Impact of Natural Disasters in Canada

Aigerim Kulzhabayeva

[1]Department of Mathematics and Statistics, York University

Abstract. *This report outlines the analysis of data from The Canadian Disaster Database to estimate the impact of natural disasters that occurred in Canada between 1900-2017. The data was analyzed using the GLS model and it was found that as the number of natural disasters in Canada continues to increase the impact on the lives of Canadians is not proportional in terms of death, injuries and utilities affected. However, we see a significant increase in the number of people evacuated, which may suggest that the impact is better mitigated through Natural Disaster management practices.*

1. Introduction

This report outlines the analysis of data from The Canadian Disaster Database (CDD) which contains detailed information on natural disasters that have happened in Canada between 1900-2017. In this report we will be estimating the impact of natural disasters by answering the following questions:

- How has the impact of natural disasters changed over time in Canada?
- Has the impact been the same across all provinces or regions of Canada?
- What might be the impact of natural disasters on Canadians in the future?

FATALITIES	The number of people killed due to a specific event.
INJURED.INFECTED	The number of people injured/infected due to an event.
EVACUATED	The number of individuals evacuated by the government of Canada due to a specific event.
NORMALIZED.TOTAL.COST	The Consumer Price Index (CPI) is used to normalize the financial data.
ESTIMATED.TOTAL.COST	A roll-up of all the costs listed within the financial data fields for a specific event
FEDERAL.DFAA.PAYMENTS	The amount, in dollars, paid out by Federal Disaster Financial Assistance Arrangements (Public Safety Canada) due to a specific event.
PROVINCIAL.DFAA.PAYMENTS	The amount, in dollars, paid out by Provincial Disaster Financial Assistance Arrangements due to a specific event.
PROVINCIAL.DEP.PAYMENTS	The amount, in dollars, paid out by a Province or Territory due to a specific event.
MUNICIPAL.COSTS	The cost, in dollars, to a Municipality due to a specific event.
OGD.COSTS	The cost, in dollars, as reported in Open Government Data sources.

INSURANCE.PAYMENTS	The amount, in dollars, paid out by insurance companies due to a specific event.
NGO.PAYMENTS	The amount, in dollars, paid out by a Non-Governmental Organization due to a specific event.
UTILITY.PEOPLE.AFFECTED	The amount of people whose utility services (power, water, etc.) were interrupted/affected by a specific event.
EVENT.SUBGROUP	The subgroup of disaster (Biological, Geological, Meteorological-Hydrological) that occurred.
Province.Territory	Province or territory in which the event took place.
PLACE	The city, town or region where a specific event took place.
MAGNITUDE	A measure of the size of an earthquake, related to the amount of energy released.
EVENT.START.DATE	The date a specific event started.

Table 1. Variables in the Data Set Natural Disasters

We define the following six variables, as *impact variables* and will use them as a response to measure the impact of the natural disaster over the years. The rest of the variables will be treated as covariates.

- FATALITIES
- INJURED.INFECTED
- EVACUATED
- NORMALIZED.TOTAL.COST
- UTILITY.PEOPLE.AFFECTED

In the following section we will introduce the Generalized Least Squares (GLS) model which we will be using to model the data. In section 3 we will start the analysis by first preparing the data and looking at overall trends. Then we will look at the missing values and define our predictor variables for the rest of the analysis. In the following sections we answer the question whether the impact increased over time in Canada by analyzing impact variables one by one. We look for the trend via plots and GLS model. We then look if the impact has been the same across all provinces by looking at the subplots of provinces separately for each impact variable.

2. Methodology

The model used to analyze the data is Generalized Least Squares (GLS) model. This model is based on linear regression model where parameters are estimates using Generalized Least Squares instead of Ordinary Least squares. In the following subsections we will overview the linear regression with OLS estimators, then move on to linear regression using GLS estimators. Finally we will at how the GLS estimator can be found using R software.

2.1. Linear regression with OLS estimators

Recall liner regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon \sim N_n(0, \sigma^2 I) \tag{1}$$

where $\mathbf{y}$ is a response vector, $\mathbf{X}$ is a data matrix, $\boldsymbol{\beta}$ is a vector of parameters, $\boldsymbol{\epsilon}$ is a a vector of errors, n is the number of total observations. We can estimate the parameter of the model by minimizing the sum of squared residuals

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

taking the derivative with respect to β we get the estimate of the parameter

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Under the following assumptions the above OLS estimator of β is Best Liner Unbiased Estimators (BLUE), meaning out of all linear unbiased estimators they have the minimum variance and therefore be most efficient.

1. The relationship between the response variable and predictor must be linear, and
2. $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I)$

which means that the errors have to be independent, identically distributed and have the same variance. However in some situation, like we had in this analysis, the errors $\epsilon_i, i = 1, ..., n$ do not have the same variances. This means that the parameters estimated are no longer BLUE and do not have the smallest variance. When this happens we can use a technique called Generalized Least squares which transforms the model to a new set of observations which satisfy the constant variance assumption, and the use the OLS technique on the transformed parameters.

2.2. Linear regression with GLS estimators

Let the linear regression model be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon \sim N_n(0, \sigma^2 V) \quad \text{where} \tag{4}$$

where $\mathbf{y}$ is a response vector, $\mathbf{X}$ is a data matrix, $\boldsymbol{\beta}$ is a vector of parameters. $\boldsymbol{\epsilon}$ is a a vector of errors and V symmetric, non-singular matrix. Then $V = KK^T$ by Choleski decomposition and we define

- $\mathbf{y}' = K^{-1}\mathbf{y}$
- $\mathbf{X}' = K^{-1}\mathbf{X}$
- $\boldsymbol{\epsilon}' = K^{-1}\boldsymbol{\epsilon}$

Then for the new model

$$\begin{aligned} E(\boldsymbol{\epsilon}) &= E(K^{-1}\boldsymbol{\epsilon}) \\ &= K^{-1}E(\boldsymbol{\epsilon}) = 0 \\ \\ Var(\boldsymbol{\epsilon}) &= Var(K^{-1}\boldsymbol{\epsilon}) \\ &= K^{-1}Var(\boldsymbol{\epsilon})(K^{-1})^T \\ &= K^{-1}\sigma^2 KK^T K^{-1} \\ &= \sigma^2 I \end{aligned} \tag{5}$$

The GLS estimator can be obtained to minimizing Squared Sum of Residual as in equation 6.

