Global Risk Assessment of Earthquakes and Effects

Machine Learning 1

Esin Isik, Sabrina Rigo, Andrea Viczian

2023-06-09

Table of contents

1. Introduction

Earthquakes are one of the most destructive natural disasters that can strike without warning, causing extensive damage to infrastructure, loss of life, and massive economic losses. While we cannot prevent earthquakes from occurring, the ability to accurately predict when and where they might occur could save countless lives and minimize the damage caused.

Therefore, the aim with this report is to contribute the significant earthquake prediction which enables to provide advanced warning of potentially catastrophic seismic events, allowing governments and communities to prepare and take necessary measures to minimize the impact of such events.

Remark:
In this report the focus is on the analysis of the primary earthquake effects, the **Total Earthquakes and Secondary Effects** section will not be discussed in further detail. During the preliminary analysis, it has been discovered that the variables listed this section do have a considerably high amount of missing values which may not produce reliable outcomes. For completeness, the information has been added in the last drop down button below.

1.1.Description of our data

Data source: https://www.ngdc.noaa.gov/hazel/view/hazards/earthquake/search (https://www.ngdc.noaa.gov/hazel/view/hazards/earthquake/search)

The Significant Earthquake Database contains information on destructive earthquakes from 2150 B.C. to the present that meet at least one of the following criteria: Moderate damage (approximately $1 million or more), 10 or more deaths, Magnitude 7.5 or greater, Modified Mercalli Intensity X or greater, or the earthquake generated a tsunami. The database can also be displayed and extracted with the Natural Hazards Interactive Map.

A short summary of the used main variables are listed below. The available information about primary and secondary deaths and damages in total numbers have been added to the dataset. The “description” field contains this data in categorical form.

Earthquake Magnitude [Mag]

Modified Mercalli Intensity Scale [MMI.Int]

Focal Depth (km) [Focal.Depth..km.]

Region

Hazard Association

Earthquake Effects

Total Earthquake and Secondary Effects

1.2.Getting the data

Setting the working directory to load the tab-separated file to R:

```
#setwd("~/FS23_ML1/Project Data/Git_repository/ML01")
eqdata <- read.csv("significant-earthquakes-database-country-region.tsv", header = TRUE, sep = "\t")
```

Having a first look at the data provided:

str data (interactive dropdown button)

1.3 Preparing the data

For improved analysis, the below named cleaning processes were performed: **Cleaning Process Exclude data dated before 1900**

- Excluding data before 1900 due to following reasons:
  - data available mostly from historical records therefore less reliable
  - the measurement quality is not reliable based on less developed methods. The modern seismometer wasn't invented until the mid-18 hundreds. Therefore, it can be suggested that these modern technologies were not widely used around the world until the 19-hundreds.
  - a lot of missing values are present from these records

Exclude data where magnitude is not available

- the variable of magnitude has key importance in this analysis, the records where it is missing are therefore too unreliable to consider.

Exclude data where the number of deaths is not available

- the death count is also a key variable within the scope of this analysis. Since there is no information available whether the NA can be treated as 0 or as true unknowns, the decision has been taken to exclude records without such value.

Enrichment Process:

Add magnitude column without decimals

- At certain stages of the analysis, it can be beneficial to consider the number of magnitude as counts.

Add two columns frequency of occurrence country and region

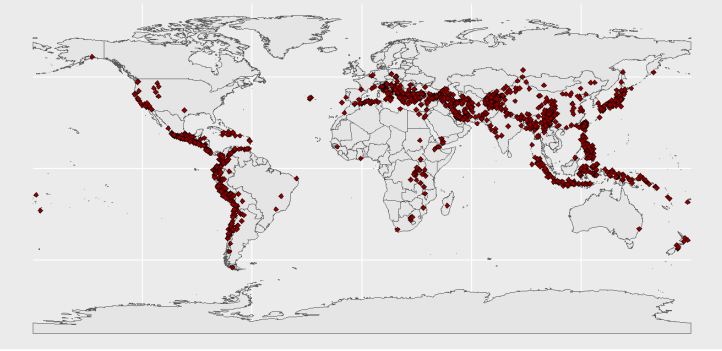
- For further analysis, two additional calculated columns have been added to the data. The values contain a number counting the occurrence of the respective country and region in the data set.

str cleaned data (interactive dropdown button)

After cleaning the dataset, **1469 observations of 44 variables** are remaining. The transformed data set represents the base on which the following analysis is performed.

2. Graphical Analysis

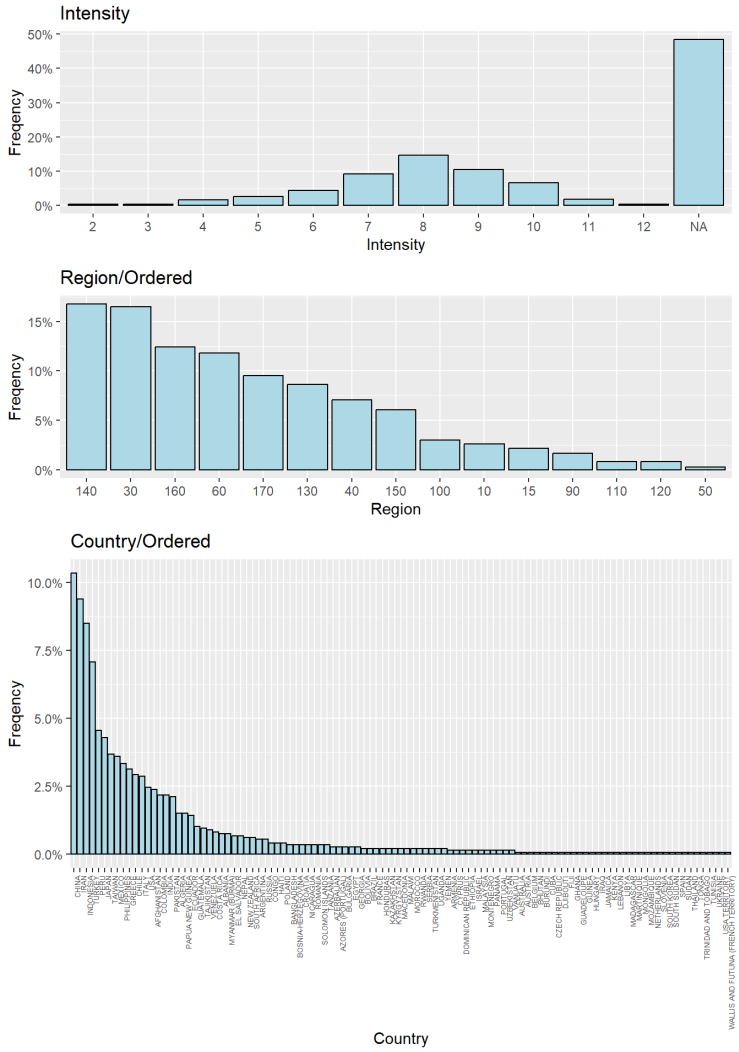
2.1 Map of earthquakes



To get a first overview about the earthquakes, a world map is created. For that, the longitude and latitude of all incidences are used. All the red squares indicate the location of an earthquake. It becomes apparent that many earthquakes happen at the border of a tectonic plate. Interesting to see is that there were a lot of earthquakes measured at the west coast of North and South America but not that many at the eastern locations. On the east Asian continent, no such behavior can be observed.

2.2 Preliminary Analysis of categorical variables

In this section, a visual overview and preliminary analysis of the categorical variables contained in this data set is provided.

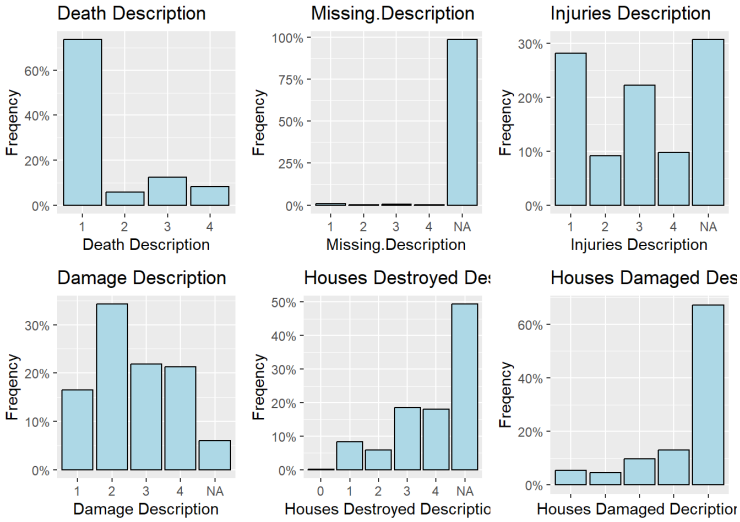


Looking at the categorical variables:
- Intensity has also a rather normal distribution, however a large number of NA values can be seen
- The region and country variables are sorted in decreasing order

Region: The below regions have the highest number of occurrences in our database, which means the highest number of earthquakes registered:
- 140: Middle East
- 30: East Asia
- 160: South America
- 60: S. and SE. Asia and Indian Ocean
- 170: Central and South Pacific
- 130: Southern Europe
- 40: Central Asia and Caucasus
- 150: 150 - North America and Hawaii

As it can be seen, the most frequently mentioned countries are also from the regions mentioned in the above regions.

Based on these charts and the World map visualizing the data based on latitude and longitude, it can be assumed that location has a high influence on the likelihood of an earthquake happening.



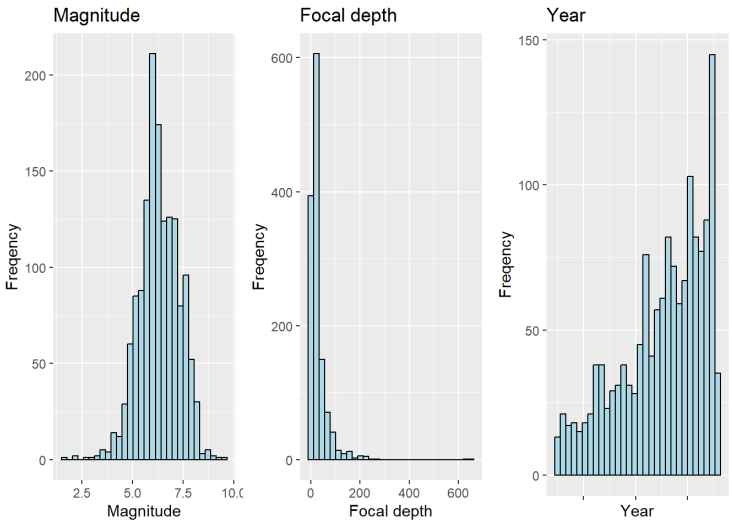
Looking at the categorical outcome variables it can be seen that:
- Variables Missing description, Houses Damaged, Houses Destroyed have a relative low number of data points.
- in all the variables except Death Description and Damage Description there is a noticeable amount of NA values
- the Death description has a rather right skewed distribution.
- the Damage description has a low amount of NA values and the distribution is more balanced.

