

RESEARCH ARTICLE

Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach

Fangyu Ding^{1,2}, Quansheng Ge^{1,2*}, Dong Jiang^{1,2*}, Jingying Fu^{1,2}, Mengmeng Hao^{1,2}

1 Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, **2** College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China

* geqs@igsnrr.ac.cn (QG); jiangd@igsnrr.ac.cn (DJ)



Abstract

Terror events can cause profound consequences for the whole society. Finding out the regularity of terrorist attacks has important meaning for the global counter-terrorism strategy. In the present study, we demonstrate a novel method using relatively popular and robust machine learning methods to simulate the risk of terrorist attacks at a global scale based on multiple resources, long time series and globally distributed datasets. Historical data from 1970 to 2015 was adopted to train and evaluate machine learning models. The model performed fairly well in predicting the places where terror events might occur in 2015, with a success rate of 96.6%. Moreover, it is noteworthy that the model with optimized tuning parameter values successfully predicted 2,037 terrorism event locations where a terrorist attack had never happened before.

OPEN ACCESS

Citation: Ding F, Ge Q, Jiang D, Fu J, Hao M (2017) Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. PLoS ONE 12(6): e0179057. <https://doi.org/10.1371/journal.pone.0179057>

Editor: Jinn-Moon Yang, National Chiao Tung University College of Biological Science and Technology, TAIWAN

Received: January 8, 2017

Accepted: May 23, 2017

Published: June 7, 2017

Copyright: © 2017 Ding et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are from the following websites, such as <https://www.start.umd.edu/gtd/>, <http://gecon.yale.edu/>, <http://www.icr.ethz.ch/data/index>, <http://www.univiena.org/unov/en/unode.html>, <http://ngdc.noaa.gov/eog/index.html>, and <http://neo.sci.gsfc.nasa.gov/>.

Funding: This research was supported and funded by the National Natural Science Foundation of China (Grant NO. 41571509, <http://www.nsfc.gov.cn/>) and Ministry of Science and Technology of China (2016YFC1201300, <http://www.most.gov.cn/>)

Introduction

Terrorism is considered to be a major threat to society [1–3]. According to the Global Terrorism Database (GTD), more than 14800 terrorism events of different types occurred globally in 2015 alone[4], which not only caused the deaths of 38430 people but also caused the world to panic[5–7]. Great efforts have been made to seek explanations of various issues related to terrorist threats[8, 9]. However, the prediction of the occurrence of a certain event is still a difficult task of great complexity and uncertainty[10, 11]. Currently, data-driven models, which are direct learning patterns from large volumes of high-dimensional training data, have achieved some successes in a variety of fields[12–15], which provide a new perspective to solve this problem.

Previous studies have showed that natural[16] and social[17] factors are used in agent-based method to simulate the potential future scenarios of violent conflict, but the underlying mechanisms are still not well-understood. Everything is connected. Perhaps, there are some regularity of terrorist attacks[18]. Thus, the typical intuition of geographers posits the question of whether the occurrences of the terrorism events may be influenced or controlled by sets of

(eng/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

geographical, natural and social issues (factors). If these patterns exist, then they can be discovered through machine learning approach.

Given the availability of global data, we collected three social elements, such as ethnic diversity, major drug regions and population density. Moreover, nighttime light are also included, which has been widely verified for monitoring economic activity[19]. Meantime, some natural elements were selected in the present study, including average precipitation, average temperature and topography. Considering the difference of these factors in space, latitude, longitude, distance to a major navigable lake, distance to an ice-free ocean, and distance to a major navigable river were used to express geospatial information. As a result, a total of 12 factors were quantified at a global scale (0.1×0.1 degree). Thanks to the excellent job of the STAR group at the University of Maryland, GTD was established, and more than 150,000 events were recorded in detail. We rasterized the location of terror events that occurred from 1970 to 2014, dividing units into 2 classes according to the frequency of a location and the consequence (i.e., the number of deaths).

The specific flowchart is shown in Fig 1. To create simulation model, we develop a series of programs and use geographic Information Systems (GIS) software to preprocess various types of data and to select a sample dataset (see Methods). Three relatively popular machine learning algorithms are used in the present study: neural network (NNET), support vector machine (SVM), and random forest (RF). Additionally, the training approach and testing of the machine learning models are described in the Methods. Notably, the research scope is limited to the region between 75°N and 55°S .

Methods

Data preprocessing

The types and sources of data used in this research are shown in Table 1. GIS software and the C++ programming language were used to preprocess the datasets, including ArcMap 10.2 (<http://www.esrichina.com.cn/>), GDAL 2.1.0 (<http://www.gdal.org/>) and Proj.4 (<https://github.com/OSGeo/proj.4>). WGS-84 was selected as the geographic coordinate system for all of the data used in this research.

Based on GTD, we can obtain the location of each incident and convert information on terrorist events around the world from 1970 to 2014 into raster data. Pixels of 0.1×0.1 degrees resolution were chosen as the units to statistically determine the number of terrorist events and the total number of fatalities. In the present study, an assessment unit is considered as having a high terrorist attack probability if there have been a terrorist attack causing casualties in the past; otherwise, it has a low probability. Class values of 1 and 0 were given for these types. In addition, we can obtain the latitude and longitude information of the centre point of each unit.