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \boldsymbol{(\epsilon')^T\epsilon} \\ &= (\mathbf{y}' - \mathbf{X}'\hat{\boldsymbol{\beta}})^T(\mathbf{y}' - \mathbf{X}'\hat{\boldsymbol{\beta}}) \\ &= (K^{-1}\mathbf{y} - K^{-1}\mathbf{X}\hat{\boldsymbol{\beta}})^T(K^{-1}\mathbf{y} - K^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T K^{-1}K^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T V^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T V^{-1}\mathbf{y} - 2\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} + \hat{\boldsymbol{\beta}}\mathbf{X}^T V^{-1}\hat{\boldsymbol{\beta}}\mathbf{X} \end{aligned} \tag{6}$$

taking derivative with respect to parameters we get

$$\frac{\partial SSR}{\partial\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T V^{-1}\mathbf{y} + 2\mathbf{X}^T V^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \tag{7}$$

then,

$$\hat{\beta}_{GLS} = (\mathbf{X}^T V^{-1}\mathbf{X})^{-1}\mathbf{X}^T V^{-1}\mathbf{y} \tag{8}$$

The estimator $\hat{\beta}_{GLS}$ is unbiased

$$
\begin{aligned}
E\left[\hat{\boldsymbol{\beta}}_{GLS}\right] &= E\left[\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\mathbf{y}\right] \\
&= E\left[\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\right] \\
&=E\left[\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta}+\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\right] \\
&=E\left[\boldsymbol{\beta}+\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\right] \\
&=E[\boldsymbol{\beta}]+\left(\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\right)E[\boldsymbol{\epsilon}] \\
&=\boldsymbol{\beta}+\left(\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\right)(0) \\
&=\boldsymbol{\beta}
\end{aligned}
\tag{9}
$$

and the variance of $\hat{\beta}_{GLS}$ is

$$
\begin{aligned}
\hat{\beta}_{GLS} &= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\mathbf{y} \\
&= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta}+\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon} \\
\hat{\beta}_{GLS}-\boldsymbol{\beta} &= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_{GLS}\right) &= E\left[(\hat{\boldsymbol{\beta}}-E[\hat{\beta}])^2\right] \\
&= E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^2\right] \\
&= E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\right] \\
&= E\left[\left(\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\right)\left(\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\right)^T\right] \\
&= E\left[\left(\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\right)\left(\boldsymbol{\epsilon}^T\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\right)\right] \\
&= E\left[\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\epsilon^T\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\right] \\
&= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}E\left[\boldsymbol{\epsilon}\epsilon^T\right]\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\operatorname{Var}(\boldsymbol{\epsilon})\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\sigma^2\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}
\end{aligned}
\tag{11}
$$

This means that the GLS parameters $\hat{\beta}_{GLS}$ are BLUE. In practice the matrix V is unknown and needs to be specified. In the following subsection we explore how to specify V matrix in R.

2.3. GLS in R

The nlme library provides variance functions (varFunc) that can be used to specify the variance structure in the model.

Function **Form**	**Explanation**
varFixed Var(ϵ_i) $= \sigma^2 \times FATAL_i$	fixed variance
varIdent Var(ϵ_i) $= \sigma_j^2$	different variances per stratum
varPower Var(ϵ_i) $= \sigma^2 \times \mid FATAL_i \mid^{2\delta}$	power of covariate
varExp Var(ϵ_i) $= \sigma^2 \times e^{2\delta \times FATAL_{ij}}$	exponential of covariate
varConstPower Var(ϵ_i) $= \sigma^2 \times (\delta_1 + \mid FATAL_{ij} \mid^{\delta_2})^2$	constant plus power of covariate
varComb Var(ϵ_i) $= \sigma_j^2 \times e^{2\delta FATAL_{ij}}$	combination of variance functions

Table 2. Standard varFunc classes functions

The two main arguments in the varFunc are *value* and *form*. The first specifies the values of the variance parameters δ. The second is a one-sided formula which specifies the variance co-variate v, and optionally, a stratification variable for the variance parameters such that every level of co-variate will have different variance parameters. As an example lets use the variables that we will meet in our analysis.Let variable FATAL be a numeric variable with observations $i = \{1, ..., n$, representing number of fatalities as a result of a disaster. Let EVENT.GROUP be a factor representing subgroup of the disaster with levels $j = \{1, 2, 3\}$- biological, hydrological and geological.

Function in R	**Form**
varFixed ($\sim FATAL$)	Var(ϵ_i) $= \sigma^2 \times FATAL_i$
varIdent (from=$\sim 1 \mid EVENT.SGR$)	Var(ϵ_i) $= \sigma_j^2$
varPower(form = $\sim FATAL$)	Var(ϵ_i) $= \sigma^2 \times \mid FATAL_i \mid^{2\delta}$
varExp(form= $\sim FATAL \mid EVENT.SGR$)	Var(ϵ_i) $= \sigma^2 \times e^{2\delta \times FATAL_{ij}}$
varConstPower(form=$\sim FATAL$)	Var(ϵ_i) $= \sigma^2 \times (\delta_1 + \mid FATAL_{ij} \mid^{\delta_2})^2$
varComb(varIdent(form =$\sim 1 \mid EVENT.GR$), varExp(form = $\sim FATAL$))	Var(ϵ_i) $= \sigma_j^2 \times e^{2\delta FATAL_{ij}}$

Table 3. Standard varFunc classes functions

3. Data Analysis

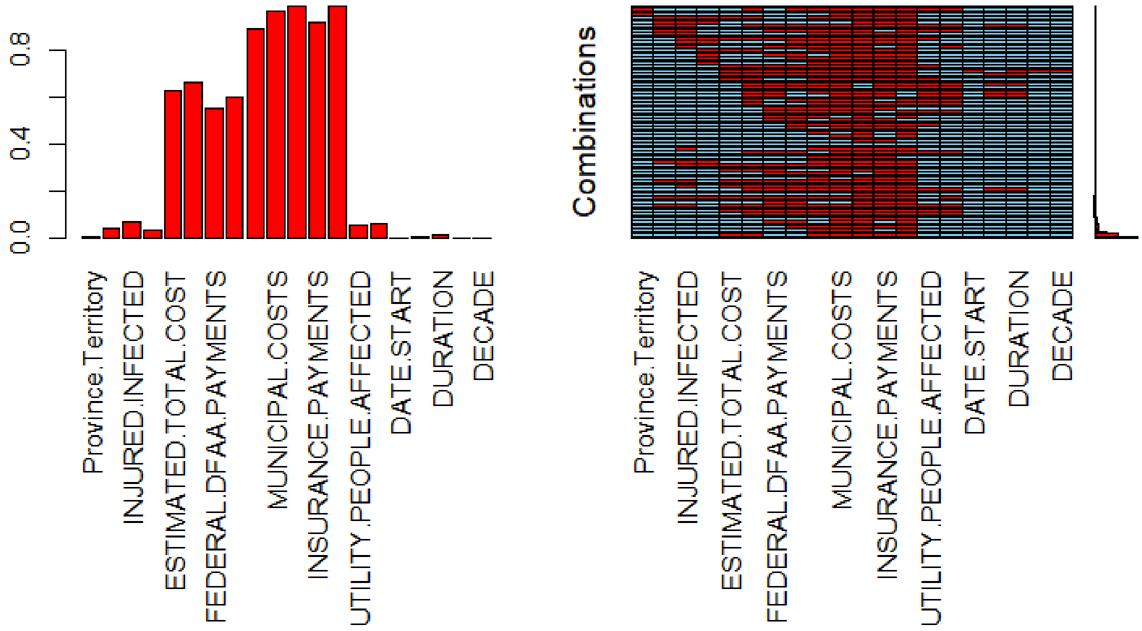
3.1. Data Preparation

During the initial investigation of the data using the xqplot() function in spida2 package we have identified the following problems

1. Variables “EVENT SUBGROUP”, “EVENT TYPE”, “PLACE”, “Province/Territory” are not of appropriate class and need to be adjusted.
2. Even start and end date variables are characters, from these two variables we could extract a single variable of “DURATION” to see if the duration of events has changed over time. These variables are strings we need to see what these numbers mean.
3. Variables of “FATALITIES” and “INJURED.INFECTED” have quite significant outliers which will need to addressed either by removal or incorporation into the model. We will be addressing these in subsections for the individual impact variables.
4. Variable associated with cost have a significant number of missing values.
5. Since we are working on the data set that spans a century, we will round the dates by creating YEAR and DECADE variables.

