Análise Exploratória da Global Terrorism Database

Trabalho desenvolvido na discipina de Inteligência Artificial

Daniel Gunna
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais
danielgunna1408@gmail.com

Felipe Coelho Silva
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais
fcs4200@gmail.com

ABSTRACT

Terrorismo é um tópico de grande importância, pois seus eventos influenciam a vida não apenas das vítimas diretas, mas também da sociedade como um todo. Isto ocorre por conta da propagação do medo, através de ameaças e terror generalizado. Compreender a natureza destes eventos é crucial neste contexto. Neste trabalho utilizamos uma base de dados, fornecida pela "National Center for the Study of Terrorism and Responses to Terrorism"(START), dos Estados Unidos. Essa base possui registros de ataques terroristas de 1970 a 2016. A exploração destes registros é dividida em duas etapas: Na primeira, são utilizadas visualização de dados para extrair e compreender os dados contextualizados, dessa forma conduzindo uma análise descritiva dos mesmos, objetivando aprofundar o entendimento em relação aos atributos da base e suas correlações. Por fim, serão apresentados modelos classificatórios para definir se um evento está associado a casualidades ou não, para este último passo, foi utilizada uma base de dados complementar para associar as condições climáticas no dia do evento ocorrido. Finalmente, conduziu-se a avaliação modelos, que indicou boa qualidade classificatória e validou a correlação dos dados climáticos com a ocorrência de casualidades.

ACM Reference Format:
Daniel Gunna and Felipe Coelho Silva. 2017. Análise Exploratória da Global Terrorism Database: Trabalho desenvolvido na discipina de Inteligência Artificial. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

O terrorismo causa grande impacto na sociedade, modificando o modo como as pessoas percebem a realidade, e está relacionado com grande aumento em níveis de estresse e medo. Como apresentado na Figura 1, os últimos anos apresentam crescimento da quantidade de eventos de terrorismo. Neste contexto, é importante compreender e explorar a natureza desses eventos.

De acordo com Henderson [6] pelo menos um ataque terrorista ocorre todo dia e, ao contrário de casualidades naturais, a quantidade de casualidades por conta de eventos de terrorismo vem crescendo ao longo dos anos, fato condizente com o que é apresentado na Figura 1.

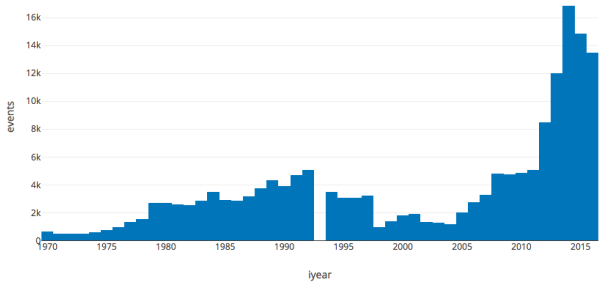

Figura 1: Distribuição dos eventos de terrorismo ao longo dos anos.

No desenvolvimento deste trabalho, escolhemos utilizar a base de dados *"The Global Terrorism Database"* fornecida e mantida pela START (National Consortium for the Study of Terrorism and Responses to Terrorism) da *University of Maryland*. Essa base de dados está disponível para acessos através da plataforma Kaggle [9], e é atualizada anualmente.

A base de dados traz informações sobre ataques terroristas de 1970 a 2016, com 170.350 registros. Cada registro possui 135 atributos associados, resumidos a seguir:

- **Eventid:** Identificador único para o evento
- **Dados Temporais:** Ano, Mês e Dia são apresentados. Também há um campo para datas aproximadas para quando não há informação da data exata de ocorrência do evento. Além disso, para eventos de múltiplos dias, há a informação de quando se deu a resolução.
- **Dados de Localização:** País, Região, Província ou Estado e Cidade. Além disso, as informações de localização também são disponibilizadas em formato de latitude e longitude e um campo para designar a localização mais especificamente (indicando prédio, edifício ou similar.
- **Informação sobre o incidente:** É apresentada uma breve descrição do ocorrido. Campos indicam se o incidente é: Grupo 1: De natureza Política, Econômica, Religiosa ou Social. Grupo 2: Para intimidar ou obter exposição midiática. Grupo 3: Não respeita as leis de guerra. Sendo estes grupos não-exclusivos.
- **Configuração do Ataque:** Há uma indicação quando há dúvidas se o incidente foi realmente um ataque terrorista, neste caso também se designa qual a categoria do incidente pode ser (Guerra, outros crimes, sem intenção, etc). Além

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

disso, se o ataque for conectado a uma série de ataques, indica-se os incidentes relacionados.

- **Informações sobre o ataque:** Tipo de ataque (assassinado, bombardeamento, etc), podendo haver múltiplos categorias associadas a um mesmo registro. Há a indicação do sucesso ou fracasso do ataque, assim como se o ataque foi suicida ou não. O tipo de arma utilizado no ataque também é indicado.
- **Alvo:** Informa-se o tipo de alvo (Empresa, Governo, Polícia, Militar, etc), grupo étnico e nacionalidade.
- **Executores:** Indica-se nome do grupo do executor associado ao ataque, se o mesmo foi capturado ou não, assim como motivo, se declarado.
- **Consequências:** Para cada registro, indica-se a quantidade de fatalidades, de feridos, danos a propriedade (e valor em dólares associado aos danos).
- **Vítimas:** Indica-se quantos foram sequestradas e-ou levadas contra sua vontade, assim como suas respectivas nacionalidades e grupos étnicos. Indica-se se houve taxa de resgate sendo paga e o valor associado a mesma. O número de resgatados-sobreviventes também é apresentado.
- **Fontes:** Indicam-se as fontes de informação sobre o incidente.