Based on G-Econ 4.0, a dataset on economic activity for the world, we can obtain the raster data of 5 factors, namely distance to major navigable lake (km), distance to major navigable river (km), distance to ice-free ocean (km), average precipitation (mm/year) and average temperature ($^{\circ}\text{C}$) from 1980 to 2008. Then, we resample the above mentioned raster data into 0.1-degree longitude by 0.1-degree latitude resolution using ArcMap 10.2 because each terrestrial observation of G-Econ was measured at a 1-degree longitude by 1-degree latitude resolution at a global scale.

The ethnic distribution map was derived from GeoEPR at a 0.01-degree longitude by 0.01-degree latitude resolution using the ArcMap 10.2 “Polygon to Raster” function. In the present study, we assume that there are no intrinsic differences between ethnic groups with respect to terror events, but the conflict of different cultural beliefs with neighbouring regions

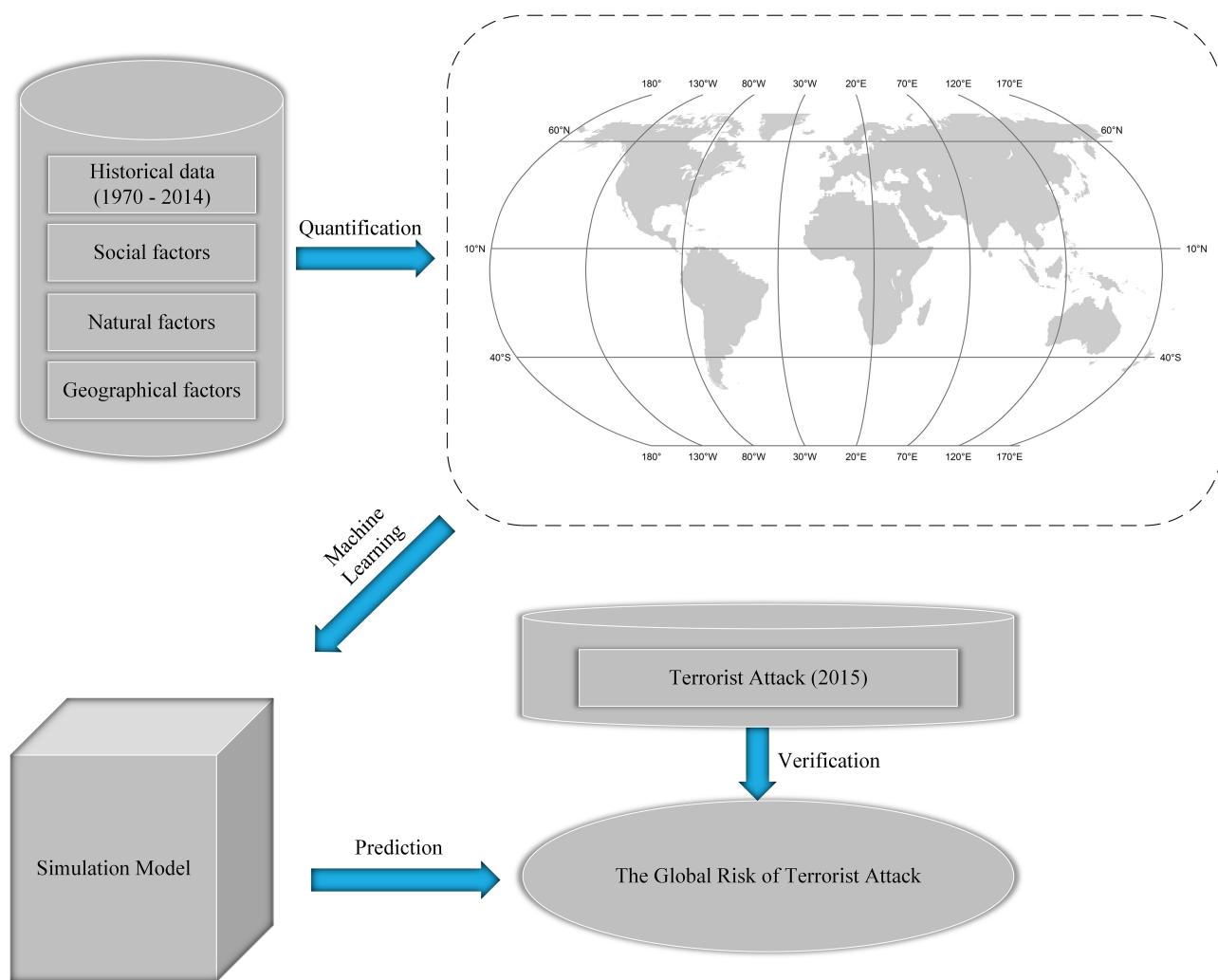


Fig 1. Workflow of the simulation the risk of terrorist attacks in all regions worldwide.

