003). In the particular case of conflict risk prediction, high sensitivity (low omission) is preferred over high specificity (low fall out).

Following our ROC analysis, the resulting new threshold value is equal to 4, which significantly improves our results for every metric (see Table 6).

Table 6. Model's statistical metrics using a cut-off point of 4.

	NP	**SN**
Accuracy	0.98786	0.98663
Kappa	0.94865	0.96068
Sensitivity	1	1
Specificity	0.9151	0.94025
Neg. Pred. Value	1	1
Precision	0.98603	0.98306
Recall	1	1
F1	0.99297	0.99146
AUC	0.95755	0.97012

With the tuned random forest, we can significantly improve our predictions for both the national and subnational level (see Table 7 and Table 8).

Table 7. Confusion matrix for the NP using a cut-off point of 4.

		Actual Conflict situation	
		Conflict (P)	No Conflict (N)
Predicted Conflict situation	Conflict (P)	**636 (TP)**	0 (FP)
	No Conflict (N)	59 (FN)	**4167 (TN)**

Table 8. Confusion matrix for the SN using a cut-off point of 4.

		Actual Conflict situation	
		Conflict (P)	No Conflict (N)
Predicted Conflict situation	Conflict (P)	**1023 (TP)**	0 (FP)
	No Conflict (N)	65 (FN)	**3774 (TN)**

Recently, a new variable, the internally displaced people (IDPs), was introduced in the GCRI indicators (Halkia et al., 2018). To evaluate the impact of the new variable on the AI GCRI predictions, we recalculate the RF

models including the IDPs. In **Error! Not a valid bookmark self-reference.**9 and table 10 below, you can find the statistical metrics using the different cut-off points.

Table 9. Model's statistical metrics using a cut- off point of 5 (including the IDP indicator).

	NP	**SN**
Accuracy	0.80213	0.82723
Kappa	0.17527	0.31614
Sensitivity	0.99788	0.9992
Specificity	0.12316	0.23069
Neg. Pred. Value	0.94366	0.98818
Precision	0.79788	0.81835
Recall	0.99788	0.9992
F1	0.88674	0.89978
AUC	0.56052	0.61495

Table 10. Model's statistical metrics using a cut-off point of 4 (including the IDP indicator).

	NP	**SN**
Accuracy	0.81838	0.8624
Kappa	0.2821	0.50621
Sensitivity	0.99311	0.99152
Specificity	0.21231	0.41452
Neg. Pred. Value	0.89883	0.93374
Precision	0.81389	0.85453
Recall	0.99311	0.99152
F1	0.89461	0.91794
AUC	0.60271	0.70302

It can be observed that the metrics are lower when the IDPs are included. The overall accuracy has decreased for both the national power and subnational model, as well as the precision.

6 Conclusions

The GCRI is a conflict risk model developed especially to support EU policy-making for conflict prevention. Based on linear and logistic regression models, the GCRI estimates the intensity or the probability of a violent conflict event. The linear regression approach (OLS) allows the researcher to investigate the marginal effect of a variable, holding all other variables of interest constant. The simplicity of its application and interpretation are the main advantages of the OLS model. However, a number of moderately strong assumptions need to be upheld such as normally distributed input data and homoscedasticity in the error terms. The logistic regression,

on the other hand, is a binary response method more appropriate for the conflict risk domain and does not present these major limitations.

The AI RF modelling approach, which is based on a decision process, extends the initial GCRI regression models, and reveals country specific indicators and their importance for conflict prevention.

The AI GCRI has been specified using the optimal hyperparameters, which were obtained through a 10-fold cross validation for each possible parameter combination. The models' dataset imbalance, inherent in the conflict modelling domain, partly addressed with the SMOTE oversampling technique indicates that the quality of input data is crucial for the models' performance. However, after the hyperparameters' tuning, we are able to increase the correct predictions for both national power and subnational conflicts.

Furthermore, we tested the internally displaced people (IDPs) variable, which has recently been introduced in the GCRI indicator basket. Even though there are good reasons to believe that IDPs are related to conflict, i.e. by providing resources for rebel groups, changing the ethnic balance and affecting the economy of the host area, the results of this experiment are inconclusive. The inconclusive results can be explained by the data quality of the IDP variable. For the IDPs variable, we have data only from 2009 to 2014 only for a small number of countries. Hence, the poor data availability makes reliable predictions difficult given the small number of cases. Moreover, most countries for which data exists, present rather small numbers of IDPs. Only in a few countries, such as Syria, Mexico or Colombia, the proportion of IDPs exceeds 10% of the total population. This high within-class imbalance may bias our results.