3.2. Missing values

It was evident for the initial exploration that the data set contains significant amount of missing values.

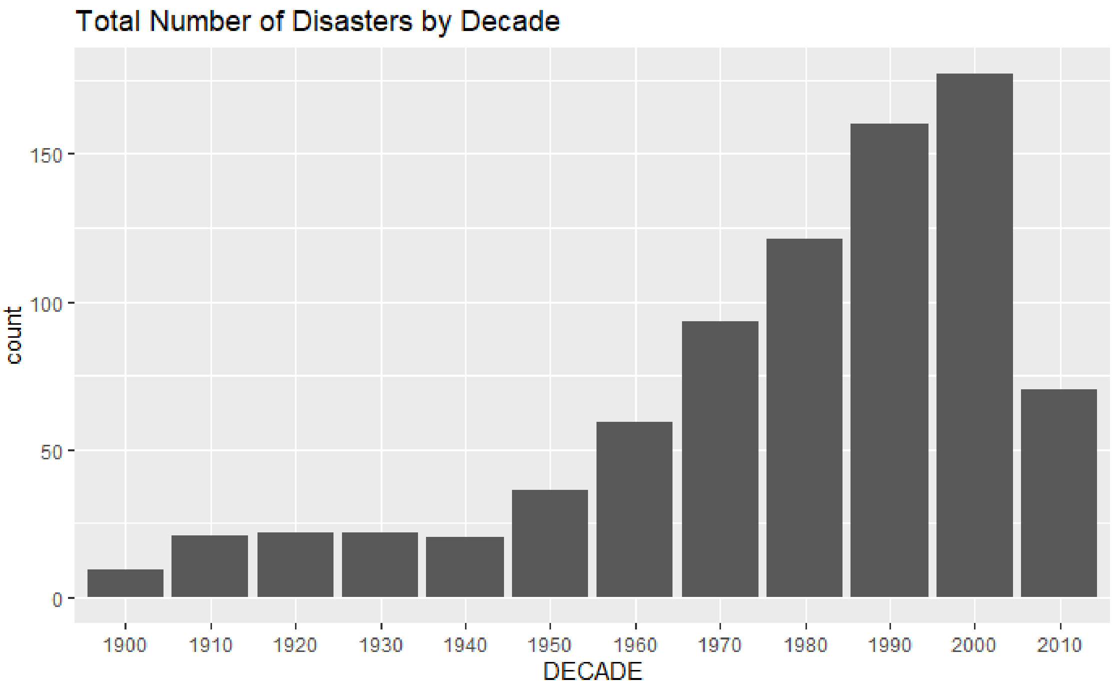


From the figure above it is apparent that the variables with most missing values are the variables associated with the cost of the disaster. Since the proportion of missing values exceeds 0.7, the values will can not be imputed. In this analysis it was decided to disregard these variables due to their unreliability. The variable of NORMALIZED.TOTAL COST,which is a cumulative of all costs will be analyzed in section 3.7. The following variables with smaller number of missing values were analyzed and the values were manually imputed using the other variables in the data set.

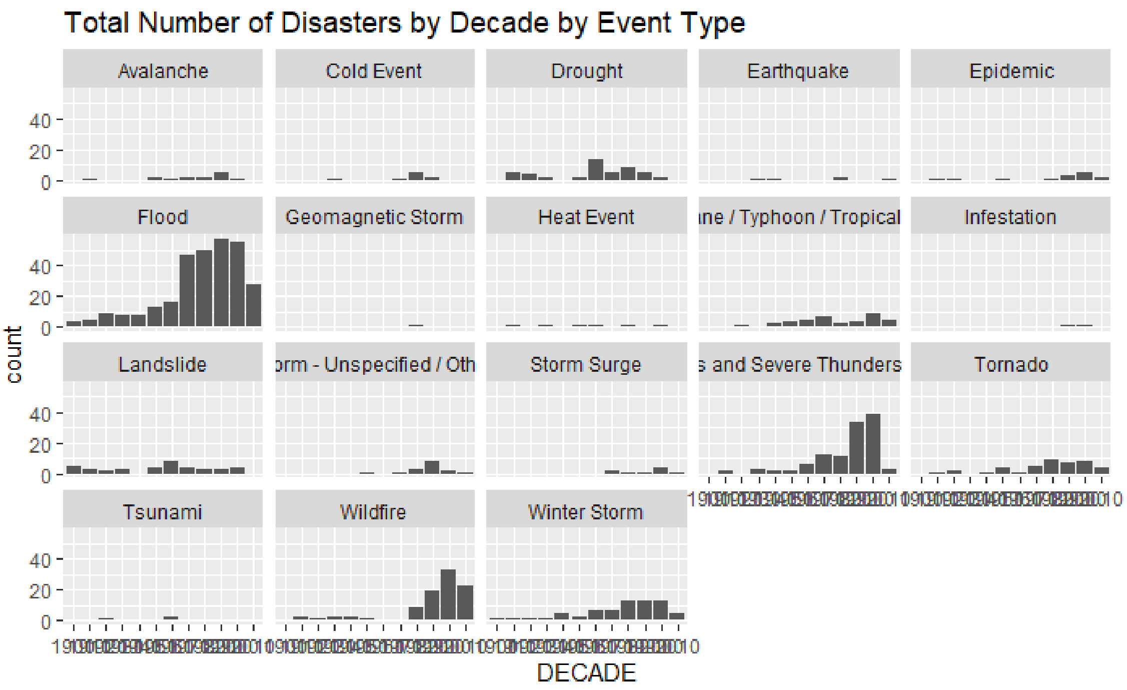
- Variable Province.Territory has 6 missing values that may be manually imputed from variable PLACE which has no missing values.
- Variable YEAR has one missing value which is a winter storm in Quebec. We impute this value using the DATE.END which is present in the data.

3.3. Exploratory analysis.

In this subsection we will plot the overall number of disasters over the decades and see if we can identify a trend.



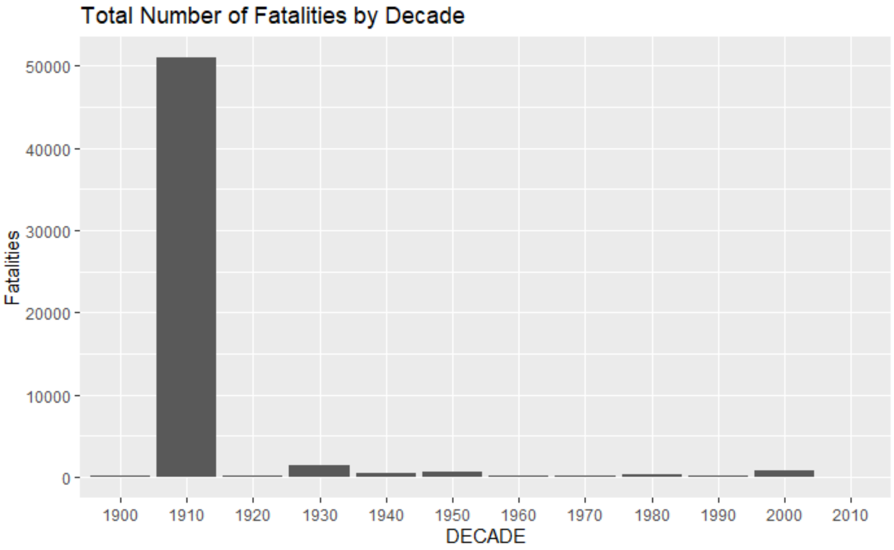
Based on the histogram above we see that overall number of disasters have been increasing. Since the variable DECADE has no missing values, based on data, this number should be accurate. To see which particular type of natural disaster has been increasing we will create subplots for the types of disasters.