<https://doi.org/10.1371/journal.pone.0179057.g001>

may result in terror events. We used 0.1×0.1 degrees resolution pixels as the units to statistically determine the number of different ethnic groups in each region.

Major drug regions were derived from world drug report and national administrative boundaries (<http://www.gadm.org/>) using the ArcMap 10.2 “Polygon to Raster” and “Resample” functions. The remaining factors (nighttime lights, population density and topography) were calculated from their respective data sources and resampled to 0.1×0.1 degrees resolution pixels using ArcMap 10.2.

Another point to note is that the twelve potential factors mentioned above have different units. To facilitate the training step of models, they were normalized as the Eq 1 below:

$$P'_j = \frac{P_j - P_{\min}}{P_{\max} - P_{\min}} \quad (j = 1, 2, 3 \dots, n) \quad (1)$$

where P'_j is the normalized value between 0 and 1; P_j is the value of the factor; P_{\min} is the minimum value; P_{\max} is the maximum value; and n is the number of data.

Table 1. Data used to obtain the factors of global terrorist attacks.

Data type	Source	Description	Publisher
Historical data	The Global Terrorism Database (GTD)	Table	National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland, https://www.start.umd.edu/gtd/
Latitude			
Longitude			
Distance to major navigable lake	G-Econ 4.0	Table	Yale University, http://gecon.yale.edu/
Distance to major navigable river			
Distance to ice-free ocean			
Average precipitation			
Average temperature			
Ethnic diversity	GeoEPR, the Ethnic Power Relations dataset, version 2014	Polygon data	Center for Comparative and International Studies (CIS), International Conflict Research, ETH Zurich, http://www.icr.ethz.ch/data/index
Major drug regions	World drug report	Table	Division for Policy Analysis and Public Affairs, United Nations Office on Drugs and Crime, http://www.unvienna.org/unov/en/unodc.html
Nighttime lights	Nighttime Lights of the World, 2013	Grid	The Earth Observation Group, NOAA, http://ngdc.noaa.gov/eog/index.html
Population density	Population density of the World, 2000	Grid	NASA's Earth Observatory, http://neo.sci.gsfc.nasa.gov/
Topography	Digital elevation model (DEM), 2000		

<https://doi.org/10.1371/journal.pone.0179057.t001>

Sampling data

The values of the attributes of a unit were gathered from the same location at each data layer and together constitute a sample row. If the location of some terrorist attacks is close, they will be recorded in the same statistical unit. Besides, small parts of units that were missing some factors' information were excluded. In the present research, pixels where terrorist attacks resulted in more than 0 of the number of casualties were selected, and the same number of pixels were randomly selected from the remaining pixels (Fig 2). This procedure ensured that sample rows of terrorist attacks that resulted in casualties occupied only half of the entire sample. Finally, we can obtain 15970 sample data. To train and test the performance of machine learning models, three-quarters of the sample data were randomly selected as training data, and the remaining data were used as the test data, which were almost evenly distributed within the whole world.

Machine-learning algorithms

We used version 3.30 of the 64-bit version of R, an open-source language and software for statistical computing[20], to build models, tune parameters and perform accuracy assessments. Three relatively popular and robust machine learning algorithms (NNET, SVM and RF) were adopted in the present study. The “caret” package[21] and built-in functions were used to train each model and tune their associated parameters within a single consistent environment.

For NNET, we use the backward propagation neural network (BPNN) algorithm, which has a powerful synaptic modification rule that will allow an arbitrarily connected neural network to develop an internal structure that is appropriate for a particular task domain[22]. Studies of BPNN suggest that increasing the number of hidden neurons improves classification accuracy[23, 24] and a large decay rate results in non-convergence[22]. Thus, the value of decay and the number of hidden units of the hidden layer were tuned when running a