Os atributos são majoritariamente nominais. Além destes, há a presença de alguns atributos de escala, como ano, mês, dia e informações como latitude e longitude.

Alguns atributos apresentados são dados não estruturados, como as manchetes das notícias utilizadas como fonte para criar cada registro/instância.

Este trabalho apresenta um estudo utilizando visualizações de dados e também procura estabelecer um modelo capaz de identificar quando uma casualidade ocorre em um evento terrorista.

2 CONCEITOS TEÓRICOS

Nesta seção descrevemos os principais conceitos utilizados no desenvolvimento deste trabalho.

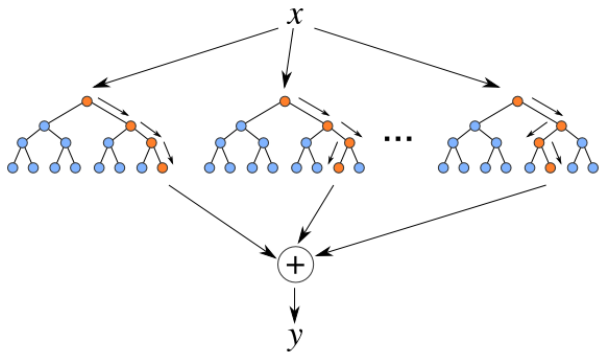

Figura 2: Ilustração do funcionamento do algoritmo random forest. X representa o conjunto de dados e Y a saída gerada.

2.1 Random Forest

Random forest, também denominado *random decision forest*, é uma técnica de aprendizagem de máquina utilizada para problemas de classificação, regressão e outros, proposta por Tin Kam Ho [7].

Esta técnica é um incremento dos algoritmos de árvores de decisão tradicionais, que utilizam um grafo árvore para mapear decisões e suas respectivas consequências, atribuindo dessa forma pesos para as variáveis presentes no modelo e para os valores que estas podem assumir. E então esses pesos serão utilizados para realizar as futuras classificações.

O algoritmo random forest constrói múltiplas árvores de decisão, cada uma com diferentes particularidades e então combina o resultado da classificação de todas as árvores para gerar o resultado final. Este processo está ilustrado na Figura 2.

2.2 Support Vector Machine

O *Support Vector Machine* (SVM), denominadas em português como máquinas de vetores de suporte, é um algoritmo de aprendizagem de máquina utilizado para classificações e regressões que, para dados n-dimensionais, procura um hiperplano que possa realizar a separação de classes objetivando maximizar a distância entre os pontos mais próximos em relação a cada uma das classe e, desta maneira, o SVM pode ser considerado como um classificador linear binário não probabilístico.

2.3 Cross Validation

Também conhecida como validação cruzada, o Cross Validation é uma técnica para avaliar a capacidade de modelos na generalização de um problema para um determinado conjunto de dados.

Para isso, o algoritmo utiliza um número de dobras K. Com a quantidade de dobras determinada, separa-se o conjunto em K partes. Em cada iteração, uma parte será utilizada como teste e as demais para o treino do modelo. Até que todas as combinações sejam testadas. O resultado da técnica é a média dos resultados obtidos em cada iteração.

Esse processo é ilustrado na Figura 3, para $K = 10$. O Cross Validation é utilizado para evitar resultados viciados (como os que ocorrem no fenômeno chamado *overfitting*).

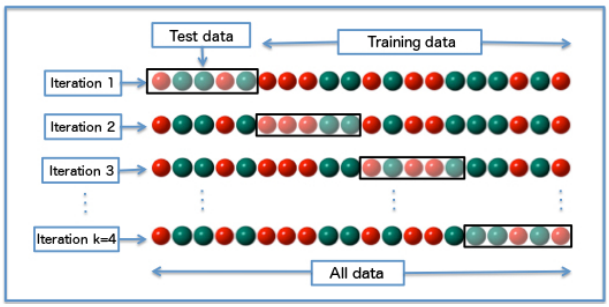

Figura 3: Ilustração do funcionamento do cross validation. K indica o número de folds (dobras) a serem utilizadas.

2.4 Grid Search

Grid Search é um método força-bruta que realiza a procura dos melhores parâmetros para o modelo. Seu funcionamento é simples: Experimenta todas as possibilidades, registrando o resultado obtido com cada um e então fornece aquele com resultado mais próximo do desejado.