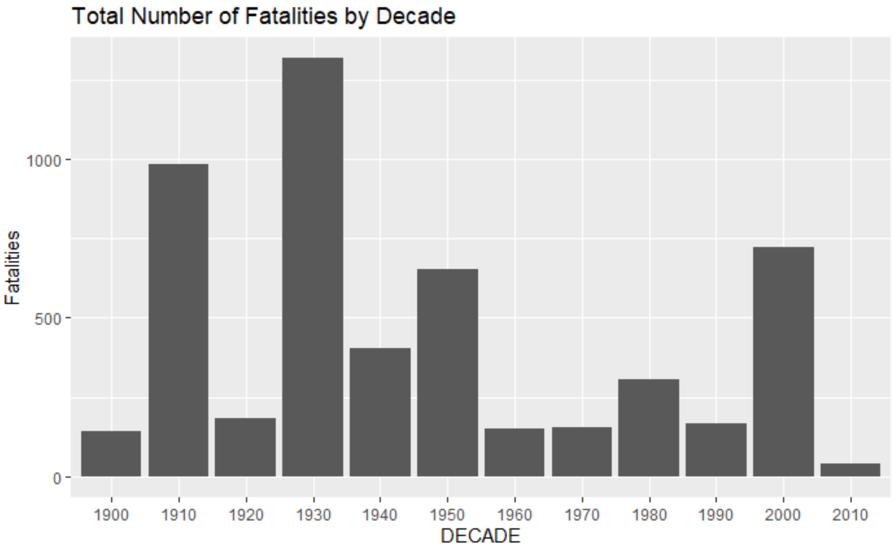
It seems that the types of event responsible for most of the increase are floods,Thunderstorms and Wildfires. To a lesser degree: winter storms, epidemics and tornados. In the next subsection we will look at each impact variable and identify whether the impact increased over the years. The nature of that data set is such that most of the relationships between the variables can be identified via two dimensional plots. As such plots will be a primary tool for identifying the trend, as a secondary tool we will be using a generalized least square model to fit the response values.

3.4. Fatalities

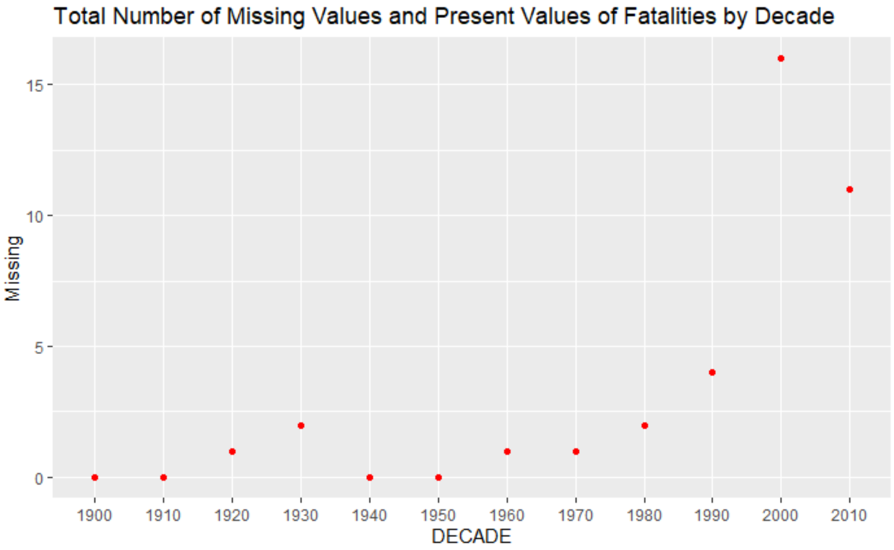
In this section we will focus on the variables FATALITIES as our outcome variable of impact. We will forst plot the total number of fatalities per decade to observe the general trend.



From the above plot we see a significant outlier of fatalities which prevents us from seeing the rest of the values on the plots. This outlier corresponds to the Spanish Flu pandemic which happened in Canada from 1918-1920. We are removing this outlier and will add it back as needed. Below is the same graph without the outlier.



In the graph without the outlier we see the number of fatalities has not been increasing proportionally to the number of natural disasters. Overall the number of fatalities seems to be decreasing. However, this maybe due to number of missing values which we plot below for comparison.



It seems that we have more missing values as the time progresses which may be the reason why the number of fatalities seems to be decreasing. At the same time it is less likely that fatalities in significant events have not been recorded. This means that any conclusion we will arrive by modelling this impact variable is not fully reliable due to the trend in missing values. For exploratory purposes we will take a look at the fatalities by looking at Subgroup of events and various Locations in separate graphs. We are grouping provinces by creating a factor with three levels: Local, Multiple Provinces, Across Canada.

3.4.1. Modelling

We will be modelling the variable FATALITIES as a response and variables YEAR as a covariate. The variance of the errors is modelled is weighted according to the exponential value of FATALITIES with separate parameters for each level of Location, since the event which happen across Canada have a lot more variability in fatalities.

```
model1<-gls(FATALITIES~YEAR+Location, data=df.fatal.model,
weights = varExp(form=~FATALITIES|Location))
```

The following is the output of the model

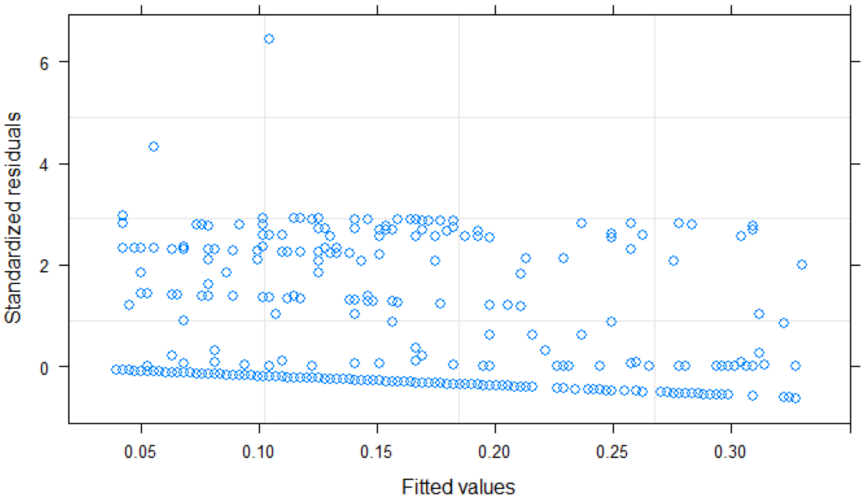
```
 Structure: Exponential of variance covariate, different strata
 Formula: ~FATALITIES | Location
 Parameter estimates:
        Local Across Canada
    0.2317444     0.1021661
```