In general, it can be said that there is more data available in the most recent years. This could indicate that more significant earthquakes happened in the last years. However, it is more likely that more data was recorded in the recent period and some significant data from earlier years might not appear in the data set.

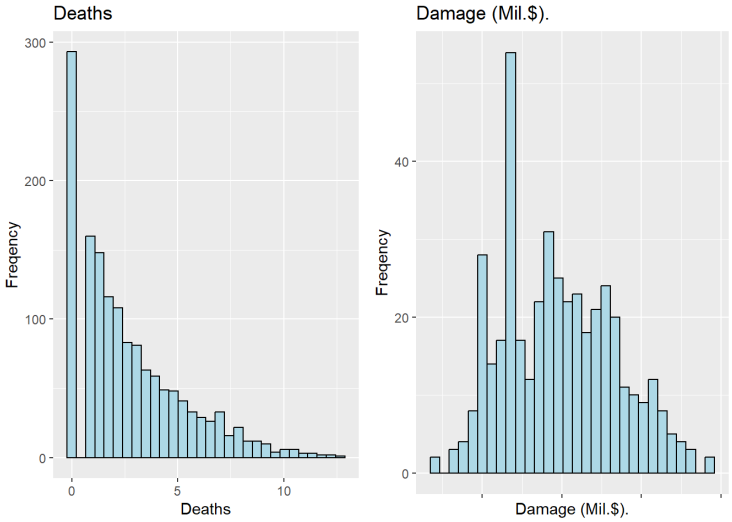
2.3 Preliminary Analysis of continous variables

In the below histograms, the continuous variables of the dataset are plotted to see the distribution of the data points.

Remark:
- Outcomes of an earthquake in terms of death and damage numbers are all count numbers, therefore the logarithms of these values in the below plots will be considered to better see the distribution of the data.
- The data regarding the injuries, house damage and missing people has not been used in the analysis due to the very large number of NULL values, as it diminishes the chances to create a useful model.



On the figures above it can be seen that:
- Magnitude is rather normally distributed
- Focal depth is left skewed
- Year is rather right skewed



Above, only the Death and Damage count numbers have been visualized, given the other variables: Injuries, Missing, Houses.Damaged, Houses.Destroyed have a considerable amount of missing values which would not produce reliable outcome in our analysis.

On the second visual, a close to normal distribution at Damage variable, and a rather left skewed distribution is visible in the case of number of deaths.

3. Linear Model

by Andrea Viczian

3.1 Characteristics of a Linear Model

Generally, linear models are never completely correct, but the interpretability of the linear model is relatively high compared to other more complex models. The danger of over fitting is generally less with linear models. For this reason, the statistical analysis is started with a multiple linear regression below.

Linear regression models are unsupervised models, therefore the aim is to predict how the dependent variable changes with changing independent variables. In regression models, the dependent variable takes quantitative measures and continuous.

3.2 Research Question Linear Model

Given that in the simple linear regression, the dependent variable shall be a continuous and numeric one, the analysis will be started by taking the magnitude of an earthquake as a dependent variable.

The magnitude of an earthquake indicates the released energy of the movement, therefore it is an important indicator of an earthquake. The below independent variables are available in the data set which influence the magnitude of an earthquake:

- location (longitude, latitude)
- focal depth of an epicenter
- time (Year) as independent variables in our model

The other variables: intensity, number of deaths, caused injuries and damages are logically either an outcome of an earthquake, as well they are all counted values, which will be looked at later in the report with other more suitable statistical models.

3.3 Data Cleaning and Graphical Analysis

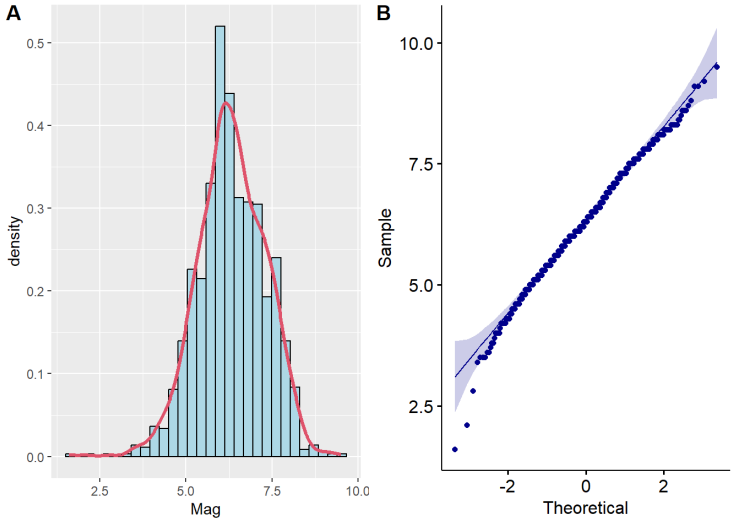
All records have a magnitude value, given that missing values were removed at the first stage of the data cleaning process.

Below, all values where year, focal dept, longitude or latitude is missing are filtered out and the analysis will be run on the below subset of the data:

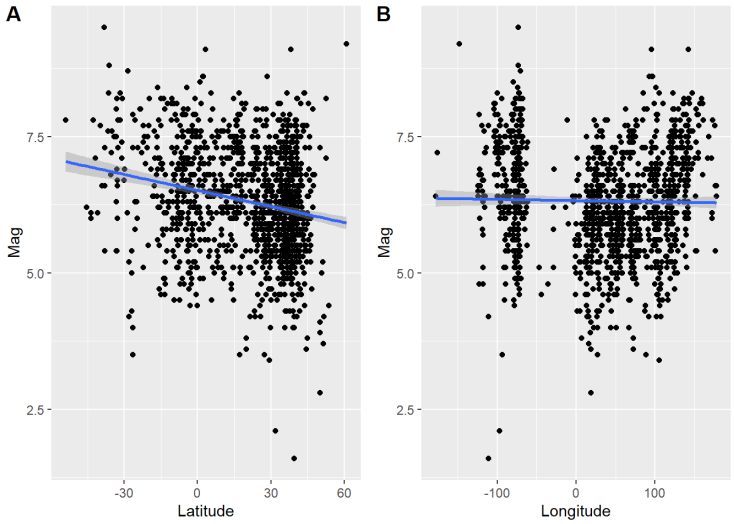
```
## 'data.frame':    1314 obs. of  5 variables:
##  $ Year            : int  1901 1902 1902 1902 1902 1903 1904 1905 1905 1905 ...
##  $ Mag             : num  8.2 6.9 7.5 7.7 6.4 7.8 6.1 7.8 6.6 7.8 ...
##  $ Focal.Depth..km.: int  33 15 33 30 9 100 10 25 20 100 ...
##  $ Longitude       : num  142.3 48.6 -91 76.2 72.3 ...
##  $ Latitude        : num  40.6 40.7 14 39.9 40.8 ...
```

The final data set eqdata.no.na.mag has 1314 rows.

First look at the distribution of the response variable, magnitude with the below histogram. As it can be seen, the highest frequency of the values is between magnitude 6 and 7, the values are decreasing towards zero and towards the value 9.9. The histogram shows a light left skewed distribution, however as shown on the Q-Q plot, the distribution is very close to a normal distribution, which will now be taken as a prerequisite assumption for the further investigation in this chapter with a multiple linear regression model.



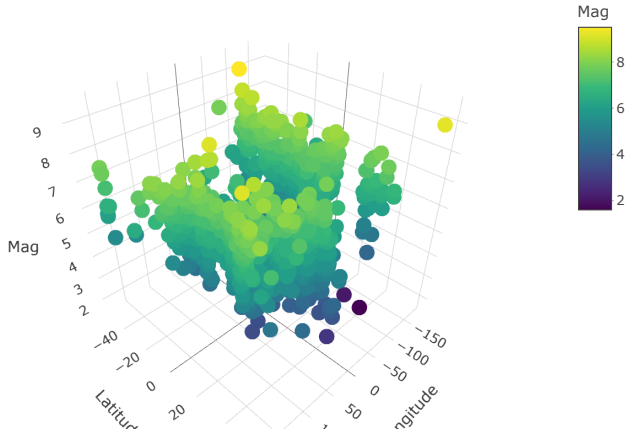
In the next step of the graphical analysis, the relationship between Magnitude and Latitude and Longitude will be analyzed:



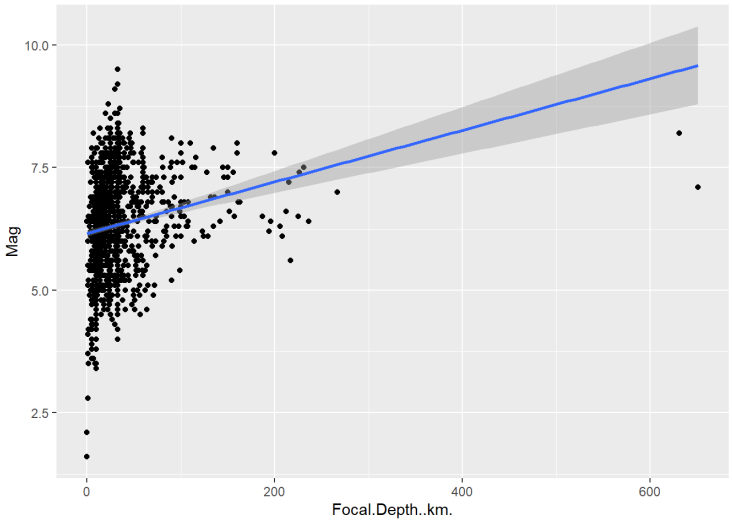
Graph A: It is clearly visible that between the latitude 30 and 60 the number of data points is increasing. This corresponds to the map shown before, this latitude corresponds to the northern hemisphere to the most populated regions: majority of the territories of North America, Europe and Asia is located in this latitude range. A negative correlation can be detected. With each unit of increasing the latitude the magnitude of the earthquakes seemingly decreasing. The smoother on this graph is close to a perfect straight line. Therefore, it may be assumed that there is no hint that the relationship between the response variable and the latitude predictor maybe non-linear.

Graph B: On the plot with the longitude values, two groups can be seen: First is located around the value -100 this value corresponds to the West coast of the North American region. The bigger group of data points is located between 0 and +150 longitude. These values correspond to the Eurasian continent. In both groups is the number of data points higher. However, here, the regression line is almost parallel with the X axis, very flat around the value of 6 magnitude. This may indicate a rather low correlation between Magnitude and Longitude. The shape of the line is close to a perfect straight line. Therefore, it may be assumed that there is no hint that the relationship between the response variable and the latitude predictor maybe non-linear.