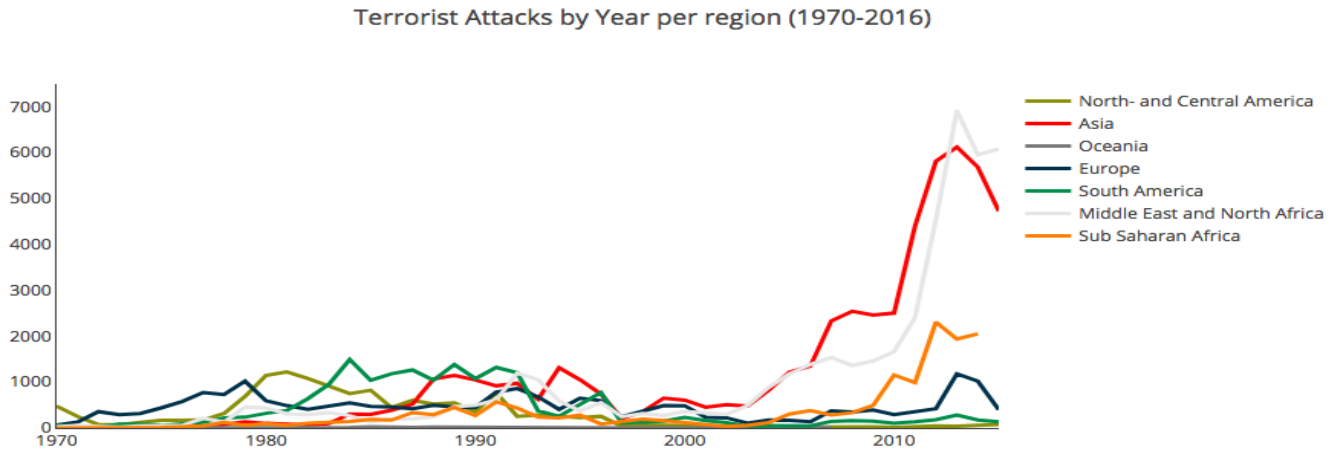

Figura 4: Número de eventos terroristas ao longo do tempo, agrupado por região.

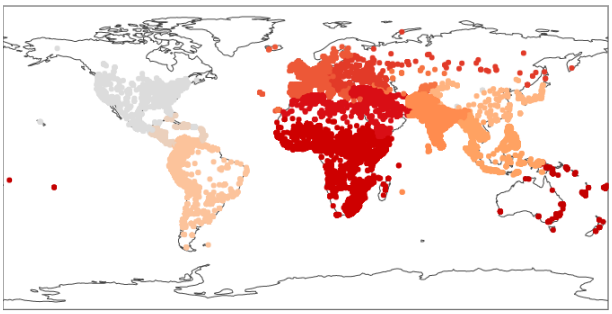

Figura 5: Distribuição geográfica dos eventos de terrorismo de 1970 a 2016.

3 TRABALHOS RELACIONADOS

Terrorismo é um assunto de projeção global e suas ocorrências estão relacionadas ao comportamento emocional das pessoas, estejam elas envolvidas ou não, por exemplo, a resiliência eomocional está fortemente correlacionada [3], mas também há estudos que analisam outros aspectos comportamentais [5] e emocionais [1].

Dada a relevância de tais eventos, é importante compreendê-los. Visualizações de dados podem ser utilizada para extrair conhecimento de dados de maneira efetiva [8] desde que otimizem os processos cognitivos reduzindo a desorganização das visualizações. [2]

As técnicas de visualização de dados são aplicadas na exploração de problemas de diversas naturezas, entre elas, muitos trabalhos são desenvolvidos para realizar a visualização de interações em redes sociais [4], trânsito [12], atividades criminosas [11] e em muitas outras áreas.

Dessa maneira, o presente trabalho pretende realizar um estudo através de visualizações e outras técnicas sobre a base de dados de terrorismo.

Este trabalho é inspirado principalmente por Wang [10], que apresenta uma análise investigativa da mesma base de dados através de visualizações, porém neste trabalho outras características da base serão exploradas e um modelo para classificar a ocorrência de causalidades nos registros de terorismo, além disso, este trabalho inclui eventos que ainda não haviam ocorrido quando o trabalho de Wang foi publicado.

4 METODOLOGIA

Esta seção apresenta os passos desenvolvidos neste trabalho, especificando a exploração visual conduzida assim como a manipulação da base de dados e a condução dos testes para avaliar o modelo construído.

4.1 Exploração Visual

Para iniciar os estudos sobre a base de dados em questão, iniciamos separando os atributos que são mais relevantes para a análise.

Consideramos que, primeiramente, a ocorrência dos eventos é muito relevante, ou seja, um estudo das características temporais e geográficas sobre as ocorrências dos eventos se faz necessária. A partir dessa observação, criamos a visualização exposta pela Figura 4.

A Figura 4 indica a progressão dos ataques terroristas ao longo do tempo, divindindo por região. Analisando o gráfico é visível que ocorreu nos últimos anos um grande aumento de ataques na região do oriente médio. Isso se deve a conflitos políticos e religiosos.

Outro fator importante que encontramos é o atributo *nkill*, que indica o número de pessoas que morreram em um evento terrorista. Para entender melhor o comportamento deste, criamos visualizações de dados. Primeiramente exploramos a relação do número de mortes com questões temporais e também com questões da característica do ataque.

Dessa maneira, foi criada a visualização exposta na Figura 6. Nela, apresentamos os eventos de terrorismo divindo-os por caracterizarem um ataque suicida (quando y=1) ou não (quando y=0). Além disso, o tamanho dos círculos representam o número de mortes associadas ao evento terrorista.

Observando essa figura é possível concluir que os ataques não suicidas são não apenas mais frequentes, mas também são associados a um número maior de mortes do que os ataques suicidas.