Overall, it is demonstrated that the random forest GCRI models are internally stable, not overfitting, and have a good predictive power. The precision and accuracy metrics are above 98%, both for the national power and subnational power conflict models.

Further research should however focus on (i) conducting an advanced evaluation of the variables' significance in order to select the best possible set of indicators, and (ii) reviewing the conflict definition according to the conflict data providers' updates.

References

Adam, Elhadi MI & Onisimo Mutanga (2012) Estimation of high density wetland biomass: combining regression model with vegetation index developed from Worldview-2 imagery. In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology XIV* Vol. 8531. SPIE, 85310V.

Adelabu, Samuel, Onisimo Mutanga & Elhadi Adam (2014) Evaluating the impact of red-edge band from Rapideye image for classifying insect defoliation levels. *ISPRS Journal of Photogrammetry and Remote Sensing* 95: 34–41.

Bean, William T, Robert Stafford & Justin S Brashares (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography* 35(3): 250–258.

Breiman, Leo (1996) *HEURISTICS OF INSTABILITY AND STABILIZATION IN MODEL SELECTION 1*. *The Annals of Statistics* Vol. 24.

Cederman, Lars-Erik & Nils B Weidmann (2017) Predicting armed conflict: Time to adjust our expectations? *Science (New York, N.Y.)* 355(6324): 474–476.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall & W Philip Kegelmeyer (2002) *SMOTE: Synthetic Minority Over-Sampling Technique*. *Journal of Artificial Intelligence Research* Vol. 16.

Colditz, René Roland (2015) An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*.

De Groeve, Tom, Peter Hachemer & Luca Vernaccini (2014) *The Global Conflict Risk Index (GCRI). A Quantitative Model*. Luxembourg:

de Santana, Felipe Bachion, Sarmento Júnior Mazivila, Lucas Caixeta Gontijo, Waldomiro Borges Neto & Ronei J Poppi (2018) Rapid Discrimination Between Authentic and Adulterated Andiroba Oil Using FTIR-HATR Spectroscopy and Random Forest. *Food Analytical Methods* 11(7): 1927–1935.

Díaz-Uriarte, Ramón & Sara Alvarez de Andrés (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7(January).

Giancristofaro, Rosa Arboretti & Luigi Salmaso (2003) Model performance analysis and model validation in logistic regression. *Statistica* 63(2): 375–396.

Guan, Haiyan, Jonathan Li, Michael Chapman, Fei Deng, Zheng Ji & X Yang (2013) Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*.

Halkia, Matina, Stefano Ferri, Inès Joubert-Boitat & Francesca Saporiti (2017) *Conflict Risk Indicators: Significance and Data Management in the GCRI*. Luxembourg: Publications Office of the European Union.

Halkia, Matina, Stefano Ferri, Marie Schellens, Michail Papazoglou, Dimitrios Thomakos & Francesca Saporiti (2019) The Global Conflict Risk Index: A quantitative tool for policy support on conflict prevention. *Progress in Disaster Science* (Under review).

Halkia, Matina, Stefano Ferri, Dimitrios Thomakos, Silvan Has & Francesca Saporiti (2018) *Global Conflict Risk Index : New Variables in 2018*.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand & Henrik Urdal (2013) Predicting Armed Conflict , 2010 – 2050. *International Studies Quarterly* 57(2): 250–270.

Janitza, Silke & Roman Hornung (2018) On the overestimation of random forest's out-of-bag error. *PLoS ONE* 13(8).

Millard, Koreen & Murray Richardson (2015) On the importance of training data sample selection in Random Forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing* 7(7): 8489–8515.

Nitze, Ingmar, Brian Barrett & Fiona Cawkwell (2015) Temporal Optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series. *International Journal of Applied Earth Observation and Geoinformation* 34: 136–146.

O'Brien, Robert & Hemant Ishwaran (2019) A random forests quantile classifier for class imbalanced data. *Pattern Recognition* 90(June): 232–249.

Pal, M (2005) Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26(1): 217–222.

Palmer, David S, Noel M O'Boyle, Robert C Glen & John BO Mitchell (2007) Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling* 47(1): 150–158.

Prasad, Anantha M, Louis R Iverson & Andy Liaw (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9(2): 181–199.