It can also be looked at a 3 dimensional interactive plot of the relationship between the longitude and latitude in terms of our dependent variable, magnitude. Logically, the results are as expected, as above plotting the longitude and latitude with the dependent variable magnitude if the graph is turned towards the corresponding axes. On the third dimension, turning the graph in the angle having Longitude on the x-axis and Latitude on the y axis, not surprisingly the distribution of the data points corresponds the world map.

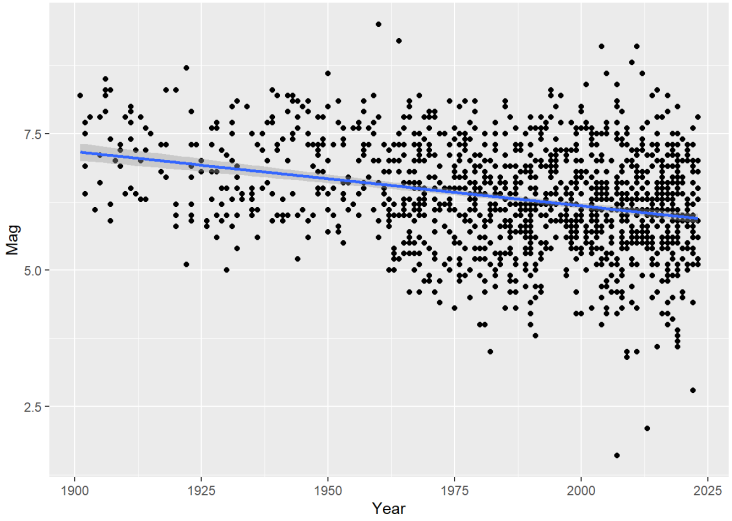


Below, the graphical analysis is continued with the variable Focal Depth:



In case of the focal debt, the values have a high density between 0 and 100 km. A few values are standing out which may be high leverage points, which means their change or removal influences our model more than the removal of other data points would influence. The distribution of the data is left skewed.

Finally, the scatter plot of the variable Year and Magnitude is inspected more closely:



The above scatter plot shows the magnitude by year. It is clearly visible that in the early years between 1900 and 1950, there are less data points visible on the chart. After the 1975, the number of data points increases. More lower values below the value of 3 magnitude can be detected. This may be explained with the advancement of the measuring technologies, or distribution of these technologies throughout the different regions. Similarly as above, the smoother on this graph is close to a perfect straight line. Therefore, it may be assumed that there is no hint that the relationship between the response variable and the latitude predictor maybe non-linear.

3.4 Fitting the Linear Model

After the above preliminary graphical analysis, the linear model with all 4 variables is fitted:
- Magnitude as Dependent,
- Longitude, Latitude, Focal debt and Years as Independent variables.

To see if Latitude and Longitude have an interaction, 2 models will be considered whereas in the second, an interaction term will be included.

The drop1() function is used to fit the linear models below, given continuous variables can equivalently be tested with the drop1() function (i.e. via F-tests) the results of a t-test or a F-test are identical:

```
## Single term deletions
##
## Model:
## Mag ~ Focal.Depth..km. + Latitude + Longitude + Year
##                  Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>                        1029.7 -310.35
## Focal.Depth..km.  1    34.365 1064.1 -269.21  43.6856 5.598e-11 ***
## Latitude          1    64.083 1093.8 -233.01  81.4643 < 2.2e-16 ***
## Longitude         1     7.477 1037.2 -302.84   9.5056  0.002091 **
## Year              1   120.852 1150.6 -166.53 153.6306 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term deletions
##
## Model:
## Mag ~ Focal.Depth..km. + Year + Latitude + Longitude + (Latitude *
##     Longitude)
##                     Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>                           1027.4 -311.33
## Focal.Depth..km.    1    31.559 1058.9 -273.57  40.1797 3.182e-10 ***
## Year                1   118.555 1145.9 -169.83 150.9379 < 2.2e-16 ***
## Latitude:Longitude  1     2.337 1029.7 -310.35   2.9751   0.08479 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.5 Interpretation and Evaluation of the Linear Model

The results of the first model show that there is a strong evidence that all of the variables have a relevant effect on the response variable. It can be seen that longitude has a lower effect, as the p-value is higher. The flat regression line between magnitude and longitude in the preliminary analysis has visually indicated this effect.

According to the results in the second model, there is no evidence of an interaction between longitude and latitude. Also in this model there is a strong evidence that focal depth and year have a strong effect on the dependent variable.

Continuing with a comparison of the models with anova:

```
## Analysis of Variance Table
##
## Model 1: Mag ~ Focal.Depth..km. + Latitude + Longitude + Year
## Model 2: Mag ~ Focal.Depth..km. + Year + Latitude + Longitude + (Latitude *
##     Longitude)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1309 1029.7
## 2   1308 1027.4  1    2.3368 2.9751 0.08479 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output clearly shows that there is no evidence that the second model with the interaction would explain the overall variability of the model better.

To find out that the variables in the first model at all has an effect on magnitude, we set a very basic model and make the comparison with anova:

```
## Analysis of Variance Table
##
## Model 1: Mag ~ 1
## Model 2: Mag ~ Focal.Depth..km. + Latitude + Longitude + Year
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1313 1253.7
## 2   1309 1029.7  4    223.99 71.185 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that there is strong evidence that the model with more parameters better fits the data.The comparison yields the result of a relative large F-value and very small p-value, which indicates a clearly better fit. The “Residual Sums of Squares” (i.e. the unexplained variance, for short RSS) of the more complex model is clearly smaller. In other words, the more complex model explains way more of the variability of these data.

The high RSS value in both models indicates that there is a strong evidence that there are other variables, not included in the data set and statistical model, which better explain the variability of our dependent variable.

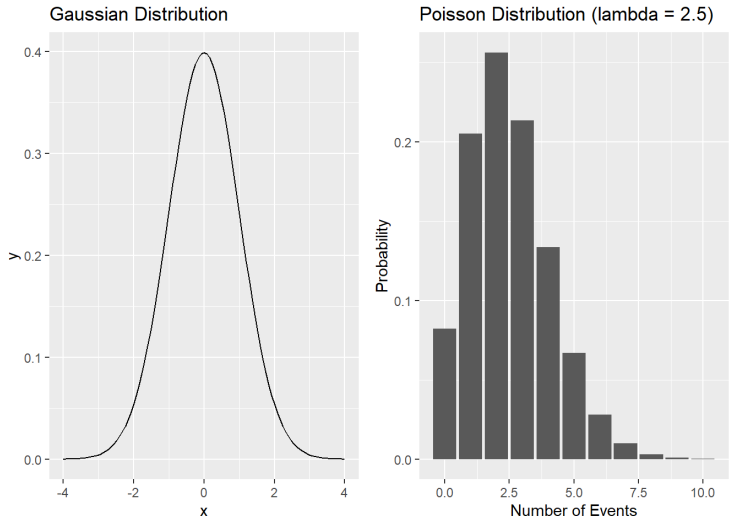
Fitting a simple multiple linear regression model gives a good overall picture. Based on the findings regarding the response variables, there is a need for further investigation with different, more complex statistical models. In the next chapter, the analysis will be continued in more detail on the relationship of the magnitude and the regional differences.

4. Generalised Linear Model set to Poisson

by Esin Isik

4.1 Characteristics of a GLM set to Poisson Distribution

A Generalized Linear Model is the same as a Linear Model with a link set to 1 and assumes a normal (Gaussian) distribution of the data. A GLM set Poisson on the other side does not assume normal distribution, but rather a Poisson distribution as visible on in the plot on the right-hand side. The link function of a GLM set to Poisson is the natural log. It is therefore suitable to be used to analyze data doesn’t have a normal distribution.



A key requirement for a GLM set to Poisson is that mean and variance of the data are equal. However, in real-case scenarios, this is often not the case. Therefore, the “quasipoisson” family will be considered in the following analysis.

4.2 Research question GLM set to Poisson

Another key requirement of fitting a GLM set to Poisson or Quasipoisson is the characteristic of the predictor in the model. The predictor can only be count data, for which there are plenty of variables given for this dataset. Having fitted a Linear Model for the Magnitude in the previous chapter, it can be sensible to dive in once more and transform the magnitude values to treat it as counts.Therefore, creating a magnitude column without decimals, it would allow to fit it into a Quasipoisson GLM and would still not loose its’ ability to be interpreted.

Moving on, it will also be analyzed which variables could have a significant influence on the number of deaths.

4.3 Fitting the Poisson GLM

Reviewing the factual background which the available variables in the dataset are based on, it would not make much sense to analyze a possible influence on the magnitude in variables that state the effects after an earthquake. This means that variables such as deaths, injuries, or material damages that occur as after-effects of an earthquake should not be considered for this model. Based on the availability in this dataset, variables that could explain a magnitude, on the other side, are the focal depth of the epicenter, the region in which an earthquake occured, and the specific country.

Fitting the previewed model could be written in this form:
$\text{Magnitude} = \beta 0 + \beta 1 \cdot \text{Focal Depth} + \beta 2 \cdot \text{Region15} + \ldots + \beta n \cdot \text{CountryArgentina} + \ldots$
$y \sim \text{Quasipoisson}(\lambda)$

Contrarily, the Quasipoisson GLM fitted on the count of deaths will be based on the after-effects. Considered variables for the model are Magnitude, Country and Damage Description.

4.4 Model Interpretation and Evaluation GLM set to Poisson

4.4.1. GLM Magnitude

Fitting the presented GLM for Magnitude yields the following results:

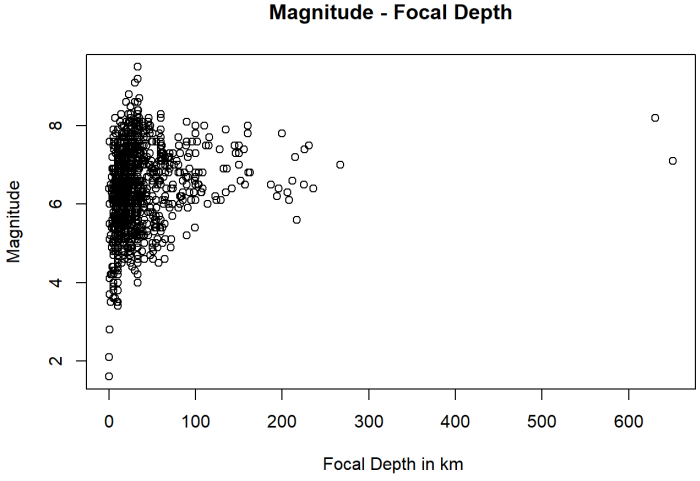
Results of Fitting the GLM - Magnitude (interactive drop down button)

Focal Depth:

```
exp.coef.focaldepth <- exp(coef(glm.fit)["Focal.Depth..km."])
cat("Coefficient for Focal Depth:", exp.coef.focaldepth, sep = " ")
```