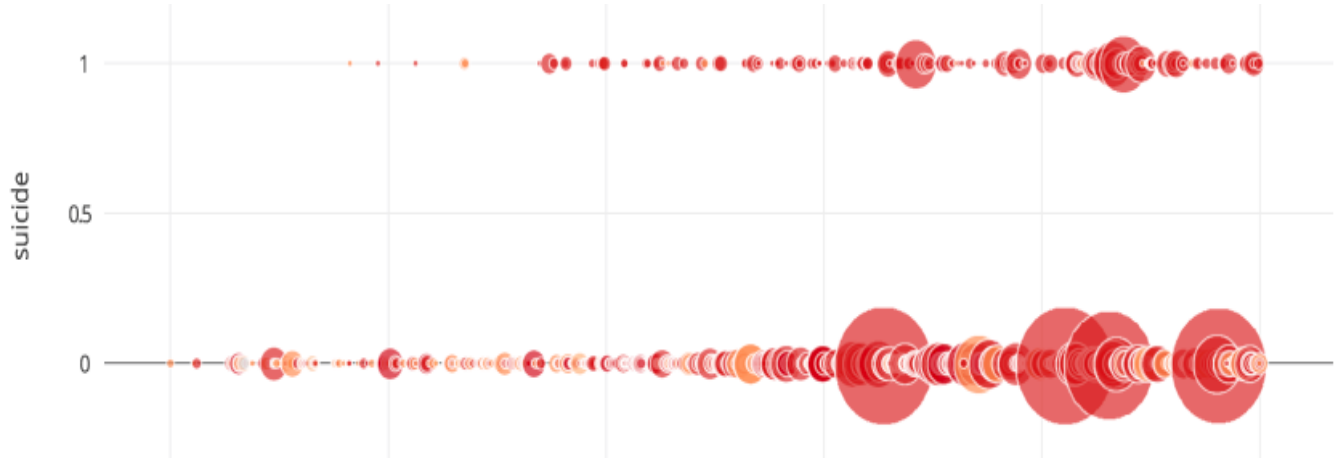

Figura 6: Ataques Terroristas Suicidas (1) em constrate com Ataques Terroristas não-Suicidas (0) ao longo do tempo (representado pelo eixo X). O Tamanho dos círculos indicam o número de mortes associados ao ataque terroristas.

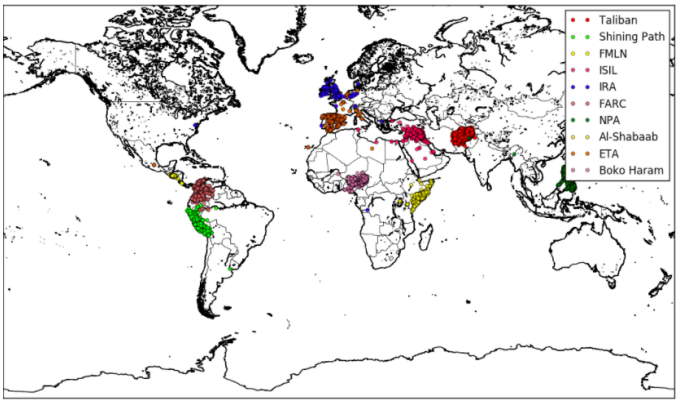

Figura 7: Distribuição geográfica dos registros de terrorismo dos 10 grupos terroristas com maior número de incidentes.

Além disso, verifica-se que a ocorrência de ataques suicidas cresceu consideravelmente a partir de 2014, e também que houve um aumento na amplitude dos ataques como um todo a partir da metade de 2014. Essa informação é condizente com o que foi dito e exposto na Figura 6.

Para compreender os eventos, acreditamos que a distribuição geográfica associada aos mesmos é muito relevante. Dessa forma, produzimos uma visualização para conseguir entender os locais de principal incidência de eventos terroristas.

A Figura 5 apresenta a distribuição geográfica, baseada nos dados de latitute e longitude fornecidos pelo conjunto de dados. Além disso, para facilitar a visualização, cada ponto, que representa um evento terrorista, apresenta uma cor de acordo com a região. É possível observar que a África e a Europa possuem eventos espalhados por todo seu território. Já na américa do sul, as atividades são concentradas em pontos específicos. Ainda sobre isso, é visível que os países orientais não possuem tantos incidentes registrados.

Uma pergunta que surge ao obsservar é: Quais são os grupos terroristas associados aos ataques? Para responder essa pergunta, separamos os dez grupos com maior quantidade de ataques registrados e disponibilizamos uma visualização sobre a distribuição geográfica de seus respectivos ataques.

Essa visualização é exposta na Figura 7. Analisando esta em contraste com a Figura 3, que contém todos os registros, é possível observar que apesar da Europa e da África possuírem ataques por todo seu território, os grupos responsáveis pelos ataques estão localizados em seu território atuam em regiões muito específicas. Dessa forma, é possível concluir que há uma grande variedade de grupos operando nessas regiões.

Além disso, Apesar dos registros de incidentes localizados na América do Norte, nenhum dos principais grupos de terrorismo está associado aos eventos.

A Figura 8 indica grande quantidade de ataques sendo conduzidos com bombas e explosões, seguido por ataques armados.