Coefficients:				
	Value	Std.Error	t-value	p-value
(Intercept)	5.253033	1.8156280	2.893232	0.0039
YEAR	-0.002588	0.0009141	-2.831752	0.0048

Table 4. Model1 Coefficients

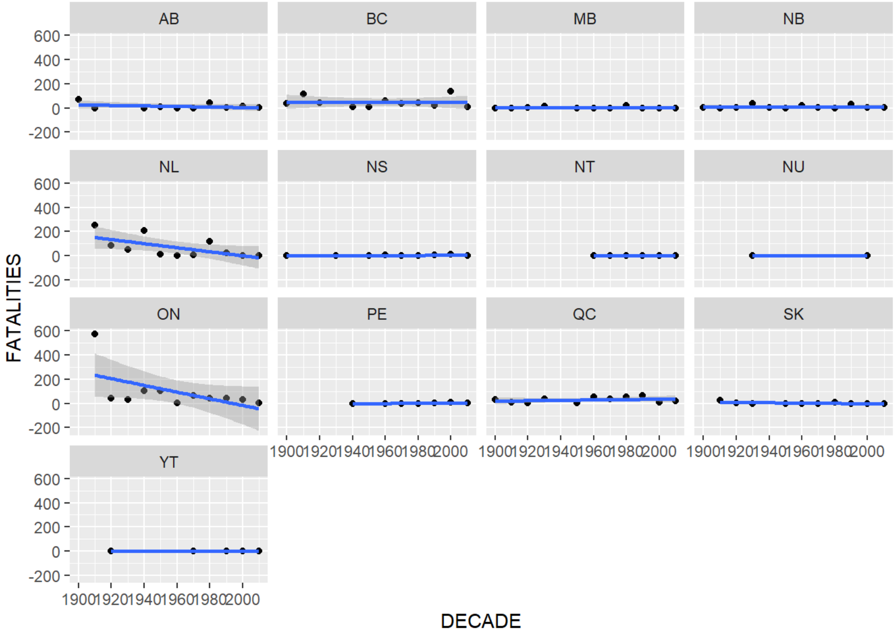
```
 Correlation:
              (Intr) YEAR
YEAR          -0.990
LocationLocal -0.151  0.010

Standardized residuals:
        Min          Q1         Med          Q3         Max
-0.62111246 -0.22999645 -0.14733156 -0.08317373  6.38258493

Residual standard error: 0.5278356
Degrees of freedom: 770 total; 767 residual
```

With the following plot of standardized residuals



The p-value for the variable YEAR is significant and confirms our initial conclusion based on the graph, the number of fatalities decreases over the years.

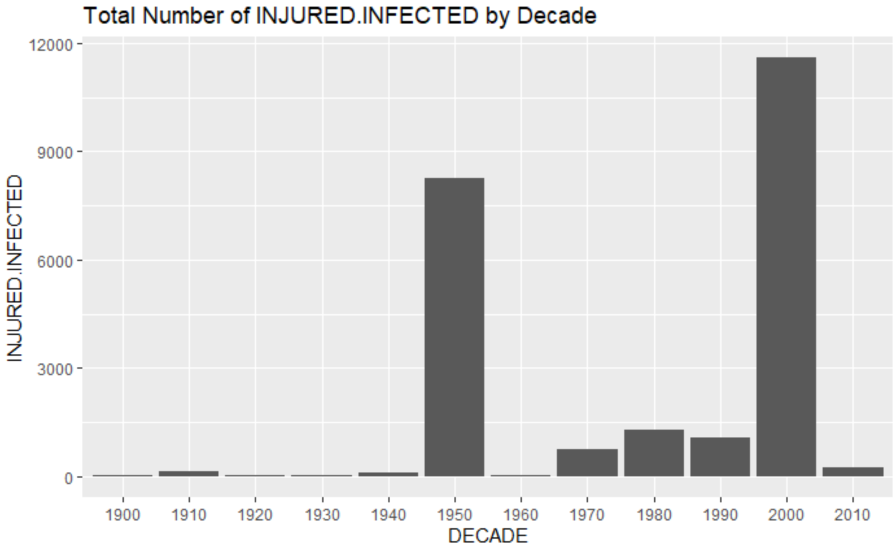
Looking at the number of fatalities per province below we see that the impact has not been the same.



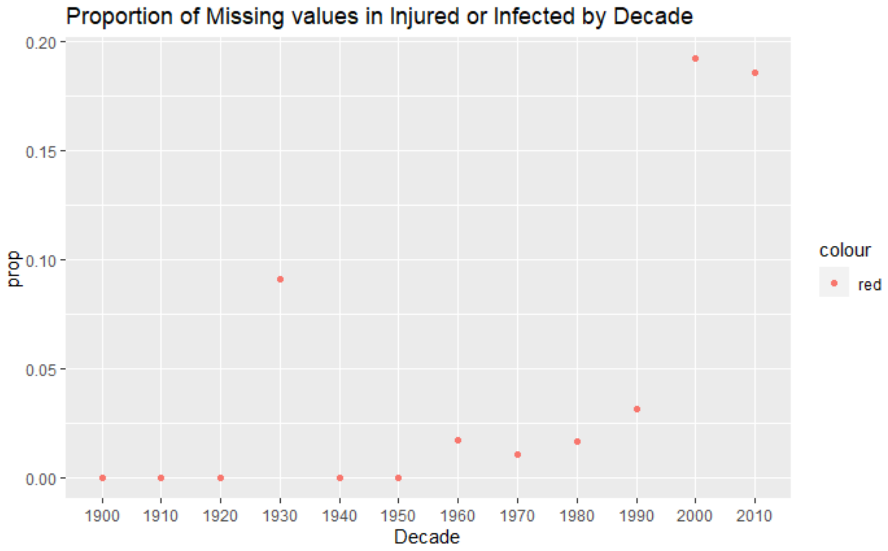
The provinces most impacted by the decline are Ontario and Newfoundland and Labrador.

3.5. Injured/Infected

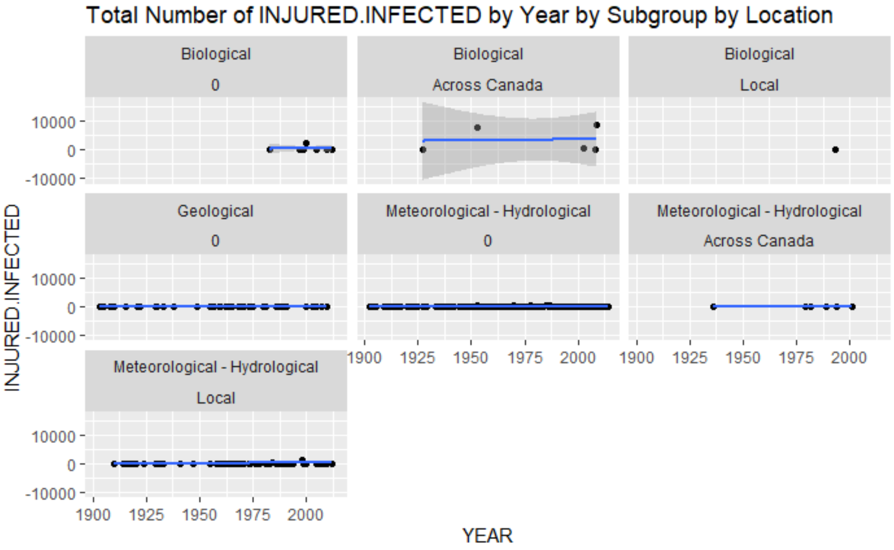
Following the steps in analysis of Fatalities, we start with looking at the general trend of the the number of people injured/infected as a function of decade.