```
## Coefficient for Focal Depth: 1.00053
```

As expected, the focal depth of an earthquake has statistically significant influence on the magnitude of an earthquake. The model reveals that increasing the focal depth by 1km, it would result in a magnitude higher by 0.05%.



However, plotting focal depth and magnitude reveals that many earthquakes, especially also strong ones, happen at a low depth more frequently. This very slow increase in magnitude the deeper the epicenter is situated, could be traced back to the few, but strong, earthquakes that happened at a very high focal depth.

Region 120:

```
## Coefficient for Region 120: 0.5387018
```

Also, high significance for Region 120 can be visible. Looking at the coefficient, it can be seen that in Region 120, countries get in average around 53.9% lower magnitudes than region 10. Being reminded that region 120 represents Northern and Western Europe, this outcome seems very plausible as Europe is not known for earthquakes with a strong magnitude.

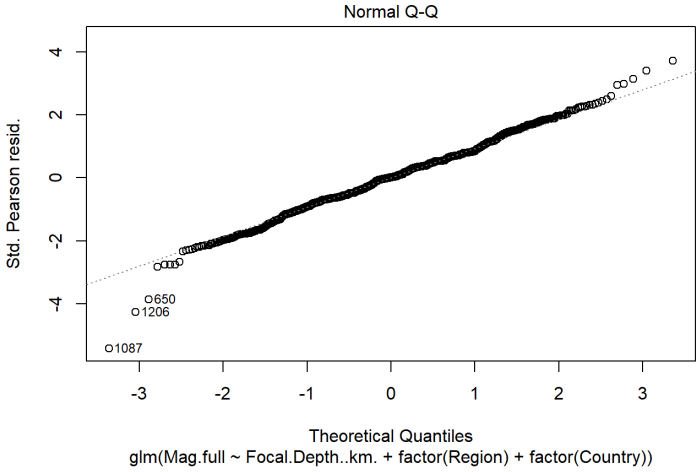
Country Japan:

```
## Coefficient for Country Japan: 1.323644
```

Looking at the country-wise significance, it shows that especially the grounds at the location of Japan, Mexico, Mongolia, Taiwan, and Turkmenistan show strong statistically significant patterns regarding the magnitude of an earthquake. Indeed, combined with common knowledge, e.g., Japan happens to be known for its pattern in high earthquake magnitudes. The model reveals that Japan has earthquakes that are on average 32.4% at a stronger magnitude than in Afghanistan.

Evaluation

The summary shows that the dispersion parameter is lower than 1. This implies that the variance increases slower than linearly. Also, the residual deviance of 166.92 is much lower than the degrees of freedom of 1203. This implies that the data shows a variability that is smaller than expected in the Poisson distribution, i.e., underdispersion.
Nevertheless, there have been 4 Fisher Scoring iterations. This speaks for the fact that the complexity of the model might be considered adequate for this data.
Furthermore, the Normal Q-Q Plot reveals the following:



The Q-Q Plot of the model shows no consistent lay-offs which would speak for a moderate fit. However, summarizing the analysis of this model, the residual deviance and its degrees of freedom differ greatly which is problematic. Therefore, in order to be able to fit a model on the magnitudes of an earthquake which shows high accuracy, more extensive data about factors that have an influential role on a magnitude should be considered. Further research has shown that these variables could be the intrinsic quality (coefficient of friction in the rock), the rupture area (whether the epicenter is in a subduction zone), the average displacement across the rupture area, and the directivity (direction of energy release (energy release in the direction of movement)).

4.4.2. GLM Deaths

The GLM for the count of deaths yields the following results:

Results of Fitting the GLM - Deaths (interactive drop down button)

Magnitude:

```
## Coefficient for Magnitude: 5.708963
```

As expected, the magnitude of an earthquake has statistically significant influence on the death count resulting from an earthquake. The model reveals that increasing the magnitude by 1 unit, it would result in a higher count of deaths by a factor of 5.7 (470%).

Country Haiti:

```
## Coefficient for Country Haiti: 44.80899
```

Looking at the country-wise significance, it shows that especially the grounds at the location of Armenia, Chile, Haiti, Indonesia, Morocco, Nicaragua, and Turkmenistan show strong statistically significant patterns regarding the count of deaths in connection to earthquakes. Strongest significance is shown for Haiti: The model reveals that Haiti has earthquakes that result on a death count that is on average 44.8 times the expected count than in Afghanistan (reference level).

Damage Description 4:

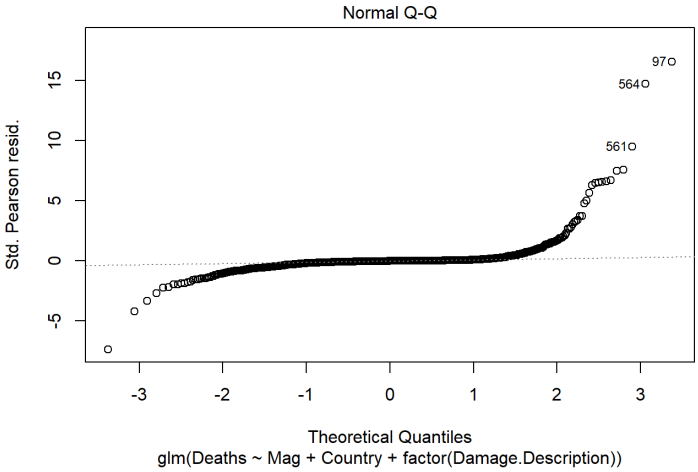
```
## Coefficient for Damage description 4: 138.6911
```

The coefficient of Damage Description 4 (25 million Dollars or more) reveals that the count of deaths are expected to be on average 138.7 times the count of deaths for an earthquake that had a Damage Description of 1 (less than 1 million Dollars)

Evaluation

The summary shows that the dispersion parameter is higher than 1. This implies that the variance increases faster than linearly. Also, the residual deviance of 4701643 is much higher than the degrees of freedom of 1270. This implies that the data shows a variability that is bigger than expected in the Poisson distribution, i.e., overdispersion.

Nevertheless, there have been 7 Fisher Scoring iterations. This speaks for the fact that the complexity of the model might be considered adequate for this data.
Furthermore, the Normal Q-Q Plot reveals the following:



The Q-Q Plot of the model shows consistent lay-offs at the beginning and at the end which would not speak for a moderate fit. Summarizing the analysis of this model, the residual deviance and its degrees of freedom differ greatly which is problematic. Testing models by varying the given variables in the data set could not produce a well-suited model to predict deaths. Therefore, in order to be able to fit a model on the death counts resulting from an earthquake which shows high accuracy, more extensive data about factors that have an influential role on the death count should be considered. These could be variables that show the amount of inhabitants in the affected region, quality of infrastructure in the region, accessibility of rescue operations, and average age of inhabitants as chances of survival can also highly depend on this factor.

5. Generalised Linear Model set to Binomial

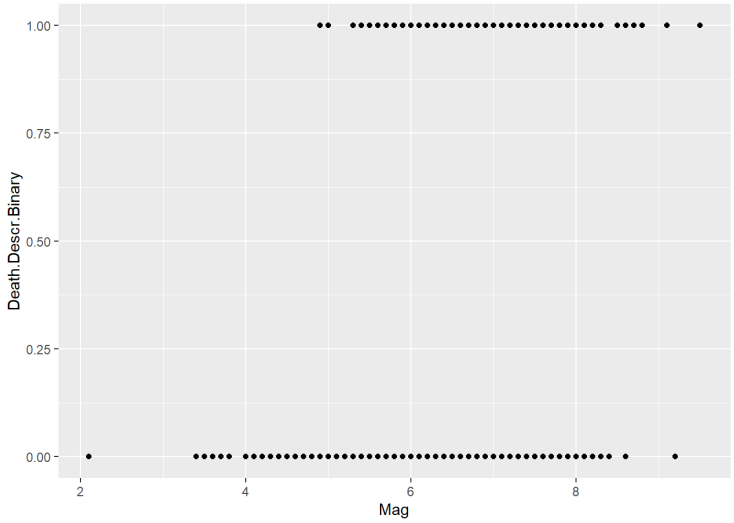
by Esin Isik

5.1. Characteristics of a GLM set to Binomial

A Generalized Linear Model set to family Binomial has the same structure as a poisson/quasipoisson GLM but has the family argument set to “binomial” which uses the logit link. A GLM set to binomial is suitable for the analysis of binary and binomial data. Binary data can include states with two options (e.g., dead, alive), whereas binomial data can stand for proportions (e.g., the proportion of incomplete records in a data set). I.e., the response variable has to be bound between 0 and 1 and does not follow a normal distribution. Binary or binomial data are therefore unsuitable to be analyzed with Linear Models. This type of GLM fitted on binary data is also often referred to as a “logistic regression”.
In a lot of real scenarios, however, there is also the possibility that there are more than two options that can occur. This kind of data can be modeled with an extension of the logistic regression: the “multinomial regression” method, which represents the characteristics of a Random Forest.

5.2. Research Question GLM set to Binomial

Within this project, the GLM binomial will be fitted to analyze Magnitude against whether the Death Description will be category 1,2 or higher. To prepare the data for the model, an additional column will be created that assigns 0’s and 1’s to the death description respectively.
The binomial GLM will be fitted to predict the odds of the death description being 1 or 2 (1-50 and 51-100), or, 3 or 4 (101-1000 and over 1000) depending on:
- the magnitude of an earthquake,
- and the Damage Description.



Plotting the Death Description against the Magnitude, a sensible distribution of the values can be detected: Death Description 3 and 4 only occur between magnitude 5 and approx. 10. The Death Descriptions 1 and two, on the other side, have a less dense distribution as the range within this data is along magnitude 3.5 and 8.5.
However, possible outliers can be detected in this plot:
- an earthquake at magnitude of approx. 2 and a Damage Description of 4 (Extreme - $25 million or more). The record identified is an earthquake that occurred in Texas in 2013. Further research has shown that it was in fact a massive explosion at a fertilizer production site and is therefore not an earthquake that occurred “naturally”.
- Also, a record is showing a magnitude over 9 but a low Death Description. The data reveals it occurring in Alaska in 1964. From this fact, it can be derived that additional information such as the number of inhabitants in the surrounding area would highly increase the explainability of the data with a fitted model to predict a Death Description. Therefore, it has to be remarked that earthquakes covered in this data also occurred in not or very few populated areas. The accuracy of the predictions could therefore be increased if this important factor would be known.

5.3. Fitting the Binomial GLM

Results of Fitting the binomial GLM (interactive drop down button)

5.4. Model Interpretation and Evaluation GLM set to Binomial

Interpreting the coefficients