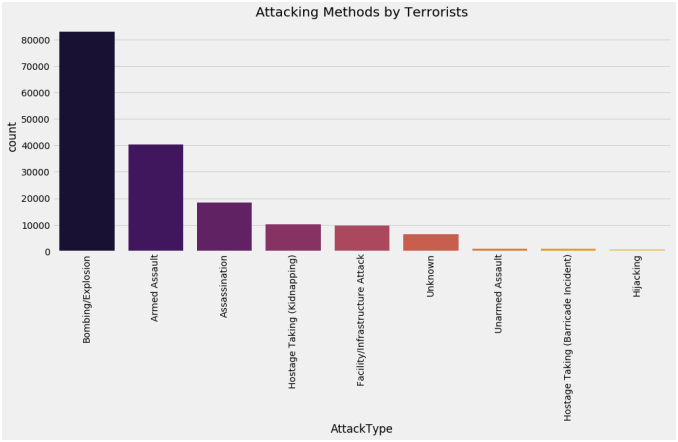

Figura 8: Número de eventos terroristas com cada tipo de armamento.

Tendo esse entendimento, também é válido analisar as consequências dos ataques e suas associações com as armas utilizadas.

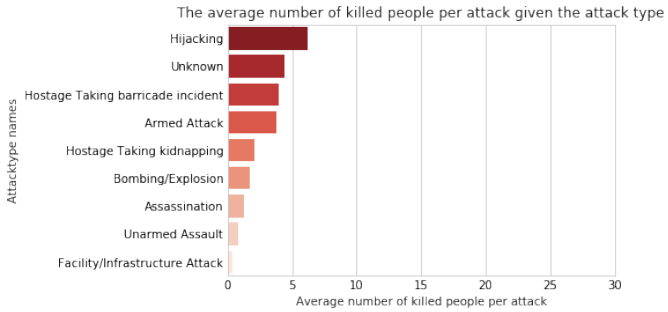

Figura 9: Média de mortes em eventos terroristas utilizando cada tipo diferente de armamento.

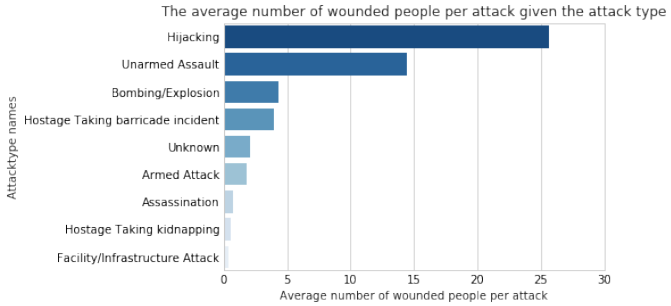

Figura 10: Média de feridos em eventos terroristas utilizando cada tipo diferente de armamento.

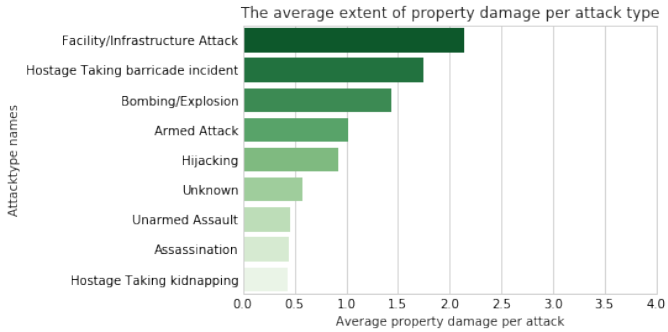

Figura 11: Média de danos financeiros em eventos terroristas utilizando cada tipo diferente de armamento.

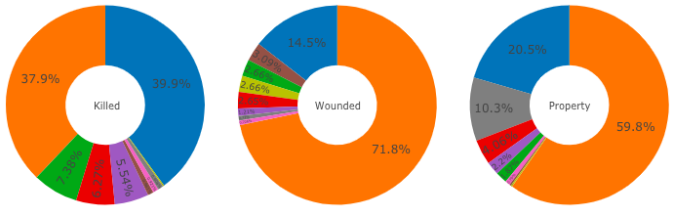

Figura 12: Porcentagem atribuída a cada tipo de armamento utilizado no total de mortes, feridos e dados a propriedades.

Com a análise das Figuras 9, 10 e 11, responsáveis pelos valores médios de cada evento terrorista, em contraste com a figura 12, que apresenta os valores percentuais totais para mortos, feridos e dados em propriedades ao longo dos anos, é possível ressaltar as seguintes conclusões:

- Bombardeamentos, apesar de possuir maior parcela no total de mortos, feridos e danos causados, em média, sequestros são mais danosos em quantidade de mortos e feridos;
- Ataques direcionados à infraestruturas diretamente também assumem maior valores médios de danos causados do que os bombardeamentos, e ao mesmo tempo pouca quantidade de mortos e feridos;
- Ataques armados estão associados com grande quantidade de mortes e poucos feridos, enquanto ataques desarmados indicam o contrário.

Para analisar a quantidade de terroristas envolvidos em ataques com diferentes tipos de armamento, foi gerada a visualização apresentada na Figura 13.

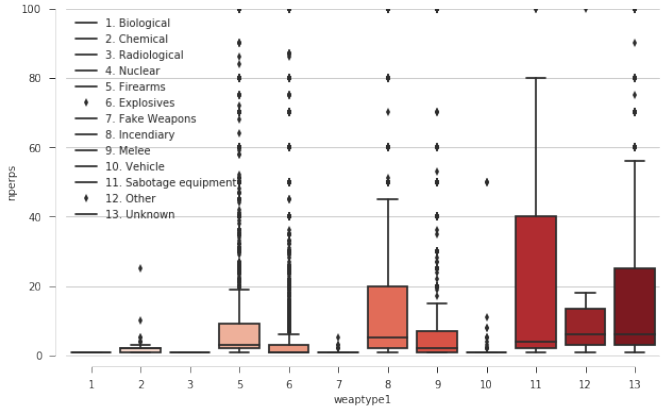

Figura 13: Média de terroristas envolvidos com cada ataque, divindo-os por tipo de armamento utilizado.

