Universite Jean Monnet - MLDM

Data Mining

A Comprehensive Analysis on Terrorist Attacks

Author:
Andrei Mardale

Contents

1 Problem Understanding

Terrorism has become one of the most important issues of security in the last couple of years [9]. More and more money is spent on improving security and a lot of governments have taken special measures in order counter this phenomenon. Be it groups of hyper nationalists or Taliban groups, the fear and terror inspired by their attacks made terrorism one of the most important threats in today's society.

In this context, this project aims to discover interesting correlations between terrorism level and different socioeconomic factors, such as the military budgets, immigration rate, population density and size.

2 Data Understanding

2.1 Dataset

For this study, I used five different datasets, which have been merged in order to select the features that would yield good correlation with the number of terrorist attacks.

The main dataset used is the one containing information about terrorist attacks, namely Global Terrorism Database [3]. It is an open source database, which contains more than 180 000 records about terrorist events from 1970 to 2017. There are around 150 variables describing each event, thus it offers sufficient information to perform a thorough analysis on this topic. This dataset was created and is maintained by researchers from University of Maryland, USA. I found this set on the internet and it is freely accessible on the website mentioned above.

The second source used is the Military Expenditure database. This is provided by the World Bank and is open source and accessible at [4]. It contains information about the amount of money spent by each country on military issues from 1960 to 2017.

The third dataset used is the Refugee population by country of asylum, offered by United Nations High Commissioner for Refugees. It is also free and available at [7]. It contains records about the number of immigrants in each country from 1990 to 2017.

The fourth dataset is about the World Population and is provided by the United Nations. It contains records about the population of every country, by year. It is available at [8]. It has roughly 300 000 entries and it also contains future predictions about the variation of the population number for every country.

The last dataset which has been used in this study is the Density of Population. It is offered by the Food and Agriculture Organization and World Bank and is available at [5]. It has a similar format as the one about the immigrants, being offered by the same World Bank website. It contains entries for the density of each country from 1961 to 2017.

2.2 Data Exploration

In this subsection, I will present a part of the exploratory analysis which has been performed on the GTD. The first operation that was performed was measuring the number of missing values for each variable. Although the variables regarding the ransoms are interesting, they can't be used because there is a high number of missing entries. As the dataset contains a lot of different variables, I select at the beginning the features that look important at a first glance.

As depicted in Figure 1 (a), the trend of the attacks seem to be ascending. Most of the attacks are bombing attacks, roughly 80.000. Second place is Armed Assault approximately 40.000. As can be observed from Figure 1 (b), by far the most dangerous area is the Middle East and north Africa. Then, there is South Asia and South America. Europe tends to be a safer place, having approximately a thirds of the number of attacks from Middle East.

Perhaps not surprisingly, the most dangerous country in terms of terrorism is Irak. As we will see later, it is the homeland of dangerous terrorist groups like ISIL. Then, we have its neighbours, Pakistan and Afghanistan. The most targeted country from EU is UK, followed by Turkey. The top most dangerous cities are Baghdad, Karachi and Lima. On the 10th position we can find Athena, a famous destination for tourists [2].

Because of the high number of available variables, a process of feature selection was performed. The main criterion was not only having a low percentage of NAs but also having an important role in the process of future prediction of attacks. Thus the main features of this dataset that were used are: **year, month, day, country, region, city, latitude, longitude, attack type, target type, group name, weapon type**.