```
## Coefficient Magnitude: 3.654098
```

```
## Coefficient Damage Description 2: 6.397946
```

```
## Coefficient Damage Description 3: 20.39144
```

```
## Coefficient Damage Description 4: 65.83646
```

```
## Coefficient Turkey: 9.668402
```

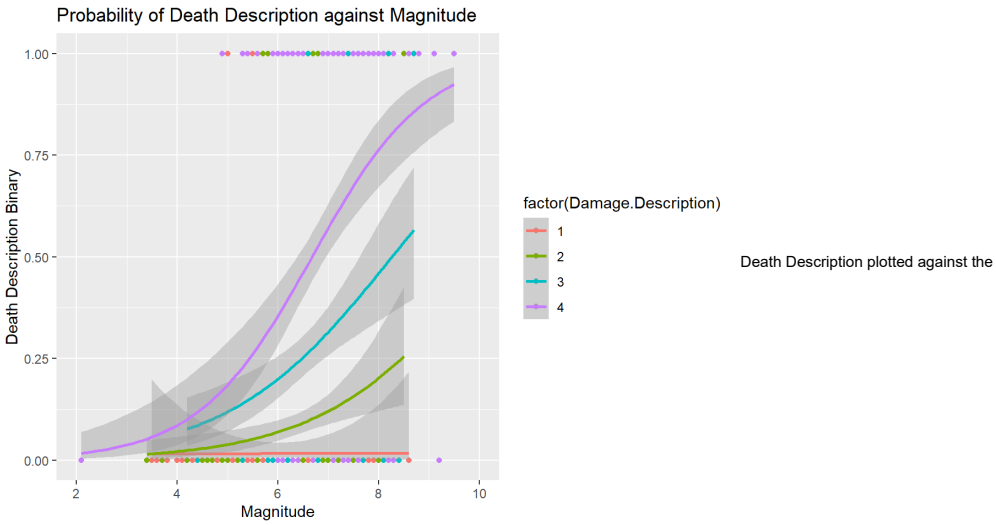
```
## Coefficient Iran: 12.4113
```

```
## Coefficient El Salvador: 31.98473
```

```
## Coefficient Italy: 6.511287
```

Several significant factors in the model could be detected:
- It is revealed that having an increased Magnitude by one unit, it is 3.6 times more likely to have a Death Description of 3 or 4 (DD 3-4). (101-1000 and over 1000)
- The odds of having DD 3-4 are approx. 6.4 times higher with a Damage Description 2 (51-100) than with a Damage Description 1 (1-50). In the same way, the odds are higher with Damage Description 3 (101-1000) by 20.4 times. For Damage Description 4 (over 1000), the odds are higher by 65.8 times.
- Regarding the location (country) the following countries have the relatively highest odds of having DD 3-4 relatively to Greece (releveled to Greece):
The odds for having DD 3-4 in Turkey is approx. 9.7 times higher than in Greece. In the same way, these odds are 12.4 times higher for Iran, 32 times for El Salvador, and 6.5 times for Italy.

Plotting the model



Death Description plotted against the Magnitude and Damage Description, it can be seen that the odds of having a Death Description of 3 and 4 (101-1000 and over 1000) rise more steeply starting from Magnitude 4. Removing the outliers has been tested and did not result in a much clearer visual.

Pseudo R squared

```
## Pseudo R squared: 0.3881445
```

The pseudo R^2 for the fitted binomial GLM can give indication on the goodness of fit on the data. Once calculated, it reveals that almost 40% of the data could be explained by this model.

Confusion Matrix

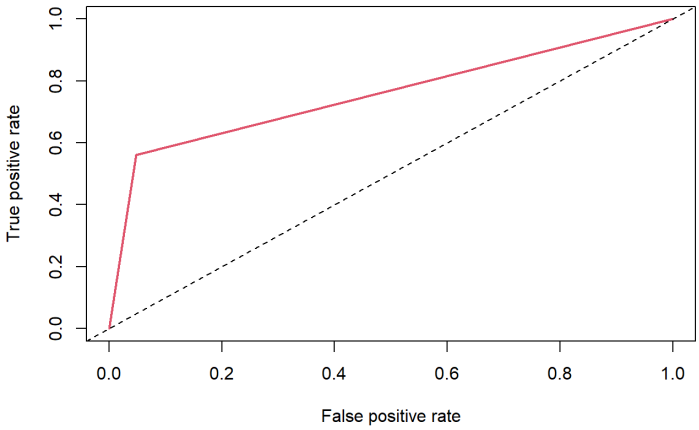
```
## Confusion Matrix:
```

```
##    fit
## obs    0    1
##   0 1036   52
##   1  128  164
```

```
## Error Rate: 0.13043
```

The confusion matrix of the model predictions reveals: 1036 observations were correctly labelled as low Death Descriptions 1-2 (1-50 and 51-100), and 164 obs. were correctly labelled as Death Description 3-4 (101-1000 and over 1000). However, 128 Death Description 3-4 observations were misclassified as Death Description 1-2 while 52 were wrongly labelled in the other sense.
Overall, the error rate of 13% is low and is an indication for a good fit.

ROC Curve




```
## ROC curve Performance: 0.7569249
```

The ROC curve of the model shows a movement towards the upper left corner which suggests a higher positive rate for a given negative rate and therefore stands for a good performance. The overall performance indicated by the Area Under the Curve, AUC, is approx. 75.7%.
As various factors in the evaluation have shown, this model could be a good fit to estimate the description of deaths followed by earthquakes. However, through the visualizations and further research, it has been discovered that the data set includes earthquakes that were caused by humans (factory explosion Texas 2013). Registered occurrences like these can massively falsify any predictions attempted with this model as the different variables interact differently that should clearly be identified as outliers (Texas: Magnitude 2, Death Description 1, Damage Description 4). Unfortunately, the additional information about cases like these is not given in this data set and therefore, estimating future events is not recommended with this model.

6. Generalised Additive Model

by Andrea Viczian

6.1. Characteristics of a GAM

GAM is the model that are the extension to smoothing splines and enables to fit models which contain several predictors simultaneously. Advantages of a GAM model are, modelling non linear relationships, modelling multiple predictors and interaction effects, further on, GAM is also a robust model for handling outliers.

Difference Generalized Linear Model (GLM) vs. Generalized Additive Model (GAM)

GAM does not assume a priori any specific form of this relationship, therefore can be used to reveal and estimate non-linear effects of the covariate on the dependent variable. GAMs assume that the relationship between the response variable and predictors is additive, meaning that the effect of each predictor is independent of the others.

This allows for more flexibility in modeling complex relationships without explicitly specifying interactions. GAMs can accommodate both continuous and categorical predictor variables. Categorical variables are typically represented by dummy variables or factor levels.

6.2. Research Question GAM

In this chapter the flexibility of the GAM model will be explored. Previously, the relationship of Magnitude and Death Description variable were investigated. In this section, the number of independent variables will be extended and fit a GAM model to see how magnitude, intensity, focal debt and regional differences are influencing the levels of number of death caused by an earthquake.

Assessing the predictive performance of a model: Cross validation

The aim is to create a model to be able to make predictions for the future. Therefore, it will be proceeded with splitting the data in train and test parts. Assessing and comparing the predictive performance of the models at the end of this chapter are expected to be enabled by this.

During the analysis issues have been encountered that were caused by missing values in the predict function at the cross validation phase. As a result, it was decided to remove the missing values and create a new subset of the data set for this analysis. This resulted having considerably less data points in the new subset of the data. The split in train and test parts needed to happen balanced across all factorized variables, such as Region and MMI.Int. (Intensity) in order to conduct the prediction and cross validation of the fitted models.

For fitting the model randomly generated subsets balanced of the data has been generated :

New train and test subset of the data (interactive drop down button)

6.3. Fitting a GAM

There are 2 variables in the data set indicating the death numbers caused by an earthquake one with the count numbers, one with categorical levels calculated from these numbers. Below, 2 models will be fit to compare the predictive power of the outputs.

Furthermore, the factorized variable Region will be included to showcase regional effects and differences. For spatial effects a 2 dimensional smoothing term has been added including the variables Longitude and Latitude.

Each following section will include their result.
The validation results using the gam.check() function are wrapped in the interactive drop down buttons below.

1. Response Variable: Fitting GAM Models with Death.Description Variable (categorical / 4 levels)

The family multinom() will be used given the categorical response variable.

Remarks to optimization:
In this GAM Model, Magnitude had an edf value of 1 therefore they have been added without a smoother to the model. This is clearly visible in the 3-D plots below.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Death.Description ~ Mag + factor(MMI.Int) + s(Focal.Depth..km.) +
##     factor(Region) + s(Longitude, Latitude)
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.82075    0.86274  -2.110  0.03567 *
## Mag                0.35468    0.07738   4.583 6.79e-06 ***
## factor(MMI.Int)3   0.35383    1.04591   0.338  0.73538
## factor(MMI.Int)4   0.10156    0.58165   0.175  0.86151
## factor(MMI.Int)5   0.11542    0.55565   0.208  0.83559
## factor(MMI.Int)6   0.36127    0.54593   0.662  0.50865
## factor(MMI.Int)7   0.21991    0.53377   0.412  0.68065
## factor(MMI.Int)8   0.61746    0.53126   1.162  0.24608
## factor(MMI.Int)9   0.94336    0.54555   1.729  0.08483 .
## factor(MMI.Int)10  1.20163    0.55915   2.149  0.03246 *
## factor(MMI.Int)11  1.80394    0.59409   3.036  0.00261 **
## factor(MMI.Int)12  2.00041    1.03704   1.929  0.05471 .
## factor(Region)15   0.06529    0.69624   0.094  0.92535
## factor(Region)30   2.02922    0.81204   2.499  0.01301 *
## factor(Region)40   0.63130    0.72185   0.875  0.38253
## factor(Region)50   2.29205    1.10882   2.067  0.03961 *
## factor(Region)60   1.38525    0.68169   2.032  0.04305 *
## factor(Region)90  -0.68026    1.18984  -0.572  0.56795
## factor(Region)100 -0.45664    1.37257  -0.333  0.73961
## factor(Region)110  0.20573    0.76611   0.269  0.78847
## factor(Region)120 -0.23313    1.09053  -0.214  0.83087
## factor(Region)130 -0.04872    0.62913  -0.077  0.93833
## factor(Region)140  0.65013    0.64490   1.008  0.31424
## factor(Region)150 -0.70192    1.49666  -0.469  0.63943
## factor(Region)160 -0.34053    1.22426  -0.278  0.78109
## factor(Region)170  2.06762    0.81577   2.535  0.01178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df     F p-value
## s(Focal.Depth..km.)    1.306  1.551 4.224 0.01590 *
## s(Longitude,Latitude) 11.070 15.076 2.349 0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.398   Deviance explained = 46.6%
## GCV = 0.82975  Scale est. = 0.73325   n = 330
```

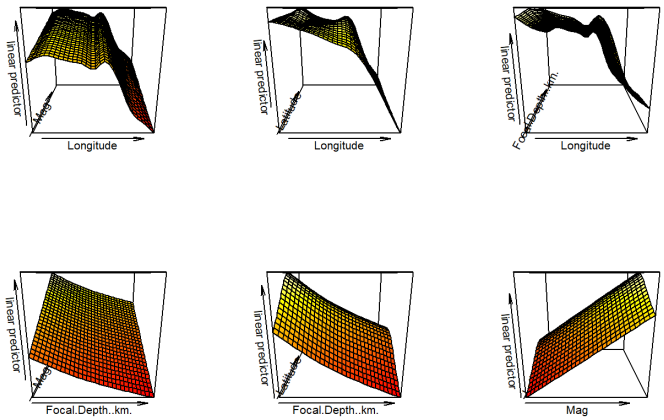
Interpretation:

The summary output in the below model indicates that there is a strong evidence that Magnitude has a linear effect on the response variable and there is some evidence that the 2 dimensional term, longitude and latitude and focal depth have a non-linear impact on the number of death cases.In the below graph, a non-linear relationship is illustrated.