Rees, Heather, Mattias Nyström, Karin Nordkvist & Håkan Olsson (2014) Combining airborne laser scanning data and optical satellite data for classification of alpine vegetation. *International Journal of Applied Earth Observation and Geoinformation* 27(PARTA): 81–90.

Santos, Miriam Seoane, Jastin Pompeu Soares, Pedro Henrigues Abreu, Helder Araujo & Joao Santos (2018) Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. *IEEE Computational Intelligence Magazine* 13(4): 59–76.

Scornet, Erwan (2017) Tuning parameters in random forests. *ESAIM: Proceedings and Surveys* 60: 144–162.

Sesnie, Steven E, Bryan Finegan, Paul E Gessler, ALISTAIR MS Smith, Ramos Bendana Zayra & Sirpa Thessler (2010) The multispectral separability of Costa Rican rainforest types with support vector machines and Random Forest decision trees. *International Journal of Remote Sensing* 31(11): 2885–2909.

Stumpf, André & Norman Kerle (2011) Object-oriented mapping of landslides using Random Forests. *Remote Sensing of Environment* 115(10): 2564–2577.

Weidmann, Nils B & Michael D Ward (2010) Predicting conflict in space and time. *Journal of Conflict Resolution* 54(6): 883–901.

Zhang, Heping & Minghui Wang (2009) Search for the smallest random forest. *Statistics and Its Interface* 2(3): 381.

List of figures

List of Tables

Annexes

Annex 1. Model indicators and data sources

Recent Internal Conflict, Neighbours with HVC, Years since HVC: For all the conflict related variables we used the Battle related deaths, One-sided violence and Non-state conflict datasets provided by the UCDP/PRIO (Available online at: http://ucdp.uu.se/downloads/)

Regime Type, Lack of Democracy: We used the Polity IV Annual Time-Series, 1800-2015 dataset provided by the Center for Systemic Peace (CSP) (Available online at: http://www.systemicpeace.org/inscrdata.html)

Level of Repression: The data are provided by the Political Terror Scale Project (PTS) (Available online at: http://www.politicalterrorscale.org/Data/Download.html)

Empowerment Rights: We used the CIRI Human Rights Dataset provided by the Cingranelli and Richards (CIRI) Human Rights Data Project (Available online at: http://www.humanrightsdata.com/p/data-documentation.html)

Government Effectiveness, Corruption, GDP per capita, Openness, Oil Production, Homicide Rate, Infant Mortality, Unemployment: For these variables we used the World Bank's indicators (Available online at: https://data.worldbank.org/)

Income inequality: The Standardized World Income Inequality Database provided by Harvard Dataverse Network was used (Available online at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/11992)

Ethnic Power Change, Ethnic compilation: The ETH Zurich provided the Ethnic Power Relations (EPR) Core Dataset (Available online at: https://icr.ethz.ch/data/epr/)

Transnational Ethnic Bonds: We used the Minorities at Risk Dataset by the CIDCM Center for International Development & Conflict Management (Available online at: http://www.mar.umd.edu/)

Food Security: Food security indicators by FAO were used (Available online at: http://www.fao.org/home/en/)

Water Stress: We used the Aqueduct Country and River Basin Rankings (Raw country scores) dataset provided by the World Resources Institute (WRI) (Available online at: https://www.wri.org/)

Structural Constraints: The structural constraints variable by the Bertelsmann Stiftung's Transformation Index (BTI) was used (Available online at: https://www.bti-project.org/en/home/)

Population Size, Youth Bulge: The data for the population size and youth bulge are provided by the UN DESA/ Population Division (Available online at: https://www.un.org/development/desa/en/)

Climate: We used the Standardised Precipitation-Evapotranspiration Index (SPEI) dataset provided by the Institutional Repository of the Spanish National Research Council (DIGITAL.CSIC) (Available online at: http://digital.csic.es/bitstream/10261/153475/14/)

Internally Displaced People: We used the Internally Displaced Persons (IDP) dataset provided by the Internal Displacement Monitoring Centre (IDMC) (Available online at: https://data.worldbank.org/)

Annex 2. Distribution of the IDPs

Figure 3. Distribution of national power conflict for different level of IDPs.