Analisando a figura chega-se às seguintes conclusões:

- Não contando os grupos "others"and "unknown", incêndios e sabotagem de equipamentos possuam taxa alto de desvio;
- Indepentemente do tipo de arma utilizada no ataque, a maioria dos ataques foram realizadas por menos de 20 terroristas.

4.2 Modelos de Classificação

Nesta seção o objeto é descrever os passos para a construção dos modelos de classificação. Uma tarefa preditiva é considerada: *Quantas causalidades ocorrem.* O objetivo dessa classificação é identificar a correlação dos atributos utilizados.

Para obter melhores resultados com o modelo de classificação, algumas alterações foram realizadas na base de dados. Nesta seção essas mudanças são descritas.

Para adicionar informações climáticas sobre os dias e locais dos eventos de terrorismo ocorridos, uma base de dados provida pela $http://www.weatherbase.com/$ foi utilizada para acrescentar os seguintes atributos:

- Temperatura;
- Pressão atmosférica;
- Cobertura por nuvens;
- Velocidade média do vento na região.

Além disso, foram removidas as instâncias da base com informações (atributos) faltantes, uma vez que este trabalho objetiva classificar as instâncias dada as informações completas.

As instâncias que indicam motivo desconhecido para os eventos foram excluídas, uma vez que não desejamos que atos que talvez não sejam terrorismo sejam analisados.

Os atributos *nkill* e *nwounds*, respectivamente número de mortes e número de feridos, possuem alguns valores como NaN, neste caso, os valores foram substituídos pela mediana do grupo. Após, estes atributos foram transformados em um único atributo: *causality*, que indica se houveram causalidades ou não.

5 RESULTADOS E DISCUSSÕES

As visualizações apresentadas forneceram uma ampla gama de conclusões que permitiram conclusões não observadas nos outros trabalhos relacionados a este, desta forma disponibilizando um melhor entendimento dos atributos da base e de suas respectivas correlações com outros atributos.

x	HC	SC
HC	0.71	0.29
SC	0.21	0.79

Tabela 1: Matriz de confusão normalizada excluindo clima. HC: houve casualidade, SC: sem casualidade.

x	HC	SC
HC	0.78	0.22
SC	0.1	0.9

Tabela 2: Matriz de confusão normalizada considerando clima. HC: houve casualidade, SC: sem casualidade.

A Tabela 1 apresenta a matriz de confusão para a predição de causalidades utilizando a base sem as variáveis sobre condições climáticas, enquanto a Tabela 3 apresenta a matriz de confusão para a base com as variáveis climáticas.

Essas tabelas indicam maior *sensibilidade* e *precisão*, para os dois grupos , na classificação utilizando dados climáticos. Além disso, a validação cruzada com 10 *folds* (dobras). Para a base sem clima, o Random Forest obteve 0.7383205317 de precisão, já para a base com clima, obteve 0.845681042507.

Também utilizamos o algoritmo *Support Vector Machine* (SVM) para experimentar outros resultados. Sua performance, avaliada pela validação cruzada com 10 *folds*, foi superior para a classificação da base sem fatores climáticos, apresentando precisão de 0.767643223, porém para a classificação com dados climáticos, sua performance foi inferior, apresentando precisão de 0.7923142234.

Em todos os testes foi utilizado *Grid Search* para a escolha dos parâmetros otimizados.

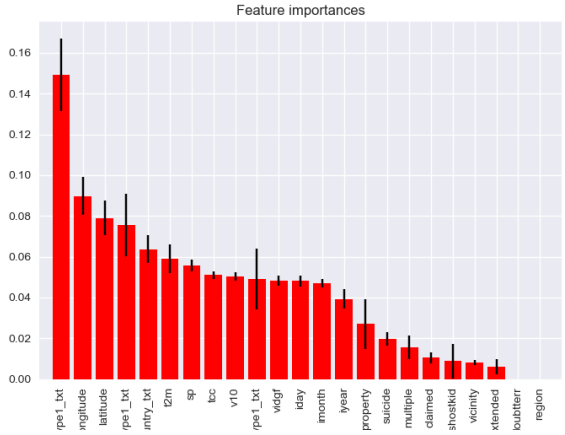

Figura 14: A importância de cada feature no processo classificatório do algoritmo Random Trees.

A Figura 14, apresenta a importância dos atributos utilizados no processo de classificação do algoritmo Random Tree. A maior relevância foi atribuída para o tipo de ataque realizado, seguido pela localização e então aos fatores climáticos.

Por fim, apresenta-se uma tabela com a comparação de performance dos algoritmos.

Algoritmo Utilizado	Precisão Obtida
Regressão Logística	0.77
Perceptron	0.53
Gradient Descendente	0.54
KNeighbors	0.70
Gaussian Naive-Bayes	0.74

Tabela 3: Precisão da validação cruzada com 10 folds para diferentes algoritmos de aprendizagem.

A tabela indica que os outros algoritmos obtiveram performance pior ao realizar a classificação do conjunto de dados com os atributos climáticos.