There is no evidence that the Intensity or Regional factor levels have an influence on the number of death caused by an earthquake. However there is a weak evidence that the higher intensity levels differ in effect from the reference level 3 and similarly that region 30 (East-Asia), 50 (Kamchatka) and 60 (S. and SE. Asia and Indian Ocean) differ from the reference level region 15 (Northern Afrca). This may however change slightly with having a different randomized set of our train and test data.

The model explains more than 45% the overall variability, based on the R squared value.

Check GAM1: Death Description + Region (interactive drop down button)



2. Response Variable: Fitting GAM Models with Death (count variable)

Now using family = “quasipoisson”: Given the response valuable is a count data, a link with log function is needed.

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## Deaths ~ s(Mag) + factor(MMI.Int) + s(Focal.Depth..km.) + s(Longitude,
##     Latitude) + factor(Region)
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -10.5148     2.2859  -4.600 6.58e-06 ***
## factor(MMI.Int)3    3.6460     3.5346   1.032 0.303253
## factor(MMI.Int)4   -5.1632     1.4523  -3.555 0.000448 ***
## factor(MMI.Int)5   -1.3554     1.4865  -0.912 0.362685
## factor(MMI.Int)6    2.2681     1.4502   1.564 0.119009
## factor(MMI.Int)7    4.3817     1.4474   3.027 0.002713 **
## factor(MMI.Int)8    1.4060     1.4478   0.971 0.332384
## factor(MMI.Int)9    2.4993     1.4481   1.726 0.085537 .
## factor(MMI.Int)10   2.8965     1.4479   2.000 0.046487 *
## factor(MMI.Int)11   5.6916     1.4485   3.929 0.000109 ***
## factor(MMI.Int)12   3.3145     1.4493   2.287 0.022997 *
## factor(Region)15  -10.1268     0.6531 -15.505  < 2e-16 ***
## factor(Region)30    6.7291     0.2663  25.270  < 2e-16 ***
## factor(Region)40    1.2752     0.2417   5.275 2.77e-07 ***
## factor(Region)50    9.1405     0.3368  27.141  < 2e-16 ***
## factor(Region)60    3.6536     0.2166  16.869  < 2e-16 ***
## factor(Region)90   25.6591     6.4186   3.998 8.32e-05 ***
## factor(Region)100  35.9392     6.7901   5.293 2.54e-07 ***
## factor(Region)110   3.7748     0.3013  12.529  < 2e-16 ***
## factor(Region)120  -6.9301     3.3052  -2.097 0.036978 *
## factor(Region)130   0.7018     0.3972   1.767 0.078375 .
## factor(Region)140   0.8051     0.2446   3.292 0.001132 **
## factor(Region)150  42.1135     6.8026   6.191 2.29e-09 ***
## factor(Region)160  19.1678     6.5919   2.908 0.003951 **
## factor(Region)170   7.0972     0.2968  23.915  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df    F p-value
## s(Mag)                  7.241  7.811 2910  <2e-16 ***
## s(Focal.Depth..km.)     8.195  8.429 1361  <2e-16 ***
## s(Longitude,Latitude) 27.084 27.960 1675  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.982   Deviance explained = 93.9%
## GCV = 1461.7  Scale est. = 10.397    n = 330
```

Interpretation:

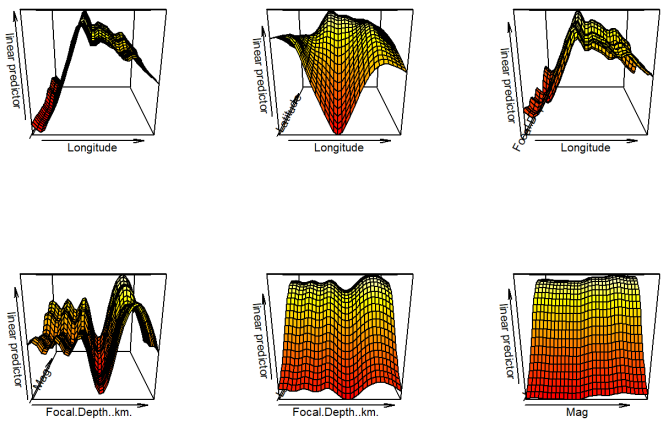
This second model with the count death numbers as response variable seem to have a strong evidence that the intensity levels and regions have influence on the number of deaths caused by earthquakes as well as that the majority of the levels and regions have a difference to the reference level 3 intensity and region 15 (North Africa).

It can also be seen that all the variables in smoother functions have a very high edf value (close to 9) which is a strong indication of an overfit.

The R-squared value is above 90%, in other words this model claims to explain 90% of the variability in the model. In the 3-D plots the overfit of the model is clearly visible.

The GAM check function indicates in the Q-Q plot, a clear non-linearity, as well as the frequency plot shows clearly a rather non-normal distribution, which are indicators of a less suitable model fit.

Check GAM2: Deaths + Region (interactive drop down button)



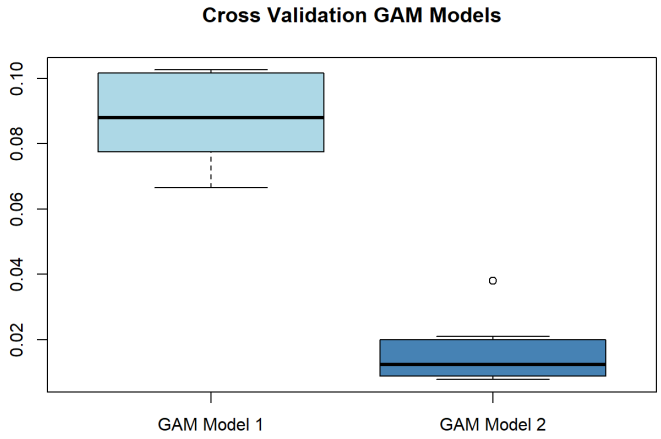
6.4. Model Interpretation and Crossvalidation GAM

The data has been balanced and divided randomly however to achieve a robust prediction value, it is recommended to run the validation process repeatedly several times. In this code were only 10 repetitions conducted to keep the compiling times in a reasonable time frame. The overall results show that increasing the run time results a more robust outcome of a predictive capacity of the model.

The cross validation process comparing our 2 GAM models yields the below results:

```
## 1. GAM Model: Death.Description response variable / Region: 0.08775004
```

```
## 2. GAM Model: Deaths response variable / Region: 0.01552855
```



Given the results it can be concluded that the **1. GAM Model**, with the categorical response variable including the Regional factor levels has a better result in the comparison. However the resulted values are in both cases relatively low, 8.7% and 1.5% therefore it can be concluded that none of the above models is suitable to make realistic predictions for future impacts caused by an earthquake.

Further on, it can confidently be said that the high proportion of he missing values, which resulted in a decreased number of records with a relative high number of predictors, had a considerable negative impact on our outcomes. Based on the low R squared values, there is a strong indication of the existence of several other factors outside of the scope of this report, not included in our model nor in the data set, which may have a high impact on the caused death numbers, influence and explain the variability of the model in higher proportion.

7. Neural Network

by Sabrina Rigo

7.1. Characterisics of a Neural Network

The idea behind an Artificial Neural Network is to create a model that is comparable to the human brain. With this a relationship between the set of data should be determined. Multiple input nodes are connected to the hidden nodes. More hidden nodes mean that the model is more complex and therefore can learn more difficult concepts. In the end stands the output and the relationship between the input and the output will be modeled.

7.2. Research Question Neural Network

With a neural network, a prediction can be made. First, the damage description will be predicted. The Magnitude, Focal Depth, Region, Deaths, Tsunami and Volcano will be used to check if a prediction can be made.

In a second step, the connection from Deaths should be predicted by looking at the Magnitude, Focal depth, damage description and the respective time. If there was a Tsunami or a Volcano, it will also be considered. The question now is how correlated these variables are when looking at the deaths. With the result predictions can be made.

7.3 Training a Neural Network

Predictions for Damage.Description

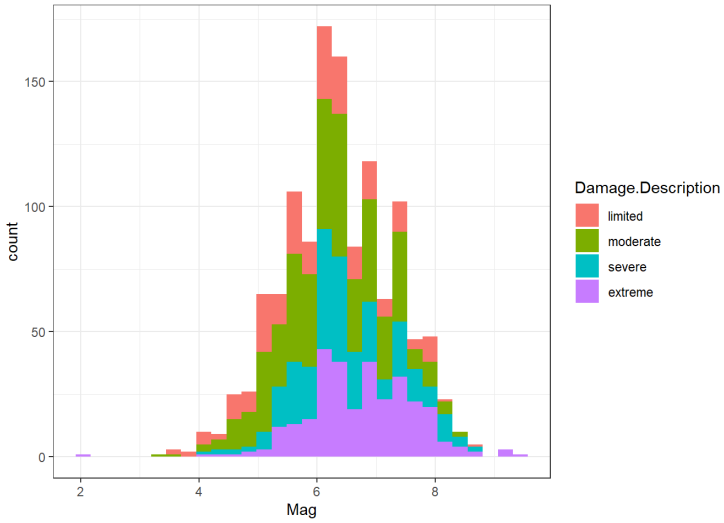
The Damage.Description contains values from 1 to 4 which stands for different categories (as already mentioned before). The column needs to be converted to a factor with different levels so that the categories can be used for the predictions. The Tsunami and Volcano values will be changed to 1 or 0 if there was a Tsunami or Volcano eruption or not. The new dataframe used for the analysis includes 1235 observations and 7 variables. All the null values were removed. The damage description is also factorised with the four levels “limited”, “moderate”, “severe” and “extreme” represented with the numbers 1-4.

A new column Type will also be created including Vulcano, Tsunami, Both, Neither. This will then also be used in chapter 8.