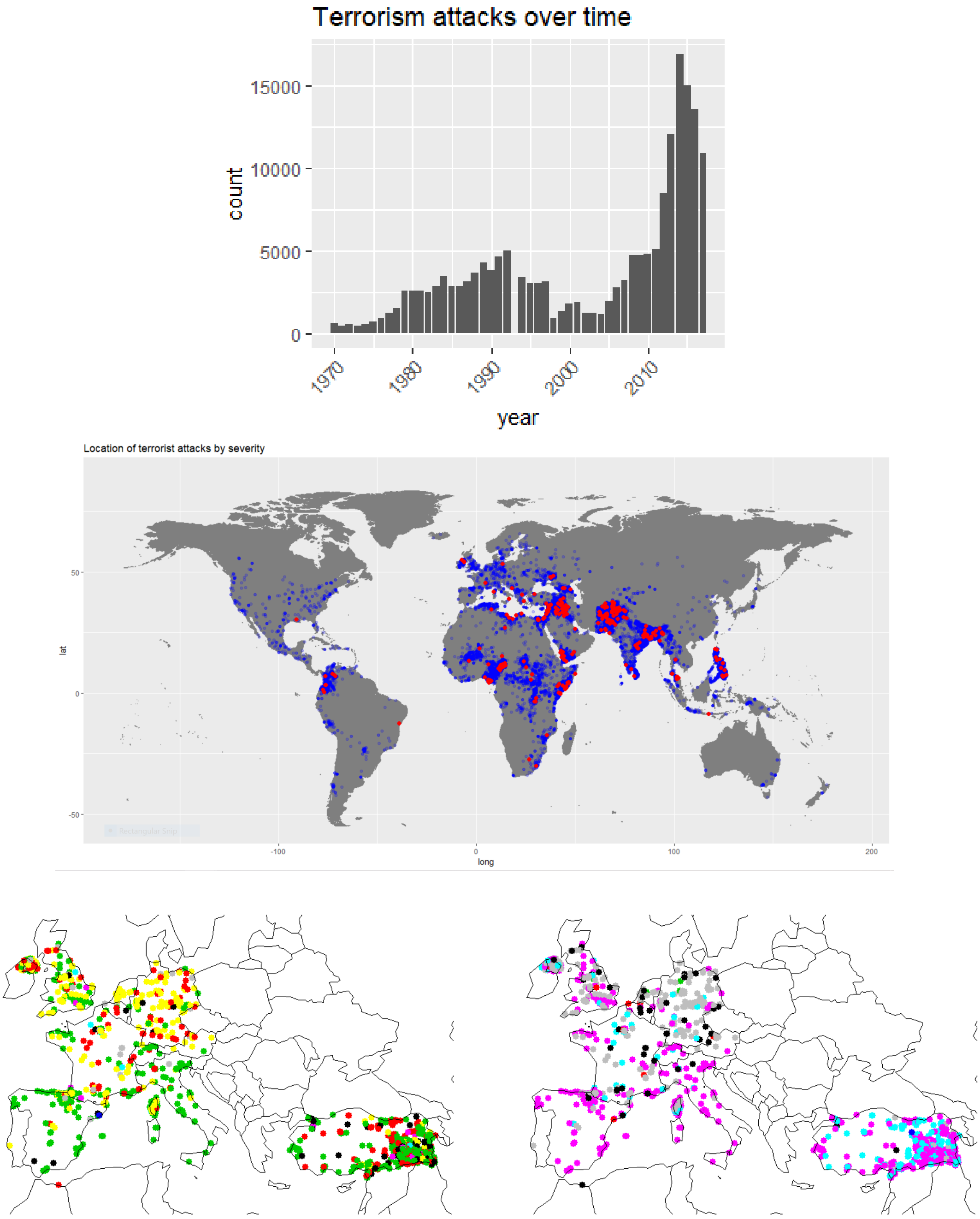

Figure 1: (a) Trends in terrorism attacks number. (b) Attacks by severity. (c) Attacks by type. (bombing, assassination, hijacking etc.) (d) Attacks by weapon. (firearms, explosives, incendiary etc.)

3 Data Preparation

As far as the data preparation is concerned, the first step was to merge the five datasets. Basically, I obtained a brand new dataset, counting the number of terrorist attacks in each country in each year, appending also the number of immigrants in that country in that year, the population, density and the military budget that country allocated. All these information were used for trying to predict the number of attacks in future, based in predictable values for the population, density, budgets and immigrants. Thus, the merge was done by performing inner joins based on the year.

The terrorist dataset is very large (180 000 entries) so I could afford to just remove the missing values. After having merged the data, I normalized the data. Tests have been performed both with and without performing outlier detection and removal and it turned out that outliers did not affect my dataset so much. **The study was restricted to Northern America and Europe.**

4 Modeling

For the modeling part, I applied four different Machine Learning approaches. The first one is highly related to what we studied this semester in Data Mining, Association Rules Mining. In the second approach, I am using different classification techniques to predict the severity of an attack. I created a new categorical feature that describes the severity of an attack. By using the data from the merged datasets, I try to predict the severity level of an upcoming attack. The third approach is based on Regression techniques to predict the number of attacks in future. Finally, I am doing clustering analysis on the data in order to try to find underlying correlation of the data.