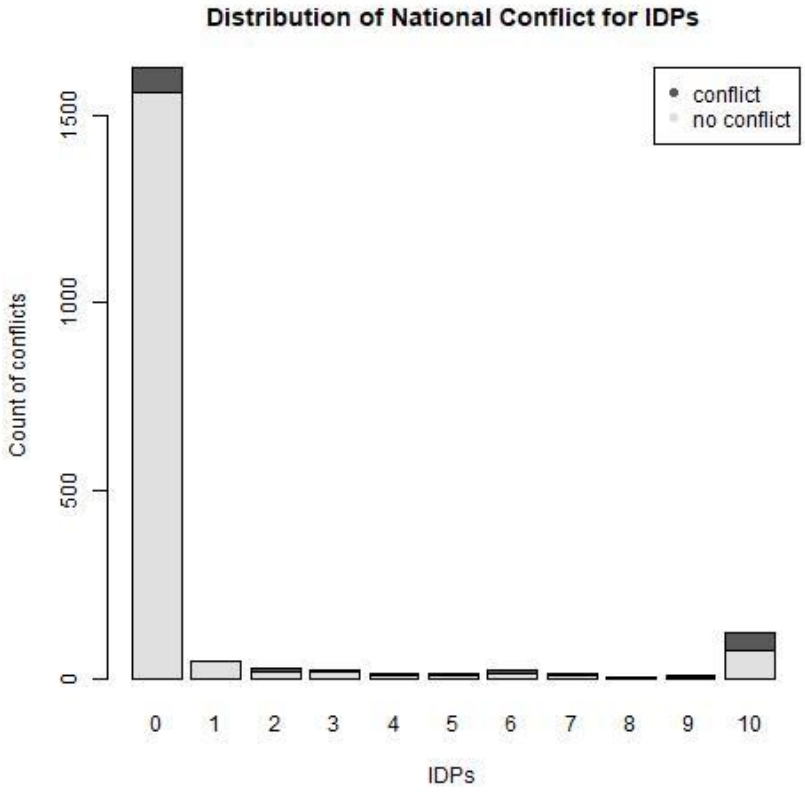


Figure 4. Proportions of national power conflict and no conflict for different levels of IDPs