6 CONSIDERAÇÕES FINAIS

As visualizações produzidas neste trabalho possibilitaram uma melhor compreensão da natureza da base de dados e da correlação de seus atributos. É necessário considerar que nem todos os atributos foram explorados e, portanto, trabalhos futuros podem considerar atributos ainda não tão bem explorados, assim como podem experimentar com a comparação dos eventos com outras informações que não apenas clima.

Além disso, neste trabalho técnicas de redução de dimensionalidade não foram explorados, o que pode ser realizado tanto para prover novas visualizações, como para explorar novas possibilidades nos modelos preditivos.

Por fim, os modelos classificatórios para indicar se houve ou não casualidades baseada em atributos foi validado e seus resultados indicam que há grande correlação entre o clima e a possibilidade de ocorrer casualidades. Futuros trabalhos podem considerar mais aspectos, como fatores econômicos, educação, mudanças políticas, eventos e outros fatores que podem estar associados a ocorrência de casualidades.

REFERÊNCIAS

[1] Adriana Camacho. 2008. Stress and birth weight: evidence from terrorist attacks. *The American Economic Review* 98, 2 (2008), 511–515.

[2] Min Chen and Amos Golan. 2016. What may visualization processes optimize? *IEEE transactions on visualization and computer graphics* 22, 12 (2016), 2619–2632.

[3] Barbara L Fredrickson, Michele M Tugade, Christian E Waugh, and Gregory R Larkin. 2003. What good are positive emotions in crisis? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of personality and social psychology* 84, 2 (2003), 365.

[4] Linton C Freeman. 2000. Visualizing social networks. *Journal of social structure* 1, 1 (2000), 4.

[5] Gerd Gigerenzer. 2006. Out of the frying pan into the fire: Behavioral reactions to terrorist attacks. *Risk analysis* 26, 2 (2006), 347–351.

[6] Harry Henderson and Harry Henderson. 2001. *Global Terrorism: The complete reference guide.* Checkmark books New York.

[7] Tin Kam Ho. 1995. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, Vol. 1. IEEE, 278–282.

[8] Christopher D Hundhausen, Sarah A Douglas, and John T Stasko. 2002. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages & Computing* 13, 3 (2002), 259–290.

[9] START. 2017. Global Terrorism Database. https://www.kaggle.com/START-UMD/gtd. (2017). [Online; accessed 25-November-2017].

[10] Xiaoyu Wang, Erin Miller, Kathleen Smarick, William Ribarsky, and Remco Chang. 2008. Investigative visual analysis of global terrorism. In *Computer Graphics Forum*, Vol. 27. Wiley Online Library, 919–926.

[11] Jennifer Xu and Hsinchun Chen. 2005. Criminal network analysis and visualization. *Commun. ACM* 48, 6 (2005), 100–107.

[12] Hao Zhang, Maoyuan Sun, Danfeng Daphne Yao, and Chris North. 2015. Visualizing traffic causality for analyzing network anomalies. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*. ACM, 37–42.

7 APÊNDICE

.