4.1 Association Rules Mining

By using Association rules mining, I am trying to obtain interesting correlations between the attacks that have occurred so far. The results turn out to be quite satisfying and helpful in trying to prevent some certain kind of attacks. The features used for this analysis are: **"country"**, **"region"**, **"attacktype1"**, **"targtype1"**, **"groupname"**, **"targsubtype1"**, **"weaptype1"**, **"nkill"**, **"nwound"**. The first step is to convert the number of kills and wounded persons in categorical features. In the following subsections I will present some results obtained when the right hand side of the association rule was set to different terrorist groups. For all searches, the parameters used for **Apriori** algorithm are **minimum support=0.01**, **minimum confidence=0.1**, **minlen=1**, **maxlen=5**.

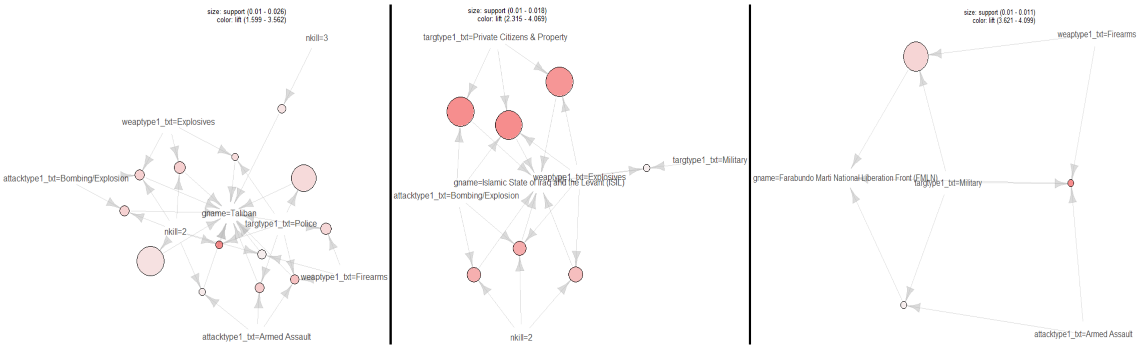
size: support (0.01 - 0.026)
color: lift (1.599 - 3.562)
nkill=3
weaptype1_txt=Explosives
attacktype1_txt=Bombing/Explosion
gname=Taliban
targtype1_txt=Police
nkill=2
weaptype1_txt=Firearms
attacktype1_txt=Armed Assault
size: support (0.01 - 0.018)
color: lift (2.315 - 4.069)
targtype1_txt=Private Citizens & Property
weaptype1_txt=Explosives
gname=Islamic State of Iraq and the Levant (ISIL)
targtype1_txt=Military
attacktype1_txt=Bombing/Explosion
nkill=2
size: support (0.01 - 0.011)
color: lift (3.621 - 4.099)
weaptype1_txt=Firearms
gname=Farabundo Marti National Liberation Front (FMLN)
targtype1_txt=Military
attacktype1_txt=Armed Assault

Figure 2: Association rules obtained for: (a) Taliban, (b) ISIL, (c) FMNL

We can observe from Figure 2 (a) that Talibans tend to attack Police forces very often, with explosives and firearms. However, most of the time, the number of killed persons is around 2. Figure 2 (b) shows that ISIS prefers to attack private citizens, with explosives, number of casualties being often around 2. Lastly, the Farabundo Marti National Liberation Front prefers to attack military troops, with firearms in armed assaults, as shown in Figure 2 (c).

4.2 Classification

In this subsection, I focused on trying to predict how dangerous can an attack be. Some other changes were made to the dataset. I summed all attacks that took place in a country, per year and created an attacks number feature. Then, I built a new feature, which tries to summarize how dangerous a country is: attacks level. If there are less than 4 attacks, the value is 1, otherwise is 2. The final dataset for classification task has 3 features: immigrants, military budget, population density and a label that needs to be predicted, attacks level. There are 532 entries in the dataset, after these preprocessing steps.