```
## 'data.frame':    1235 obs. of  7 variables:
##  $ Mag               : num  8.2 6.9 7.5 7.7 6.4 7.8 6.1 7.8 6.6 7.8 ...
##  $ Focal.Depth..km.  : int  33 15 33 30 9 100 10 25 20 100 ...
##  $ Damage.Description: Factor w/ 4 levels "limited","moderate",..: 2 4 3 4 4 3 3 4 3 1 ...
##  $ Region            : int  30 40 100 40 40 130 130 60 130 30 ...
##  $ Tsu               : num  1 0 1 0 0 0 0 1 0 0 ...
##  $ Vol               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths            : int  18 86 2000 2500 4880 2 4 19000 120 11 ...
```

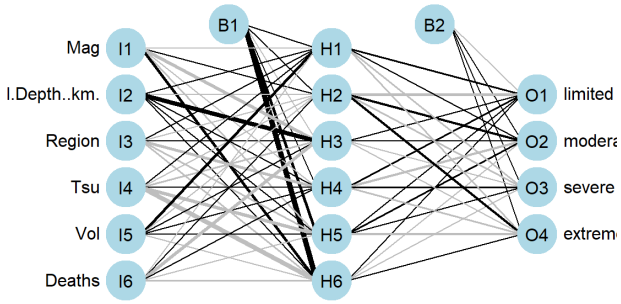
```
## [1] "limited"  "moderate" "severe"   "extreme"
```

The following plot shows the distribution of Magnitude and the colors of the damage description. It makes clear that with any number of magnitude the damage of the four different categories can happen. This could be the case since the data only shows significant earthquakes. Technically no earthquake is just a light movement of the tectonic plates.



Now the data is divided into training and test data and an NNet is created with 6 hidden layers. The output is the four different categories from the damage description.

create NNet code (interactive drop down button)




```
##           true
## pred       limited moderate severe extreme
##   extreme        0       16     22      42
##   limited        5        7      3       3
##   moderate      37       57     29      18
##   severe         1        0      0       1
```

```
## [1] 0.153527
```

The number of 6 hidden layers was chosen here since this gives the highest number of accuracy that was able to be achieved. With 30%, it is still quite low. Another point made clear is that there are no predicted values for the limited damage description. This could be the case since only 1235 observations for this neural network could be used.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction limited moderate severe extreme
##   limited        5        7      3       3
##   moderate      37       57     29      18
##   severe         1        0      0       1
##   extreme        0       16     22      42
##
## Overall Statistics
##
##               Accuracy : 0.4315
##                 95% CI : (0.3681, 0.4967)
##    No Information Rate : 0.332
##    P-Value [Acc > NIR] : 0.0007986
##
##                  Kappa : 0.1907
##
##  Mcnemar's Test P-Value : 1.116e-13
##
## Statistics by Class:
##
##                      Class: limited Class: moderate Class: severe
## Sensitivity                 0.11628          0.7125      0.000000
## Specificity                 0.93434          0.4783      0.989305
## Pos Pred Value              0.27778          0.4043      0.000000
## Neg Pred Value              0.82960          0.7700      0.774059
## Prevalence                  0.17842          0.3320      0.224066
## Detection Rate              0.02075          0.2365      0.000000
## Detection Prevalence        0.07469          0.5851      0.008299
## Balanced Accuracy           0.52531          0.5954      0.494652
##                      Class: extreme
## Sensitivity                  0.6562
## Specificity                  0.7853
## Pos Pred Value               0.5250
## Neg Pred Value               0.8634
## Prevalence                   0.2656
## Detection Rate               0.1743
## Detection Prevalence         0.3320
## Balanced Accuracy            0.7208
```

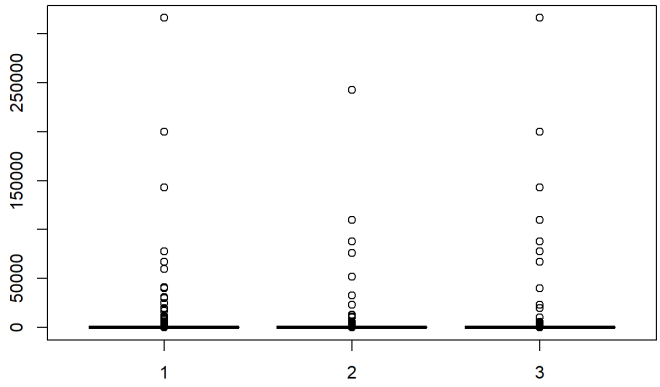
The overall statistics shows that there is an accuracy of 43%. The higher the number of hidden layers is set, the higher the accuracy gets. The accuracy rate needs to be higher than the no information rate to even consider these predictions. A lower no information rate would be desired in this case.

The confusion matrix shows that there are a few misclassifications in the predictions. The moderate and the extreme classes show the best results.

The ROCR validation with the class type does not work in this case since there are 4 different classes.

Predictions for Deaths

Since this first Neural Network did not seem to be a perfect model a second Neural Network was created by looking at the predictions that can be made for the deaths. The dataframe with 1235 observations is used again.

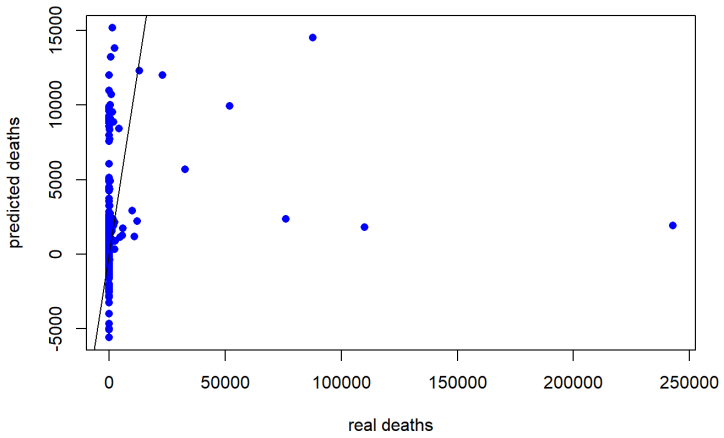


The boxplots show the separation into the train and test group as well as a sample fraction of 2%. Here, it already becomes apparent that the train (1) and test (2) are not including the same type of data. It also shows that most data points are close to 0, with only a few exceptions.

Usage of neuralnet code (interactive drop down button)

After dividing the data into train and test, it can than be used to train the neural network. This ANN was created with 3 hidden layers since this is giving the best results. It is clear to see that with these 6 input nodes (Magnitude, Focal depth, Damage Description, Region, Tsunami, Volcano) the model is trained to give the deaths output.

Afterwards, the scaled prediction will be created and this will be plotted with regard to the real deaths.



7.4. Evaluation and Adjustment

The predictions for the damage description could be used for the prediction of moderate or even extreme earthquakes. However, it would make more sense to try and get more data before using it since the model is pretty weak.

The prediction for the deaths looks good in the plot above. To check if the model is useful for a prediction, it needs to be checked how correlated the true values and the predicted values are.

```
##            [,1]
## [1,] 0.05129658
```

With these 5% it is clear that the true values and the predicted values are far off from each other. This could be the case because there are some extremely high values which only exist once. Another reason could again be the small size of observation that we were able to use for this analysis. The many missing values or not knowing what exactly they mean is a big problem.

In general, it can be said that in reality the deaths are mostly a small number but the predicted values are spread out. There are even some minus values which makes no sense for a death prediction.

It was tried to improve the model with a softplus function that smoothes out the results.

This model takes far more steps than the first one and the Error decreases slightly.

```
##            [,1]
## [1,] 0.09699469
```

With a new increased correlation of 9% the model became a little better but is still far off from what is acceptable.

```
## Root Mean Squared Error (RMSE): 3640.44
```

```
## Mean Absolute Error (MAE): 2190.576
```

```
## R-squared: 0.01987592
```

The Root Mean Squared Error/Mean Absolute Error/R-Squared also show again that this prediction is not the best as it is in general being more than 3640 deaths off in the predictions.

8. Support Vector Machine

by Sabrina Rigo

8.1 Characterisics of Support Vector Machines

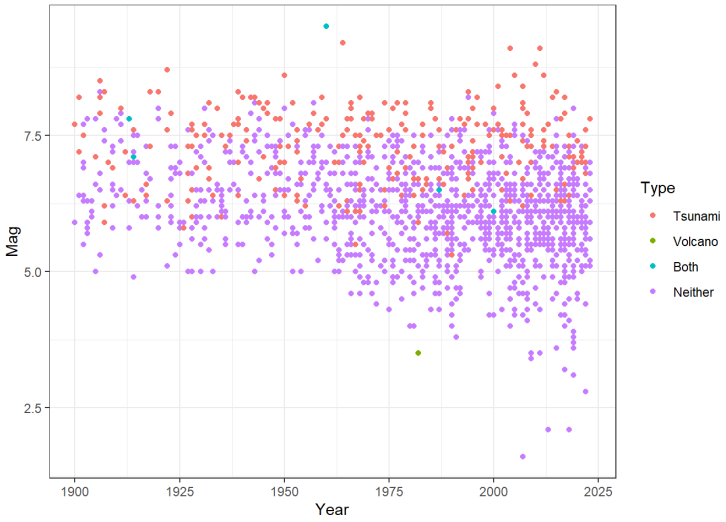
A Support Vector Machine in short SVM can be thought of as a surface where a line is made between the data points. In a perfect dataset, is is possible to create a line between different groups. The goal is to divide the space in equal homogeneous partitions. The data can be separable which is the simplest case of an SVM. However, an SVM can also be used if the data is not linearly separable.

8.2 Research Question SVM

The question to answer in this case is how the data can be splitted regarding the different types. There are earthquakes with Tsunami and Volcano eruptions. To make the classification even clearer, a new column type is created with the identification if there was a Tsunami, Volcano, Neither or Both.

8.3 Training a SVM

Magnitude and Type per year In the first step, it is necessary to look at the data to be separated. The focus is on the Magnitude per earthquake in each year. The types can be seen by the different colors.

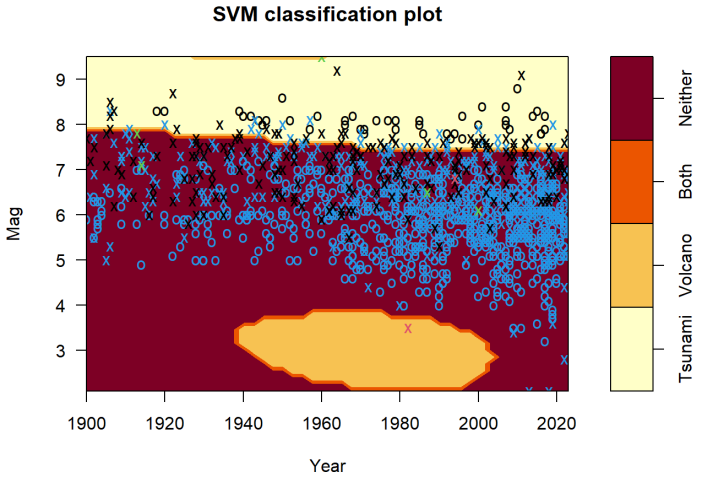


This plot shows that it is not possible to linearly separate the data points. So, in a further step, the data is partitioned and put into train and test data. With this, a model can be trained to show a support vector machine with the respective categories.