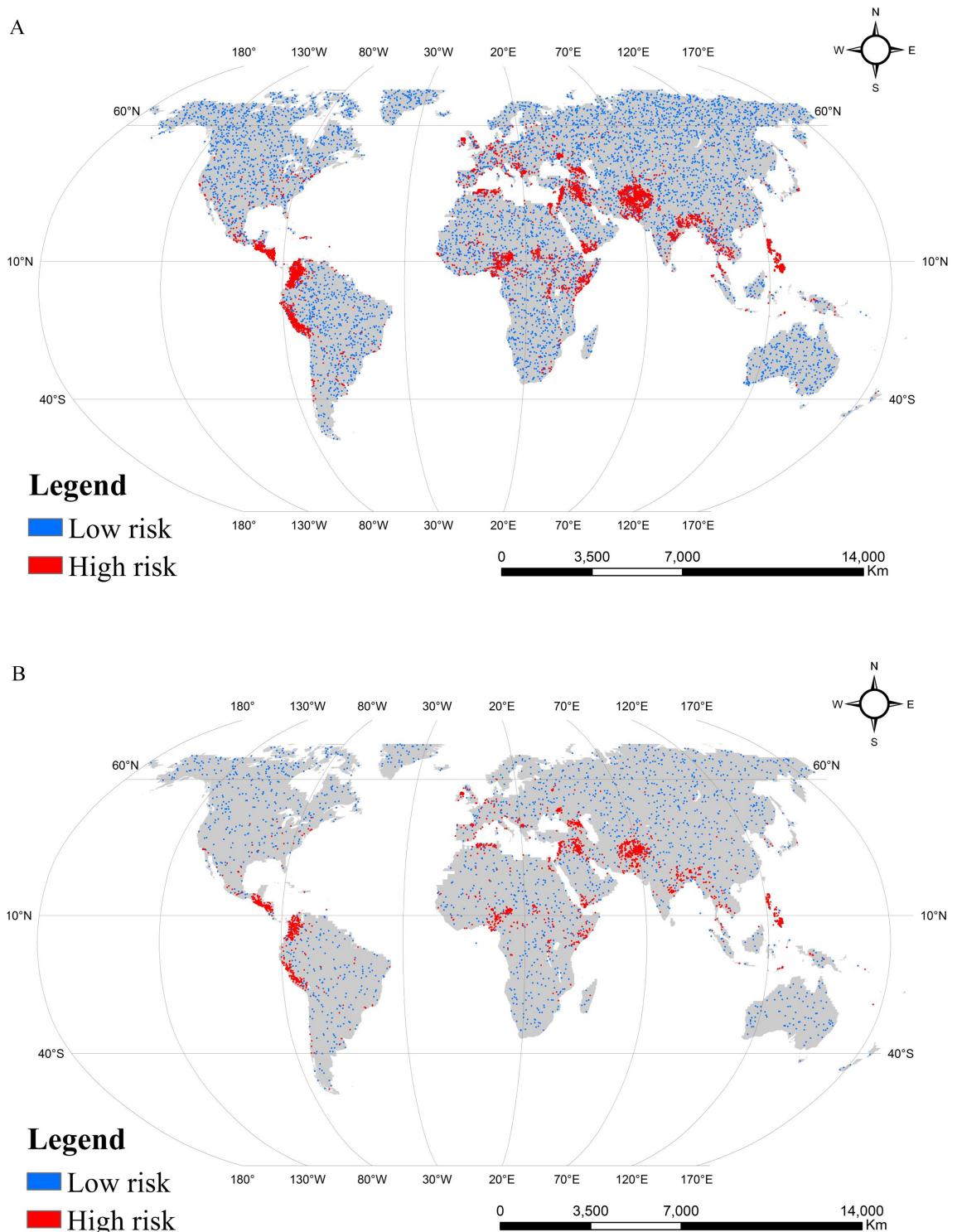


Fig 2. Spatial distribution of samples. (A) Training samples with 11,978 assessment units. Among the units, 5,989 have experienced terrorist attacks that resulted in casualties (= high risk), whereas 5,989 have not (= low risk). (B) Validation samples with 3,992 assessment units. Among the units, 1,996 have experienced terrorist attacks that resulted in casualties (= high risk), whereas 1,996 have not (= low risk).

<https://doi.org/10.1371/journal.pone.0179057.g002>

backward propagation neural network with sigmoid activation functions in the hidden and output layers.

For SVM based simulation, a radial basis function (RBF) kernel was considered for use in this research. Previous studies of the SVM algorithm had emphasized that two parameters (cost, sigma) used in the RBF kernel have an impact on classification accuracy[25]. A large cost means assigning a higher penalty to errors[26], whereas increasing sigma affects the shape of the separating hyperplane[27]. Therefore, we needed to fine-tune the cost and sigma parameters when performing simulation with the RBF kernel.

For simulation built with the RF model, the default number of trees (500) was used because values larger than 500 were unable to significantly improve the performance of the RF algorithm[28]. The m_{try} parameter controls the number of variables randomly sampled as candidates at each split and is known to have an effect on the overall classification accuracy[29]. For the binary partitioning, a small number of randomly selected variables are available, such as the value of the m_{try} parameter, which equals the square root of the number of variables within a dataset[30]. In the present study, only the m_{try} parameter was tuned when running the RF algorithm.

Results

Before using machine learning models to simulate the risk of terrorist attacks, we need to tune some parameters. The area under the curve (AUC) is used as a metric for measuring the performance of models. Larger AUC values indicate higher accuracy. Based on this criteria, we selected models with optimized tuning parameter values. For the NNET model, the highest AUC value (0.966) was obtained with 30 hidden units and a decay value of 0.1 ([Fig 3A](#)). For the simulation using SVM model, a total of 5 values for the cost parameter (2, 4, 8, 16 and 32) and 6 values for the sigma parameter (0.08, 0.09, 0.1, 0.12, 0.14 and 0.16) were examined.

Based on the highest AUC obtained (0.962), a cost value of 32 was selected for the SVM model, whereas the value for the sigma parameter was set at 0.09 ([Fig 3B](#)). A total of 12 m_{try} parameter values (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12) were examined for the simulation based on RF model. The highest AUC value (0.974) was obtained with a m_{try} value of 2 for the RF model ([Fig 3C](#)). Visualizing the resampling distributions, the RF model has better performance than the NNET and SVM models during the cross validation process on training samples. Moreover, the optimized models achieved an AUC value of over 0.95, which proved that there were some patterns between terrorist attacks and 12 factors ([Fig 3D](#)).