The same methodology has been applied on all Classifiers: data was split in training (80%) and test (20%). Then, on the training set, I used a k-Cross Validation technique, with k = 5. After that, the best parameters were chosen according to the previous step. Finally, a testing step is performed to measure the test accuracy.

4.3 Regression

In the regression setup, I will use a slightly different set of features. I have number of attacks, military budget, emigrants percentage and total population of the world. All these values are represented per year, globally.

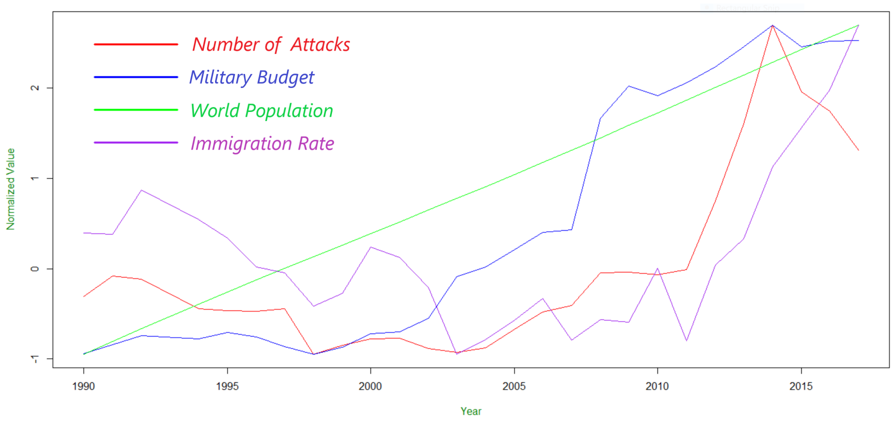

Figure 3: Correlation trends between terrorist attacks and analyzed features.

We can observe in Figure 3 that the number of terrorist attacks is highly correlated with the military budget, world population and immigration rate. This is a cue that I can do a regression analysis using this features. The Pearson correlation values of terrorist attacks with regard to Military Budget, Immigrants and World Population are **0.8058181, 0.7009036, respectively 0.6878820**.

4.4 Clustering

In this task, I tried to get some intuition if there is a link between the geographical position of the attacks and the different kind of characteristics an attack has. The data considered as been filtered by the following filters: the year of the attack should be greater than 1990. The countries selected are mostly from Western Europe plus an outlier, Turkey. The group names were replaced by labels, so that I could consider who committed the attack, in the clustering analysis. The algorithm used for clustering was **K-Means** with different values for the parameter k. The results can be obtained in the following set of figures 4.

For a small size of $\mathbf{k = 3}$, the main centers are the area around UK, Turkey and Central part of Europe. This figure does not give a lot of intuition, as perhaps, the groups attacking Turkey have difficulties in transporting man and weapons straight to the UK. For $\mathbf{k = 4}$, the main two centers from Turkey and UK still remain, but now the attacks in the last central group are divided into two groups. The part related to Germany and the one related to Spain. Meanwhile, France remains a mix of the two groups. It is interesting though to notice the small green island from Switzerland which belongs to the Turkish group. When we set **k = 5**, the great Turkish group is divided into two parts. The eastern and the western part. Checking the minorities map of Turkey [1], we can observe that in the east there are Turks, while in west there are indo-european populations. The same indo-european populations can be observed in Spain, too [6]. Going for **k=6**, the same division in Turkey, while UK remains a cluster. Now Spain has its own cluster, while central part of Western Europe has a mixture of points. This is caused by the relaxed regulations regarding human travelling between these countries. People can freely travel from a country to another.

5 Evaluation

Classification

In this subsection, the results obtained with different classifiers have been evaluated in terms of training time accuracy an test time accuracy.

Classifier	Hyper-Parameters	Train Acc.(%)	Test Acc.(%)
SVM (rbf)	σ = 40.019, C = 1	86.59	78.50
SVM (poly)	degr = 3, C = 1	*60*	*60.74*
DeepBoost	iterations = 50, depth = 3, beta = 0.0039, λ = 0.063	**90.35**	**79.43**
Random Forest	mtry=2	83.76	78.51
Decision Trees (CART)	cp = 0.01538462	83.29	78.5
k-NN	k = 5	88	78.7
Neural Network	size = 3, decay = 1e-04	81.64	73.81

Table 1: Results of different models for the classification problem.