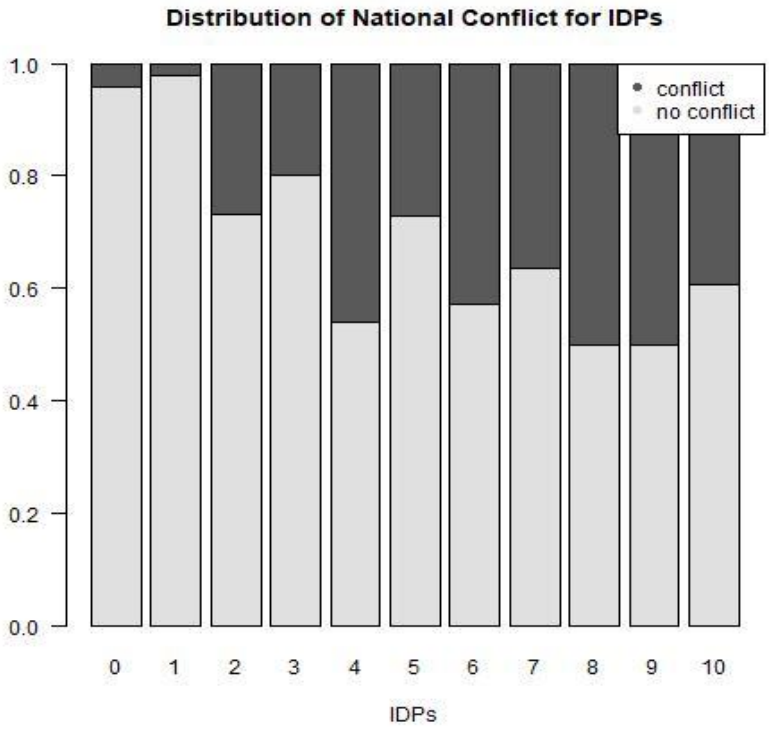


Figure 5. Distribution of subnational conflict for different level of IDPs

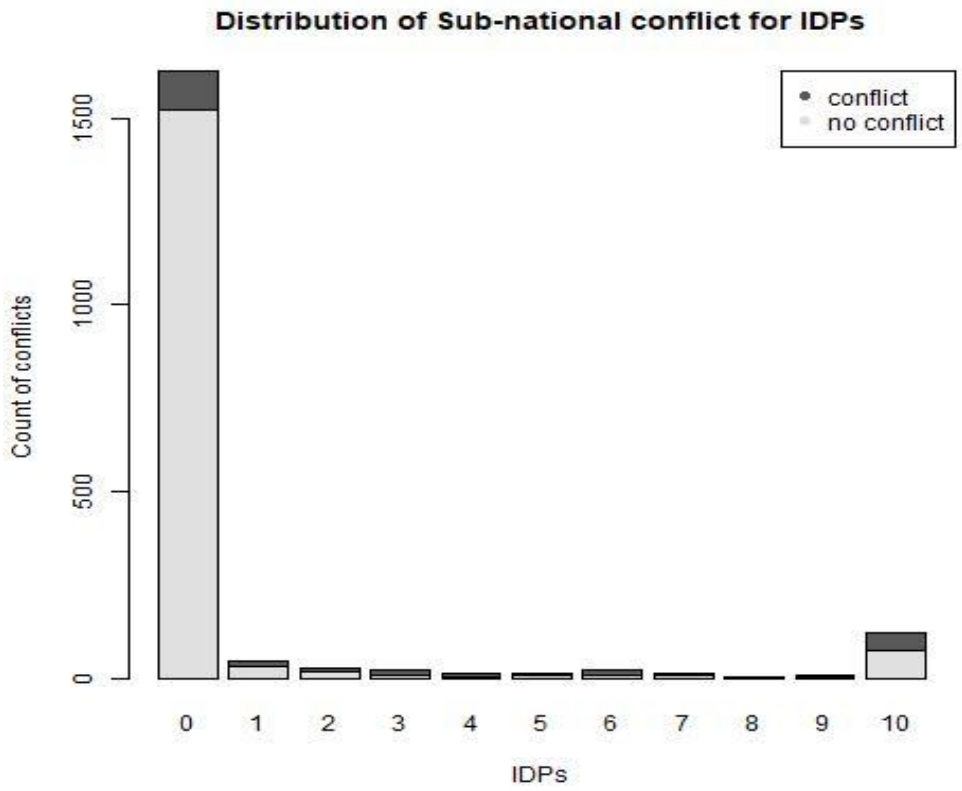
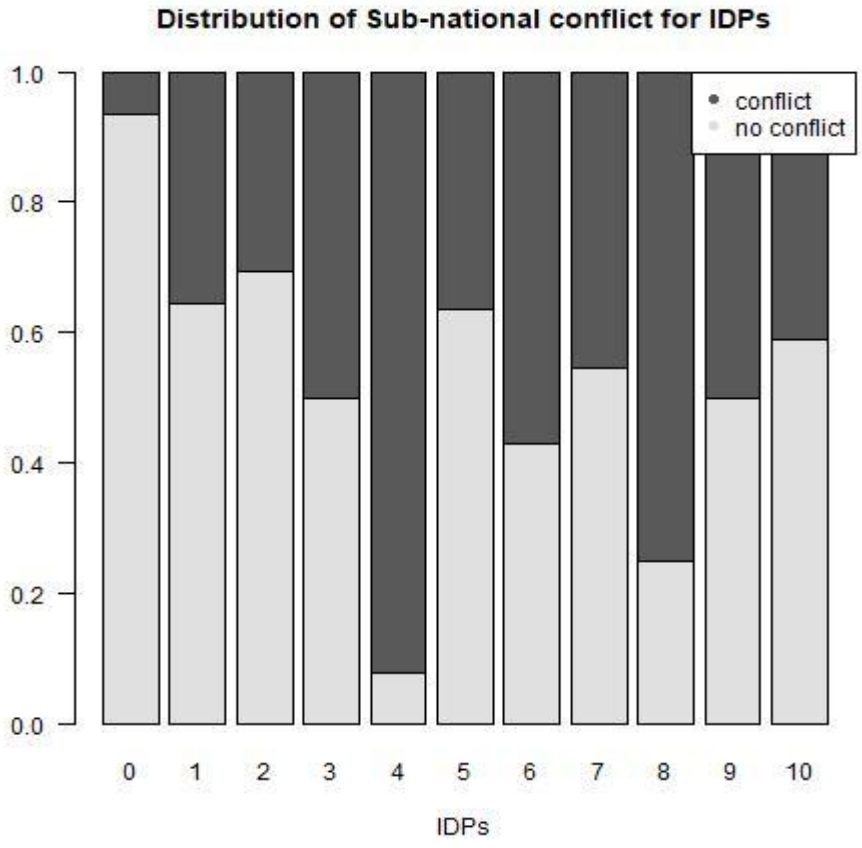


Figure 6. Proportions of subnational conflict and no conflict for different levels of IDPs



GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30028-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/004232

ISBN 978-92-76-14306-2