SVM Type Code (interactive drop down button)

```
##
## Call:
## svm(formula = Type ~ Mag + Year, data = train, kernel = "radial",
##     cost = 10, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  10
##
## Number of Support Vectors:  455
##
##  ( 246 202 5 2 )
##
##
## Number of Classes:  4
##
## Levels:
##  Tsunami Volcano Both Neither
```

The summary reveals the number of classes and how many observation belongs to each class. The number of Support Vectors is relatively high if we take into consideration that there were 1469 observations in total used in the beginning.



The SVM classification plot shows that the most points belong to the group Neither. That means that neither a Tsunami nor a Volcano occurred when the earthquake happened. The type Both barely exists in this plot.

Cross Validation

To make sure that what we see in the plot is correct, a cross validation is used. For the cross validation, the data is divided into 5 subgroups and then tested for the accuracy. The model will be trained and evaluated 5 times and then the accuracy will be displayed.

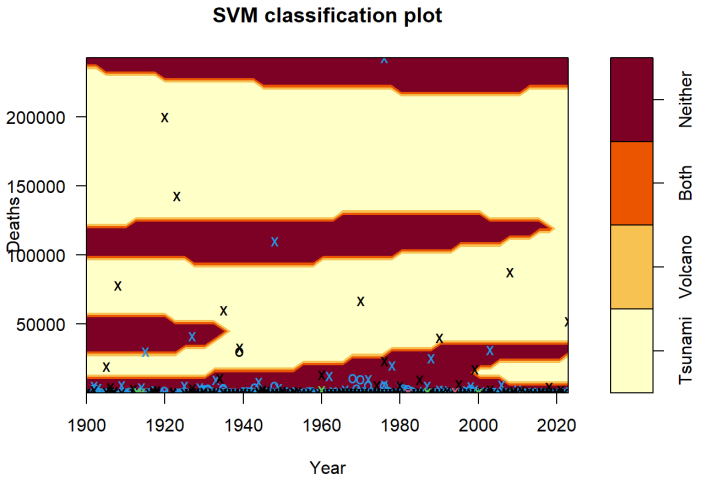
Cross Validation SVM1 Code (interactive drop down button)

```
##
## Call:
## svm(formula = Type ~ Mag + Year, data = train, kernel = "radial",
##     cost = 10, cross = folds, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  10
##
## Number of Support Vectors:  455
##
##  ( 246 202 5 2 )
##
##
## Number of Classes:  4
##
## Levels:
##  Tsunami Volcano Both Neither
##
## 5-fold cross-validation on training data:
##
## Total Accuracy: 84.01279
## Single Accuracies:
##  84.8 82 84.8 85.6 82.86853
```

This gives a total accuracy of 84%. That means that in total 84% of the data points could be added to the right group.

Deaths and Type per year
To have another look at the Types of the earthquake, a plot with the deaths per year was created.

```
##
## Call:
## svm(formula = Type ~ Deaths + Year, data = train, kernel = "radial",
##     cost = 10, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  10
##
## Number of Support Vectors:  588
##
##  ( 336 245 5 2 )
##
##
## Number of Classes:  4
##
## Levels:
##  Tsunami Volcano Both Neither
```



This shows that in most of the cases, not many deaths are predicted to happen. And in most cases, the earthquake will neither include a Tsunami or Volcano eruption.

This information can be used when thinking about building close to a coast.

Cross Validation SVM2
To get an idea of the accuracy of this second support vector machine, the cross validation is considered again.

```
##
## Call:
## svm(formula = Type ~ Deaths + Year, data = train, kernel = "radial",
##     cost = 10, cross = folds, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  10
##
## Number of Support Vectors:  588
##
##  ( 336 245 5 2 )
##
##
## Number of Classes:  4
##
## Levels:
##  Tsunami Volcano Both Neither
##
## 5-fold cross-validation on training data:
##
## Total Accuracy: 79.45643
## Single Accuracies:
##  83.6 76 78.8 80.4 78.48606
```

The accuracy of this second support vector machine is slightly less than in the first. 79% of the trained data was added to the right classes. The number of Support vectors also got higher and is now 588.

8.4. Conclusion SVM

Within this whole evaluation it is made clear that the majority of earthquakes result in neither a Tsunami nor a Volcano. There are 455 and 588 support vectors in our models which lie on the margin or even violate the margin of the identifier. It clearly is hard to find a good fit for the separation of these points. Looking at the Magnitude per year frame, it can be seen that the earthquakes with a higher Magnitude could lead more likely to a Tsunami. When looking at the deaths per year, it is clear that earthquakes are most likely to not result in many deaths and belong to the class neither.

However, it would be necessary to look at more data to give a really precise prediction.

9. Optimisation Problem

General optimization problems that can be addressed with this data set revolve around the risk mitigation regarding earthquakes.
Specifically, these include strategies that can be developed in order to minimize damage and causalities by predicting locations and strengths of future earthquakes. Parts of the previous analysis have indicated that the data set shows higher availability in records in the last 70 to 50 years. As it has been confirmed that modern earthquake measurement systems haven't been widely used until before the named time period, it can be said that earthquake predictions are a field of study today that necessitates more data than available to this day. Highly accurate or reliable earthquake predictions are therefore not yet available.

Furthermore, as the significant earthquakes data set contains records of earthquakes that had a significant strength and/or after-effects, it can be a sensible choice to further research and combine any other available earthquake records. In this way, a larger base to forecast the time frame of future earthquakes could be given.
In connection to extending the data base, further research of relevant variables to take into consideration have been named in previous parts of the analysis.

The significant earthquakes database gives indication on after-effects of an earthquake which also serves the opportunity to recommend conducting analyses for the optimization of infrastructure design. To mitigate damage and causalities, it is crucial to investigate earthquake prone regions and to provide a stable and suitable infrastructure to minimize overall damage. However in chapter 5, it has been determined that "unnatural" earthquakes occurrences caused by, e.g, explosions, are also present in the data base. Unfortunately, these cannot be specifically identified. Also, an important information missing is the estimated present population at an earthquake's epicenter. Both of these characteristics of the data should be seen as a disclaimer when attempting to fit a model for any prediction with this data.

Similar to the previous point, another chance for optimization is the allocation of resources for seismic monitoring. As the world map in chapter 2.1. has shown, significant earthquakes follow a specific pattern that can be identified as high friction points between tectonic plates. Within these regions, it is of high importance to situate technologically advanced seismic monitoring technologies in a way that covers most, if possible all, of the high risk zones. Therefore, it can be sensible to conduct further analyses to determine the optimal allocation of these resources.

Concludingly, the significant earthquake database poses various opportunities to optimize the risk mitigation of earthquake after-effects. However, as elaborated, the data base alone should not be used to produce prediction models as it also contains sparse and incomplete information. Therefore, determining additional data that adds value to this extensive data base could bring promising outcomes.

10. Conclusion

As showcased in the preliminary graphical analysis fitting a linear model indicated that there is a strong evidence that longitude, latitude, focal depth and year have an effect on the dependent variable, magnitude of an earthquake. There was no indication having an interaction between longitude and latitude. However the comparison of our linear models has shown a high unexplained variance which lead to a further investigation with more complex statistical models.

In chapter 4, going deeper into the analysis of the earthquake magnitude, it has been identified that the magnitude shows a slight increase of 0.05% if the focal depth of the epicenter is increased by 1 kilometer. Furthermore, it could be seen that countries Japan, Mexico, Mongolia, Taiwan, and Turkmenistan show strong statistically significant patterns regarding a strong magnitude of earthquakes. Also, the model could confirm that Central and Northern Europe have on average earthquakes at a lower magnitude. Overall, the model evaluation has shown that the available information in the data set do not suffice to create a high performing model. It has been recommended to research on further variables that presumably have an influential role on a magnitude. These variables could be the intrinsic quality, the rupture area, the average displacement across the rupture area, and the directivity of the earthquake.

Concerning the attempt to develop a model to predict causalities associated with earthquakes, it could be revealed that increasing the magnitude by 1 unit result in an increased death count with a factor of 5.7. Regarding the Damage Description, it was gained the insight that the count of death is increased at a factor of 128 in relevance to a Damage Description of 1. Also, it has been shown that Haiti shows the significantly highest death count in case of an earthquake. However, the evaluation of the model showed that a well-suited model to predict deaths could not be produced with this data set. Therefore, more extensive data about factors that have an influential role on the death count should be considered: amount of inhabitants in the affected region, quality of infrastructure in the region, accessibility of rescue operations, and average age of inhabitants to map the chances of survival.

Modelling the odds of Death Descriptions with regards to Magnitude, Country and Damage Description, it is revealed that having an increased Magnitude by one unit, it is 3.6 times more likely to have a Death Description of 3 or 4 (DD 3-4). (101-1000 and over 1000). Regarding the country, the odds for having DD 3-4 in Turkey, Iran, El Salvador and Italy have been determined to be significantly higher than in Greece. The total

performance of the binomial model has been assessed with over 70%. During the analysis however, it has been detected that the previously named unnatural earthquakes are present in the data base and could greatly falsify any prediction about the Death Descriptions and therefore, prediction which regards to this variable are not recommended within this data set.

The cross validation of two General Additive Models (GAMs) has revealed a strong evidence that magnitude has a positive linear relationship with the death numbers caused by an earthquake and proves a weak evidence of a non-linear effect of longitude as well as focal depth.The fitted model revealed a rather low influence of higher intensity levels and several regional indications on the Death Description variable. The GAM model fit resulted only 8.7% explaining the overall variability of the Death Description variable, therefore it can be concluded that it is not reliable to make realistic predictions for future impacts caused by an earthquake.

The Neural Network predictions on the deaths and damage description are difficult to use for further analysis. The accuracy was at any point under 50% which is not enough to make meaningful interpretations. However, it shows that a moderate damage description is most likely.

The Support Vector Machine made it possible to separate the data into the 4 types Tsunami, Volcano, Both, Neither. As already mentioned above it was possible to train the data and after cross validation we got an accuracy of around 80% for both models. These predictions show that most earthquakes with high magnitude were classified to the group Tsunami. At places where the magnitude is the highest the chances of a Tsunami happening are bigger. The deaths are in most of the cases not that high. It can also be said that with neither a Tsunami nor a Volcano less deaths happen.

In conclusion, it has been determined that the earthquake database is highly informative but shows great sparsity with regards to the completeness of the data set and the detailed information about the nature of the events. The considerable amount of missing values made it challenging to fit reliable statistical models. Removing the missing values in certain cases has reduced the number of records considerably which has impacted the outcomes and made the predictive models less reliable. It was also noted that many other political or socio-economical factors are not contained in this data set which however may contribute to improve the future predictions in terms of estimated death and damage. Therefore, the data base alone should not be used to produce prediction models but should rather be considered as an additional source to a more extensive and complete data set.