Based on a comparison between machine learning models applied to training datasets, we find that the RF model obtained the highest AUC (1) value, followed by the NNET (0.980) and SVM (0.976) models ([Fig 3E](#)). The same general trend was observed for machine learning models applied to validation datasets; the RF model had the highest AUC (0.976) value, followed by the NNET (0.971) and SVM (0.966) models ([Fig 3F](#)). In addition, the McNemar[31] test was used to assess whether a statistically significant difference existed between different machine learning models when using the validation samples. The McNemar test shows that the difference between the machine learning models based on the validation data was not significantly different ($p > 0.05$).

According to the results of the comparative analysis, we utilized the RF model, which obtained the best performance, to predict the risk of terrorist attacks in all regions worldwide. The results demonstrated that the high risk zones of terrorist attacks are mainly concentrated in South Asia, the Middle East, Central Africa, Northern Africa, Northwestern South America, Southern North American, Southern Europe and Western Europe ([Fig 4A](#)).

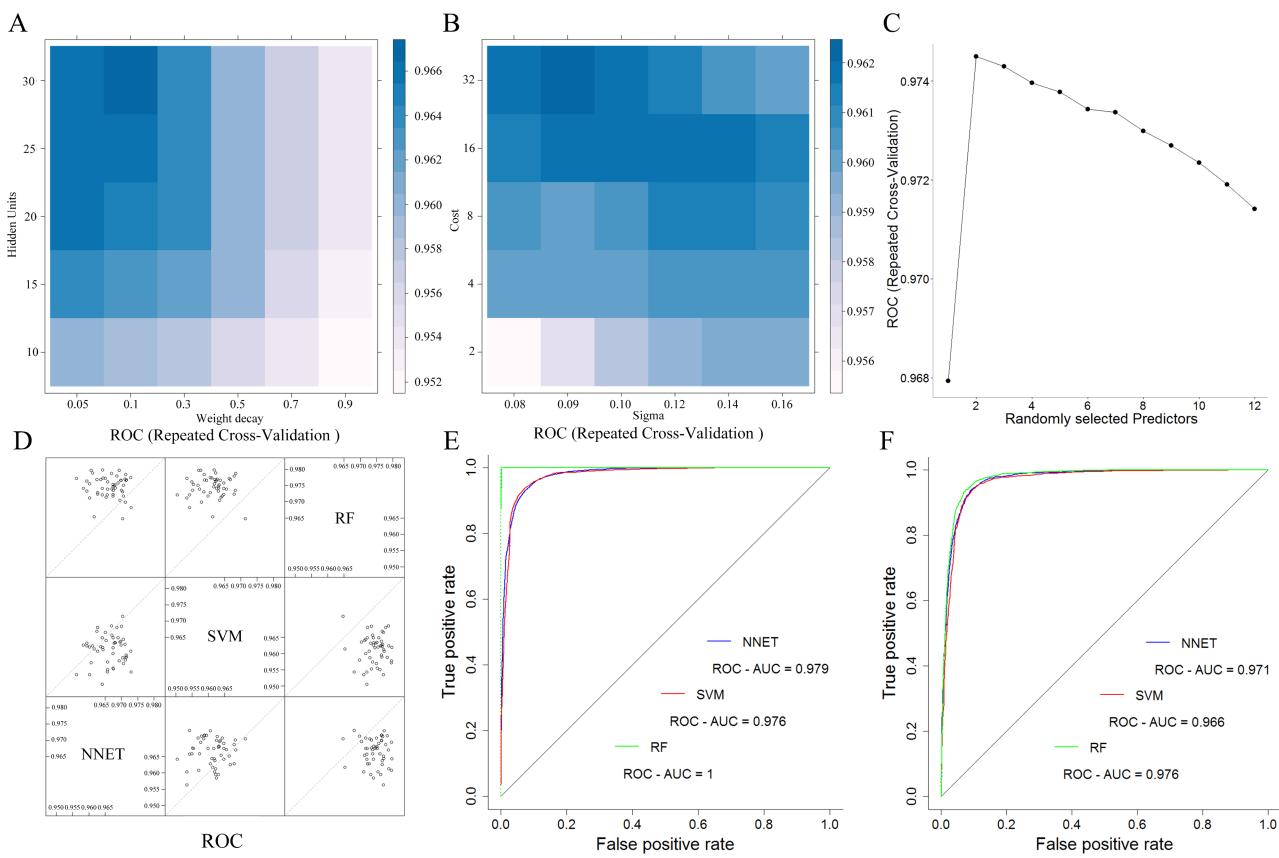


Fig 3. Selecting the optimal models using five repetitions of a 10-fold cross-validation. (A) Tuning the parameters of NNET. (B) Tuning the parameter of SVM. (C) Tuning the parameter of RF. (D) Comparing the performance of multiple models based on the same versions of the training data during the cross-validation process. (E) Receiver operating characteristic (ROC) curves of NNET, SVM and RF applied to training samples. (F) ROC curves of NNET, SVM and RF applied to validation samples.