The results obtained during the classification process are displayed in Table 1. We can observe that the best results were obtained using the DeepBoost model. On the other hand, the worst results are obtained with a SVM using polynomial kernel. All in all, most of the classifiers trained seem to overfit the training set.

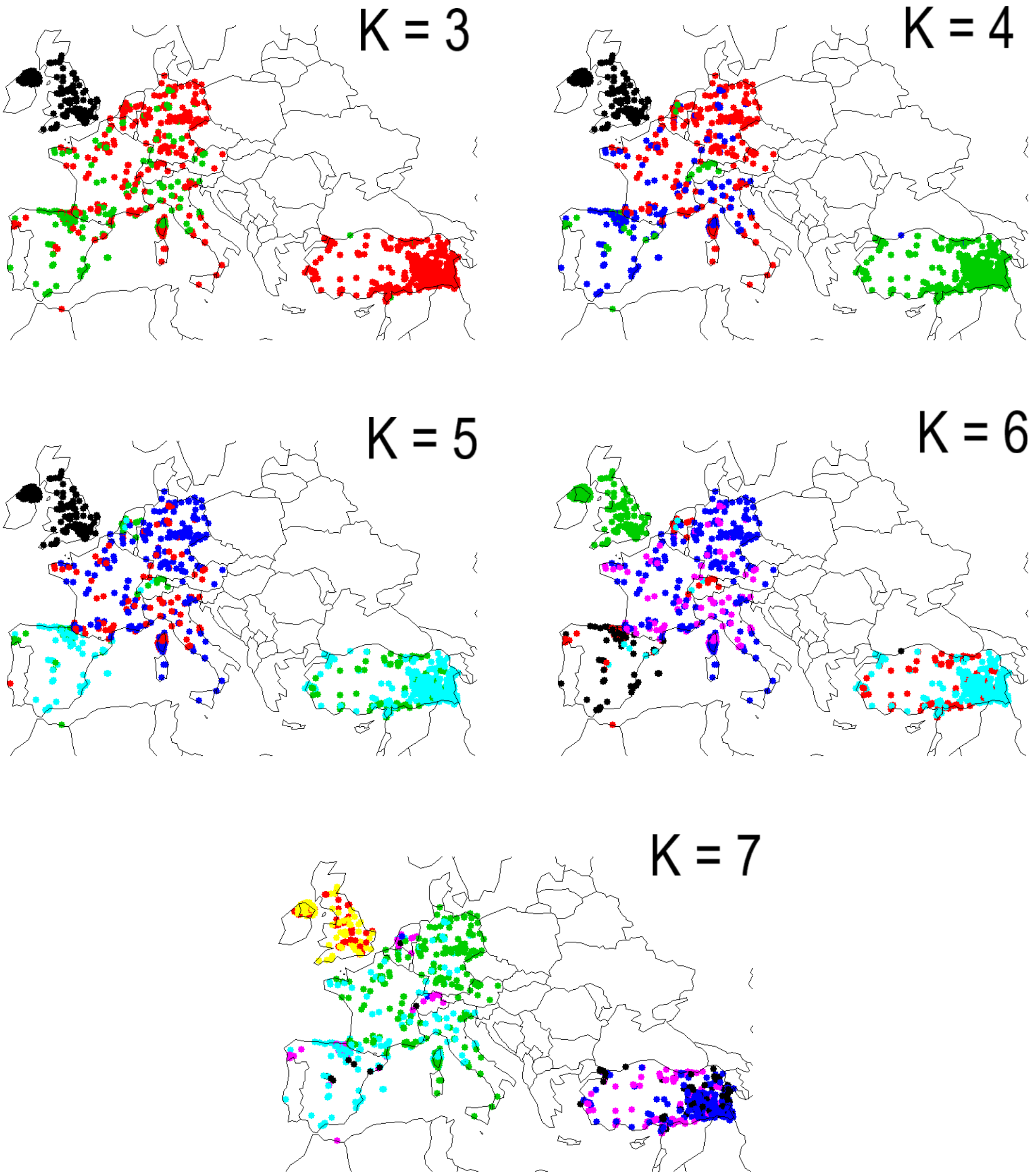

Figure 4: Attacks in Western Europe and Turkey, grouped in clusters of different sizes.