It looks like the number of people injured/infected increases with time, and 1950 and 2000 being the decades with significantly higher number of infections/injuries. Next we proceed to the missing value to see if this trend can be attributed to them.



Just like with missing values for variable FATALITIES, the missing values are again more occurring towards the latter decades, which means that the total number can only increase towards the end. This means that the general trend is that the total number of people Injured/Infected is increasing. The following plot shows the total number of injured/infected by location and subgroup of events. This will help us define the model



We see most of the spread in the Subgroup of Biological event Across Canada, which happen to be epidemics. So from the graph we can say the the number of people infected/injured can be significantly affected by the increased number of epidemics. We will account for these by including them in the model.

3.5.1. Modelling

We will be modelling the variable INJURED/INFECTED as a response and variables YEAR as a covariatee. The variance of the errors is modelled is weighted according to the exponential value of INJURED/INFECTED with separate parameters for each level

Coefficients:				
	Value	Std.Error	t-value	p-value
(Intercept)	-1.1856858	1.8338959	-0.6465393	0.5181
YEAR	0.0006221	0.0009252	0.6724209	0.5015

Table 5. Model1 Coefficients

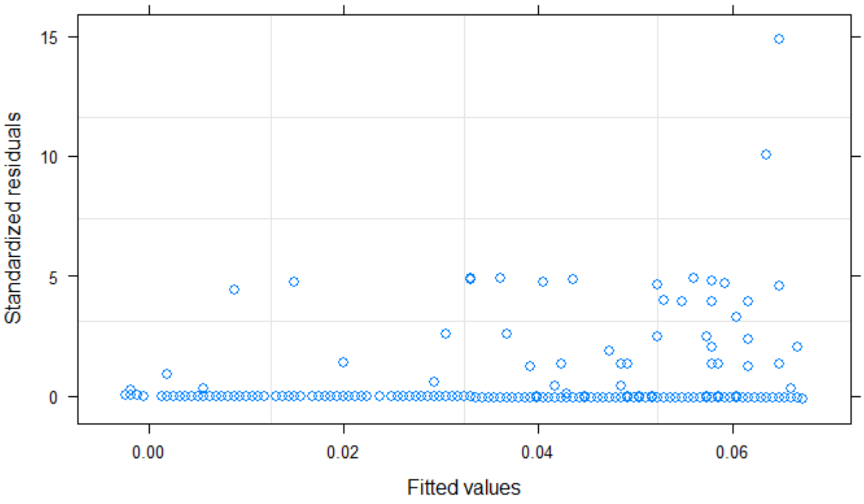
of EVENT.SUBGROUP, since we saw that the epidemics have a lot more variability in INJURED/INFECTED.

```
model21<-gls(INJURED.INFECTED~YEAR, df, na.action = na.omit,
weights = varExp(form=~INJURED.INFECTED|EVENT.SUBGROUP))
```

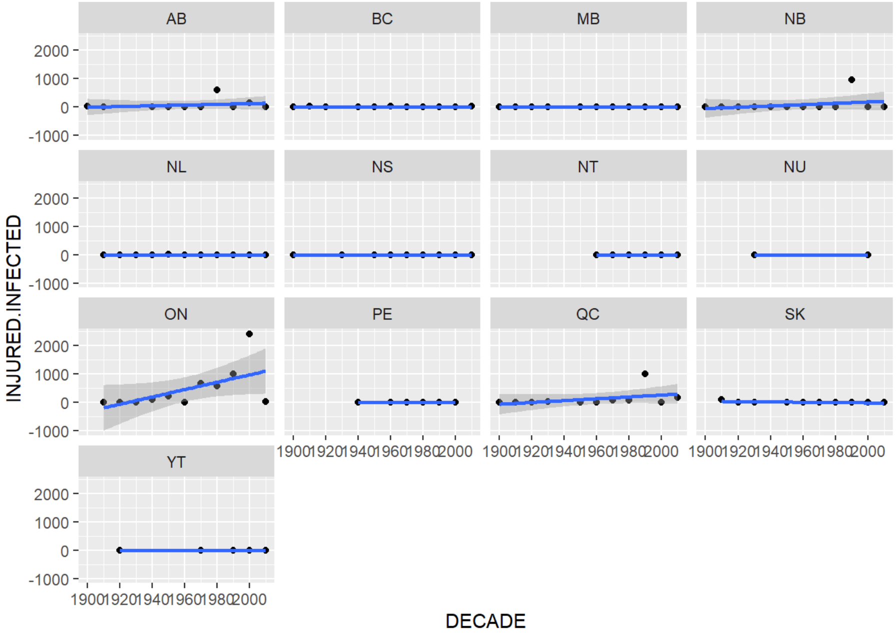
The following is the output of the model

```
Variance function:
 Structure: Exponential of variance covariate, different strata
 Formula: ~INJURED.INFECTED | EVENT.SUBGROUP
 Parameter estimates:
Hydrological   Geological   Biological
  0.11898491   0.03859589   0.21915004
```

[H] The following is the plot of the standardized residuals.



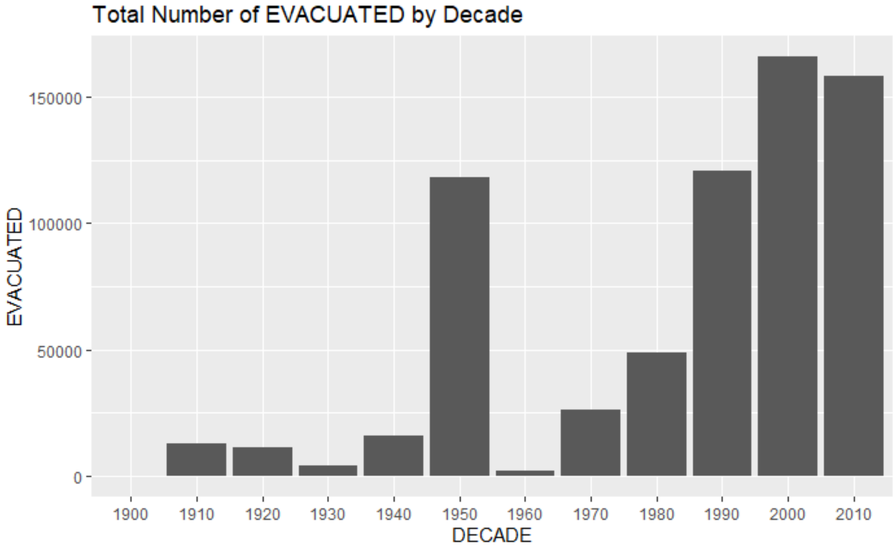
It seems that variable YEAR is not significant in this model. Looking at the subplots of provinces.