<https://doi.org/10.1371/journal.pone.0179057.g003>

We divide the validation results derived from terrorism events that happened in 2015 into 4 categories (A, B, C, D) and emphasize that some terrorist attacks that happened in units that are missing some factors' information are excluded in this step. Generally, 11,632 terrorist attacks happened in high risk regions, and the prediction accuracy of the RF model was 96.6%. More importantly, the RF model successfully predicted 2,037 terrorism event locations where a terrorist attack did not occur between 1970 and 2014 but occurred in 2015 (Fig 4B). In addition to this, the optimal model performed fairly well in predicting the places where terror events might occur in 2013 and 2014, with success rates of 96.0% and 94.7%, respectively.

Furthermore, we were also interested in understanding the mechanism between the 12 factors and terrorist attacks. Thus, we tried to statistically analyze the discriminatory power of each factors. The measurement was performed using the “Boruta” package[32], the default value (0.01) of confidence level was used and the maximal number of importance source runs was set to 100. The Fig 5 shows that the above geographical, natural and social factors are essential for simulation the risk of terrorist attacks in global regions. Additionally, the result suggests that the population density, nighttime lights and the location (latitude, Longitude) of the units play more important roles than the remaining factors in distinguishing whether a pixel belongs to a high risk unit, which is similar to that derived from F-score method. The population density factor has the highest mean importance (66.49), followed by latitude

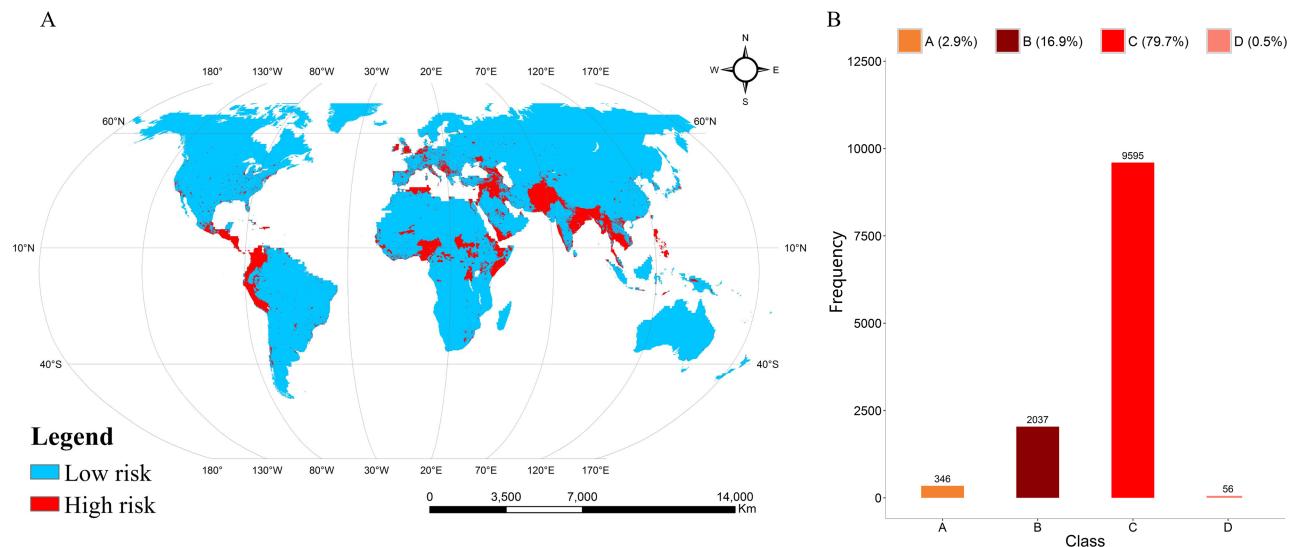


Fig 4. Simulation and verification. (A) The risk of terrorist attacks predicted by the RF-based model. (B) Verification of the prediction accuracy of the RF model using terrorist attacks data in 2015. Terrorism event locations that have no history of terrorist attacks in the past and lie in the low risk region of the prediction map belong to A. Terrorism event locations that have no history of terrorist attacks in the past and lie in the high risk region of the prediction map belong to B. Terrorism event locations that have a history of terrorist attacks and lie in the high risk region of the prediction map belong to C. Terrorism event locations that have a history of terrorist attacks and lie in the low risk region of the prediction map belong to D.