Regression

For the quality assessment of the regression models trained, two measures have been used: the mean squared error and the R^2 score. On one hand, the mean squared error shows how well has the model been able to learn the underlying structure of the data. The best model is the Polynomial regression, which has a training Mean Squared Error of 0.38 and a test time error of 0.1. At the same time the variance explained by the population, military budget and immigration rate is very good, denoted by the good R^2 score.

Model	**Train MSE**	**Train R^2**	**Test MSE**	**Test R^2**
Linear Regression (attk∼pop)	0.78	0.72	0.59	0.95
Linear Regression (attk∼pop + popˆ2)	0.46	0.91	0.20	0.03
Linear Regression (attk∼pop + mil_budg+immig)	0.38	0.85	0.1	0.8
SVM (kernel = linear, C = 1)	0.5	0.94	0.11	0.76
Random Forest (mtry = 3)	0.33	0.96	0.2	0.94
Bayesian Regularzied NN (neurons = 2)	0.38	0.9	0.29	0.02
Bagged CART	0.55	0.83	0.29	0.79
Monotone MLP NN (hidden1 = 5 and n.ensemble = 1)	0.42	0.8	0.56	0.87
Neural Network (L1 = 5, L2 = 0 and L3 = 0)	0.43	0.87	0.35	0.89
Penalized Linear Regression ($\lambda1 = 1$ and $\lambda2 = 4$)	0.44	0.88	0.18	0.69
Supervised PCA (thrs = 0.1 and components = 2)	0.56	0.8	0.18	0.6

Table 2: Evaluation of different regression models in terms of Minimum Squared Error and R squared values.

k	**WSS(%)**	**BSS(%)**	**BH**	**CH**
3	17.04	82.96	5.68	10256.62
4	10.87	89.13	2.99	10352.22
5	8.52	91.48	1.7	11294.62
6	7.3	92.7	1.22	10687.95
7	5.43	94.57	0.78	12208.34

Table 3: Quality measures of the different values for K.

Clustering

For the K-Means algorithm, I measured the **WSS** - Within Sum of Squares, **BSS** - Between Sum of Squares, **CH** - Calinski and Harabasz clustering quality index and **BH** - Ball and Hall clustering quality index. It is difficult for a clustering algorithm to say if the result is good or not, as the values from Table 3 demonstrate.

The WSS value decreases as the number of clusters increases. At the same time, the BSS values increases with the size of k. The Ball - Hall index represents the mean of all cluster's mean dispersion. For the Calinski and Harabasz index, the higher the value, the better.

Bibliography

[1] Ethnolinguistic map of turkey. https://en.wikipedia.org/wiki/Minorities_in_Turkey#/media/File:Ethnolinguistic_map_of_Turkey.jpg. Accessed: 2019-03-13.

[2] Global terrorism analysis. https://www.kaggle.com/laurenstc/global-terrorism-analysis/data. Accessed: 2019-02-19.

[3] Global terrorism database. https://www.start.umd.edu/gtd. Accessed: 2019-03-12.

[4] Military expenditure. https://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS. Accessed: 2019-03-12.

[5] Population density (people per sq. km of land area). https://data.worldbank.org/indicator/en.pop.dnst. Accessed: 2019-03-12.

[6] Preroman populations in spain. http://3.bp.blogspot.com/_S6cO9MTuZBo/S9AjC2IxP0I/AAAAAAAAABM/-uCzCRgSSao/s1600/mapa-de-los-pueblos-prerromanos.gif. Accessed: 2019-03-13.

[7] Refugee population by country or territory of asylum. https://data.worldbank.org/indicator/SM.POP.REFG. Accessed: 2019-03-12.

[8] United nations - world population. https://population.un.org/wpp/Download/Standard/Population/. Accessed: 2019-03-12.

[9] Zhang, X., Jin, M., Fu, J., Hao, M., Yu, C., and Xie, X. On the risk assessment of terrorist attacks coupled with multi-source factors. *ISPRS International Journal of Geo-Information 7*, 9 (2018), 354.