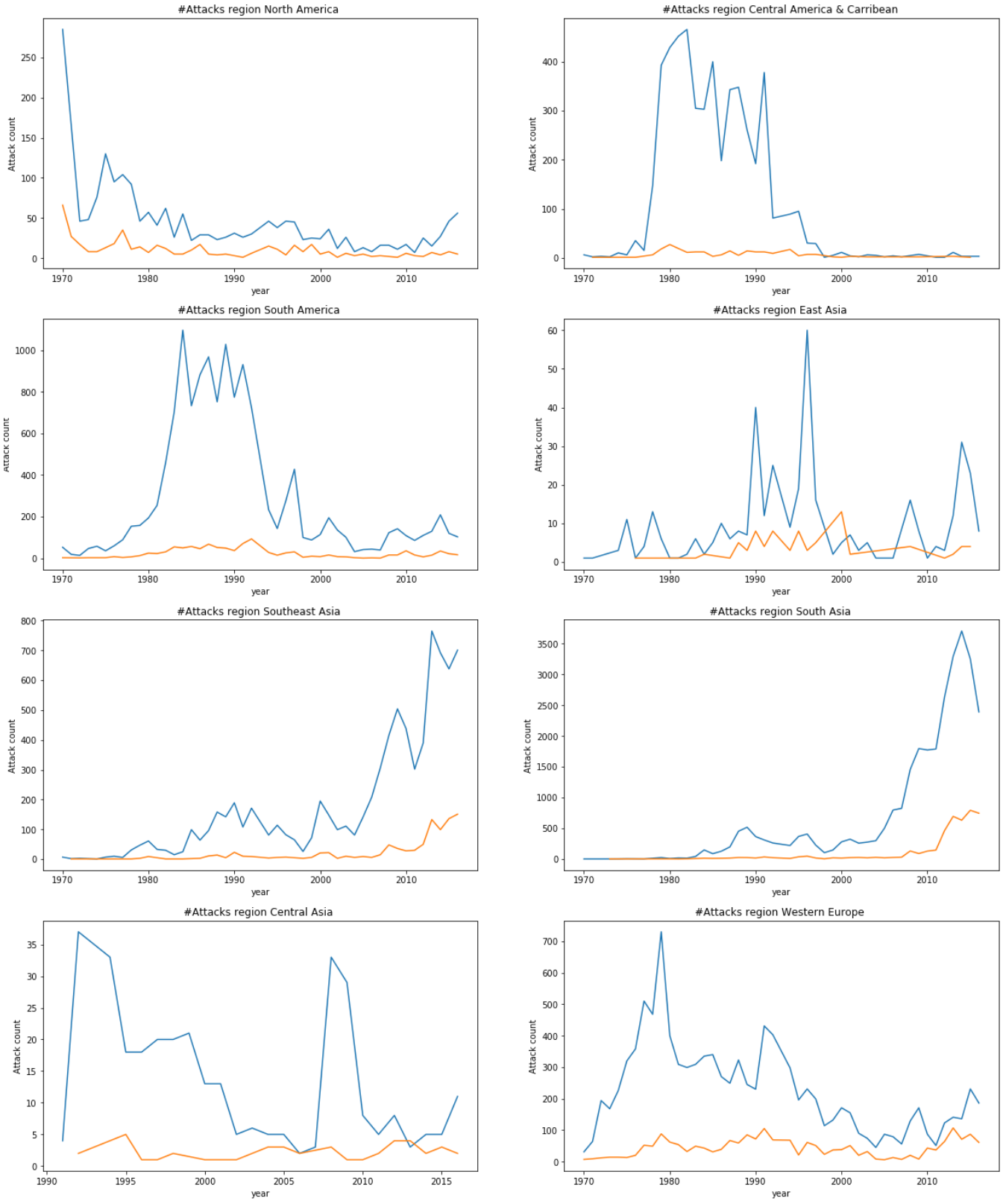

Figura 15: Ataques terroristas por região. Azul representa os ataques bem sucedidos e o laranja o contrário.

.

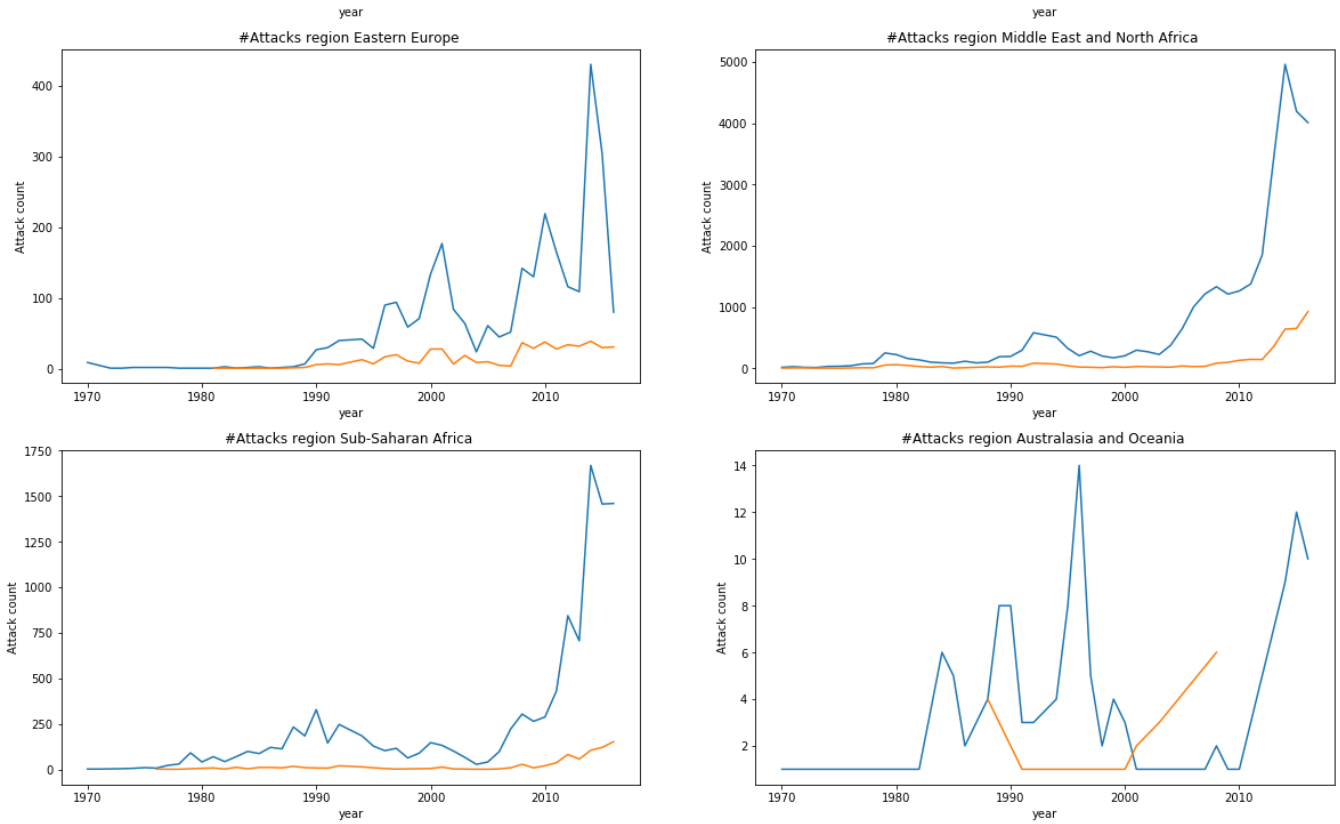

Figura 16: (Continuação) Ataques terroristas por região. Azul representa os ataques bem sucedidos e o laranja o contrário.