<https://doi.org/10.1371/journal.pone.0179057.g004>

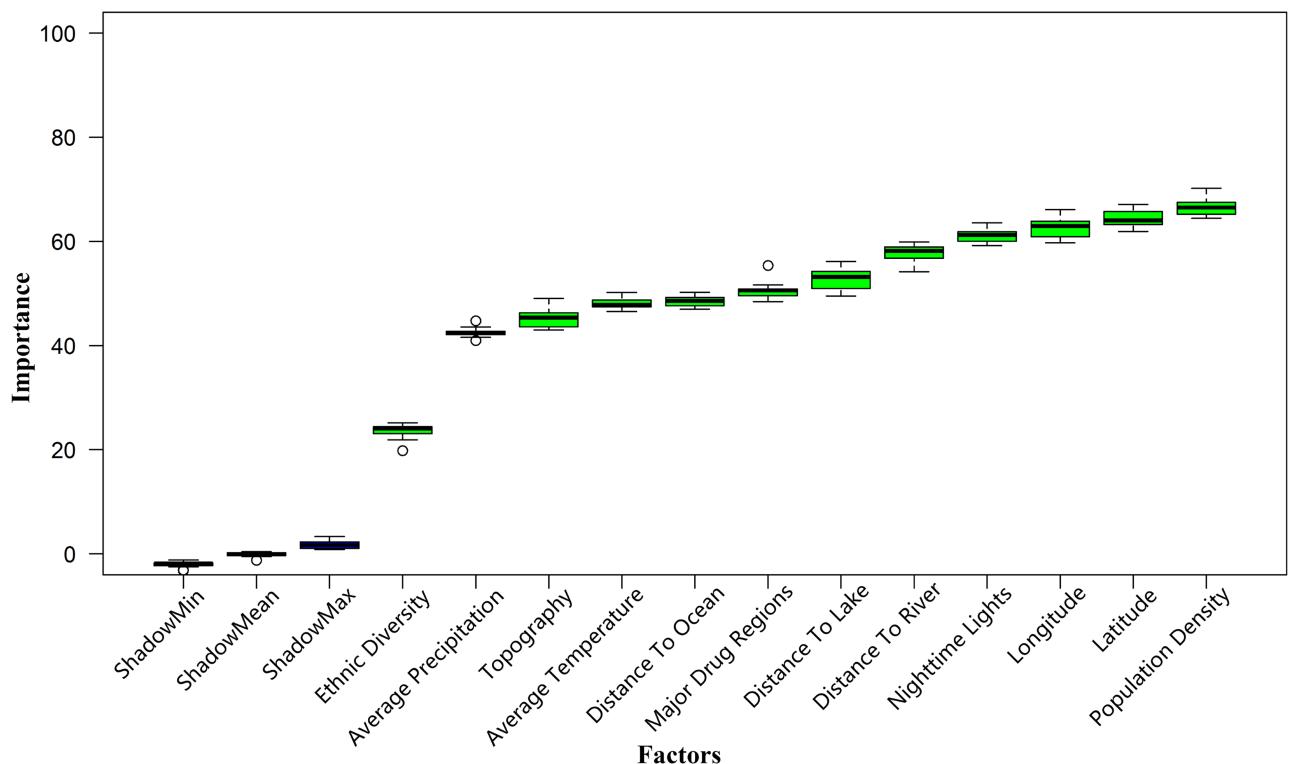


Fig 5. Measuring the discriminatory power of each factors.

<https://doi.org/10.1371/journal.pone.0179057.g005>

(64.45), longitude (62.50), nighttime lights (61.13), distance to a major navigable river (58.88), distance to a major navigable lake (53.28) and major drug region (52.86). Given the mean importance of major drug region, it may be reasonable and explainable by the reference 8, which writes that terrorist groups tax the drug trade as a source of funds.

Conclusions

In this work, we demonstrate a novel strategy based on large amounts of data that uses machine learning models that are relatively typical and robust, though not the most advanced, to simulate the risk of terrorist attacks in all regions worldwide, which proved to be reasonable. Notably, the successful prediction of the risk of terrorist attacks using machine learning models was critically dependent on large volumes of high-dimensional data, including geographical, natural and social factors, which are quantized into space using dataset preprocessing (see [Methods](#)). Additionally, we recognize the contributions of other researchers made to the data source ([Table 1](#)). These data may not be the most representative in simulation risks of terrorist attacks worldwide, but it can be downloaded for free and quantified into the space. Moreover, the importance of each factors is summarized based on Boruta algorithm, the results of which are data-driven and the specific reason for those need further study.

Unfortunately, some terror events happened during this research, such as the Nice attack, the Brussels bomb attacks, the Orlando nightclub attack, the Wuerzburg train attack, and the Munich shootings, which caused a large number of casualties. To further verify the correctness of the prediction, we tested the locations of these terrorist incidents against the prediction results, which shows that the above terrorist attacks happened in high risk areas. Moreover, no significant difference was found among machine learning models (NNET, SVM and RF) based on the validation samples.

The next step will focus on the finer spatial and temporal scales to analyze the terror events, such as the distribution of terrorist attacks in different regions and the characteristics of terror attacks between the different years. Additionally, we will make better use of information from the GTD instead of using only the position and the number of casualties of each event when using machine learning models with rich data, which might reveal more key principles hidden in data.

Acknowledgments

We thank Dr. Kun Liu and Dr. Gang Lin in Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences for their help in writing this paper. We like to thank the editors and anonymous reviewers for their helpful remarks.