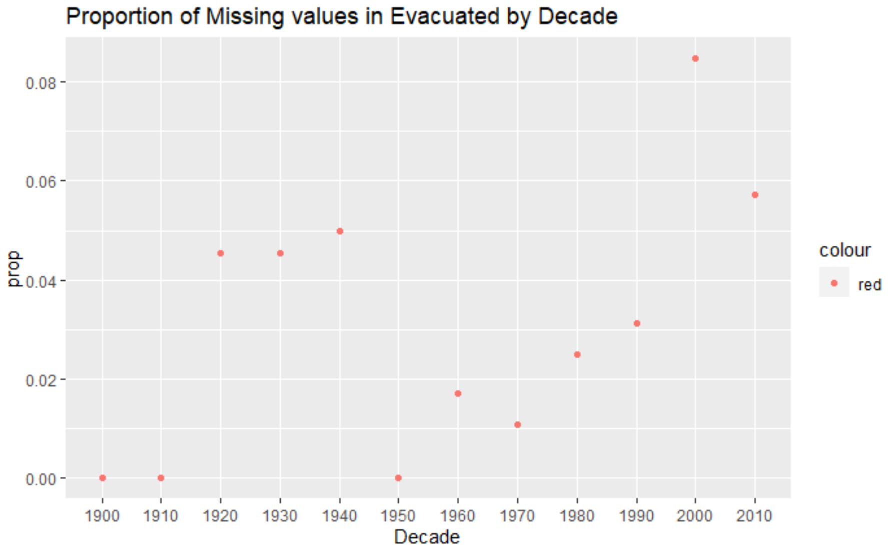
It seems that the number of people injured and infected increased the most in Ontario.

3.6. Evacuated

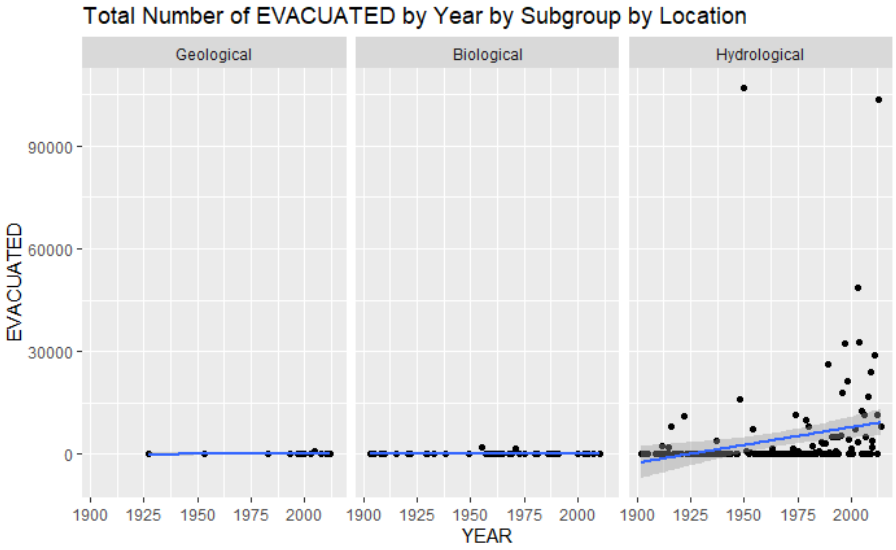
We will conduct the same analysis on the response EVACUATED.



The number of people evacuated over time is increasing which may be due to increased number of floods and wildfires that we saw in section 3.3. To check these number against missing values we plot the missing values below.



There seems to be more missing values in this variable, however the missing values are pretty sporadic relative to the other two. Moving on to total number of people evacuated by the event subgroup.



The last subplot corresponding to the number of people evacuated due to hydrological events is quite interesting and shows a prominent positive slope.

3.6.1. Modelling

We will be modelling the transformed, square root of EVACUATED variable as a response and variables YEAR and EVEN.SUBGROUP with three levels as a covariates. The variance of the errors is modelled is weighted according to the exponential value of YEAR with separate parameters for each level of EVENT.SUBGROUP, since we saw that the epidemics have a lot more variability in EVACUATED. In addition we have an interaction term of YEAR and EVENT.SUBGROUP.

```
model32<-gls(EVACUATED^(1/2)~YEAR*EVENT.SUBGROUP, evac.group,na.action
```

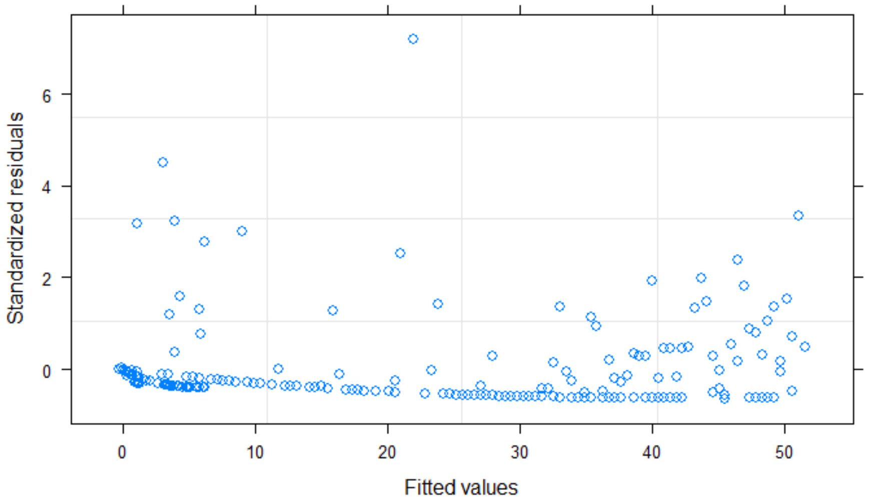
The following is the output of the model

```
Variance function:
 Structure: Exponential of variance covariate, different strata
 Formula: ~YEAR | EVENT.SUBGROUP
 Parameter estimates:
Meteorological - Hydrological                    Geological
                  0.010245471                   0.009440459

EVENT.SUBGROUP Geological                         -0.541  0.541
EVENT.SUBGROUP Meteorological - Hydrological      -0.246  0.246  0.133
YEAR:EVENT.SUBGROUP Geological                     0.534 -0.534 -1.000
YEAR:EVENT.SUBGROUP Meteorological - Hydrological  0.242 -0.242 -0.131
```