Author Contributions

Conceptualization: QG DJ.

Data curation: FD DJ.

Formal analysis: FD DJ.

Funding acquisition: DJ.

Investigation: DJ.

Methodology: FD DJ.

Project administration: FD QG DJ.

Resources: FD QG DJ.

Software: FD DJ JF.

Supervision: QG DJ.

Validation: FD DJ.

Visualization: FD MH.

Writing – original draft: FD.

Writing – review & editing: QG DJ.

References

1. Beck U. The terrorist threat world risk society revisited. *Theory, culture & society*. 2002; 19(4):39–55.
2. Wein LM, Baveja M. Using fingerprint image quality to improve the identification performance of the U. S. Visitor and Immigrant Status Indicator Technology Program. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(21):7772–5. <https://doi.org/10.1073/pnas.0407496102> PMID: 15894628;
3. Schneider CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ. Mitigation of malicious attacks on networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108(10):3838–41. <https://doi.org/10.1073/pnas.1009440108> WOS:000288120400009. PMID: 21368159
4. LaFree G, Dugan L. Introducing the global terrorism database. *Terrorism and Political Violence*. 2007; 19(2):181–204.
5. Stecklov G, Goldstein JR. Terror attacks influence driving behavior in Israel. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(40):14551–6. <https://doi.org/10.1073/pnas.0402483101> PMID: 15448203.
6. Sharot T, Martorella EA, Delgado MR, Phelps EA. How personal experience modulates the neural circuitry of memories of September 11. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104(1):389–94. <https://doi.org/10.1073/pnas.0609230103> PMID: 17182739;
7. Hersh ED. Long-term effect of September 11 on the political behavior of victims' families and neighbors. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110(52):20959–63. <https://doi.org/10.1073/pnas.1315043110> PMID: 24324145;
8. Makarenko T. The crime-terror continuum: tracing the interplay between transnational organised crime and terrorism. *Global crime*. 2004; 6(1):129–45.
9. Li Q, Schaub D. Economic globalization and transnational terrorism a pooled time-series analysis. *Journal of Conflict Resolution*. 2004; 48(2):230–58.
10. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Utility, probability, and human decision making*: Springer; 1975. p. 141–62.
11. Bakker R, Hill DW Jr, Moore WH. Modeling terror attacks: A cross-national, out-of-sample study. *Understanding Terrorism (Contributions to Conflict Management, Peace Economics and Development, Volume 22)* Emerald Group Publishing Limited. 2014; 22:51–68.
12. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature*. 2015; 518(7540):529–33. <https://doi.org/10.1038/nature14236> PMID: 25719670
13. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016; 529(7587):484–9. <https://doi.org/10.1038/nature16961> PMID: 26819042
14. Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016; 533(7601):73–6. <https://doi.org/10.1038/nature17439> PMID: 27147027
15. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. *Science*. 2015; 350(6266):1332–8. <https://doi.org/10.1126/science.aab3050> WOS:000366162400032. PMID: 26659050
16. Ball P. The Transformation of War: Modelling Modern Conflict2012. 48–52 p.
17. Taylor G, Frederiksen R, Vane RR, Waltz E, editors. *Agent-based simulation of geo-political conflict*. 2004 Innovative Applications of Artificial Intelligence; 2004.

18. Johnson NF, Spagat M, Restrepo JA, Becerra O, Bohorquez JC, Suarez N, et al. Universal patterns underlying ongoing wars and terrorism. *Eprint Arxiv Physics*. 2006.
19. Ma T, Zhou C, Tao P, Haynie S, Fan J. Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities. *Remote Sensing of Environment*. 2012; 124:99–107.
20. Development BR. Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2010.
21. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. 2008; 28(5):1–26. WOS:000260799600001.
22. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cognitive modeling*. 1988; 5(3):1.
23. Tjoa MP, Krishnan SM. Feature extraction for the analysis of colon status from the endoscopic images. *Biomedical Engineering Online*. 2003; 2(1): 9.
24. Ciresan DC, Meier U, Gambardella LM, Schmidhuber J. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*. 2010; 22(12):3207–20. https://doi.org/10.1162/NECO_a_00052 PMID: 20858131
25. Alpaydin E. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*: The MIT Press; 2004. 28 p.
26. Pal M, Mather PM. Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generation Computer Systems*. 2004; 20(7):1215–25.
27. Huang C, Davis LS, Townshend JRG. An assessment of support vector machines for land cover classification. *Int J Remote Sens*. 2002; 23(4):725–49.
28. Breiman L, Cutler A. Random forests-classification description. Department of Statistics, Berkeley. 2007; 2.
29. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002; 2(3):18–22.
30. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. *Ecology*. 2007; 88(11):2783–92. PMID: 18051647
31. Eliasziw M, Donner A. Application of the McNemar test to non-independent matched pair data. *Statistics in medicine*. 1991; 10(12):1981–91. PMID: 1805322
32. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *Journal of Statistical Software*. 2010; 36(11):1–13. WOS:000282057200001.