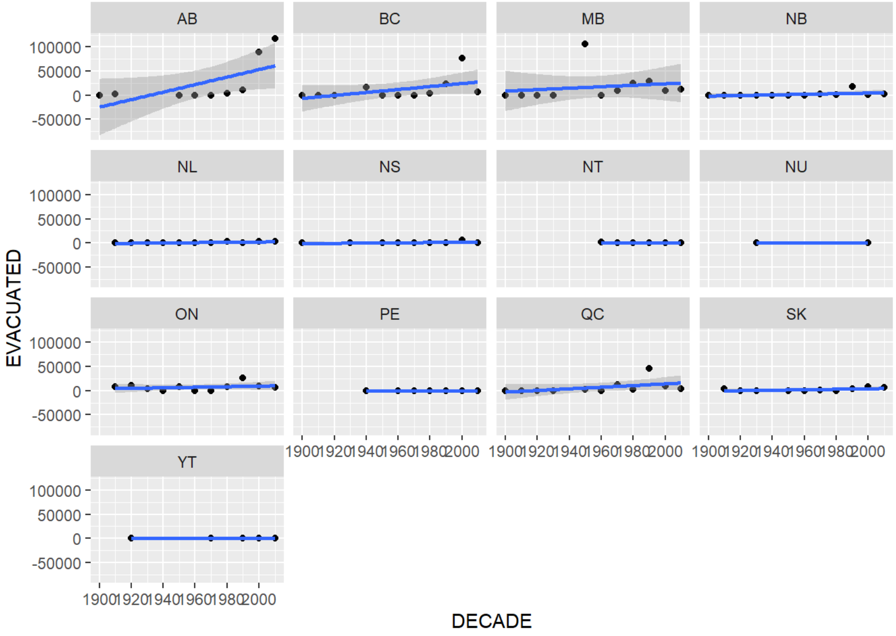
Coefficients:				
	Value	Std.Error	t-value	p-value
(Intercept)	-32.2273	57.84447	-0.557137	0.5780
YEAR	0.0167	0.02924	0.570870	0.5687
EVENT.SUBGROUPGeological	-77.4691	106.91264	-0.724602	0.4695
EVENT.SUBGROUPMeteor-Hydro	-848.0722	235.43899	-3.602089	0.0004
YEAR:EVENT.SUBGROUPGeological	0.0410	0.05480	0.748054	0.4553
YEAR:EVENT.SUBGROUPMeteor-Hydro	0.4460	0.12091	3.688922	0.0003

Table 6. Model Coefficients

From the table of coefficient we see that the interaction term of YEAR*SUBGROUPMeteorological - Hydrological is highly significant. The following is the plot of the standardized residuals of the model.



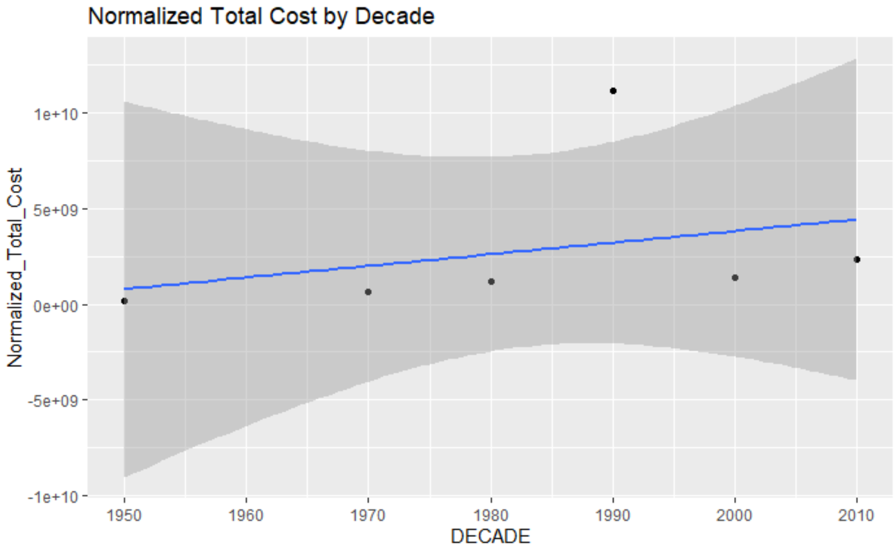
From the plots and the model we can conclude that the number of people evacuated increases for the Meteorological - Hydrological events. Looking at the subgraphs of the provinces below.



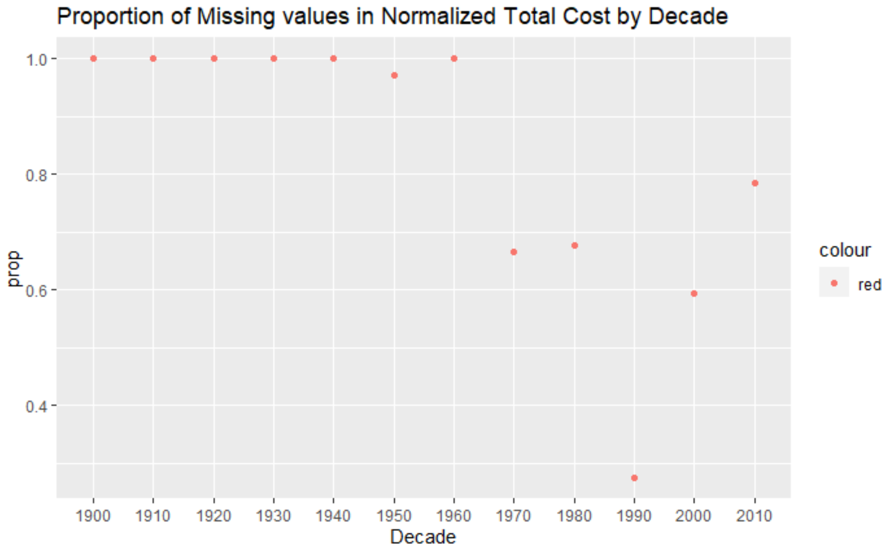
The number of people evacuated is mostly in Alberta, British Columbia and Manitoba, which is consistent with the types of disasters that have been increasing.

3.7. Normalized total cost

Now we turn to the financial impact of natural disasters. First we plot the normalized total cost variable and look whether the total dollar amount has increased over the decades.



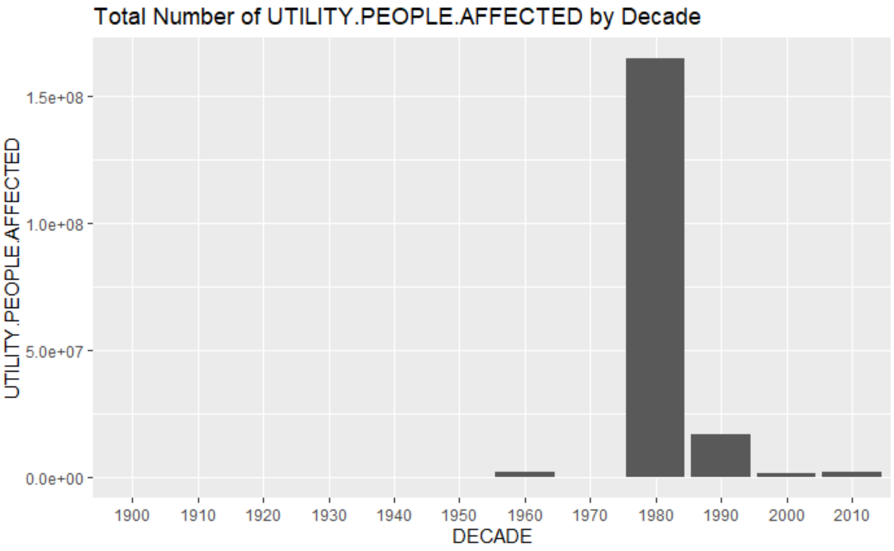
It seems that the normalized total cost increases with time, however we must keep in mind that the normalized total cost is a composite variable with substantial number if missing values. We take a look a the missing values in this variable disregarding the accumulated missingness.



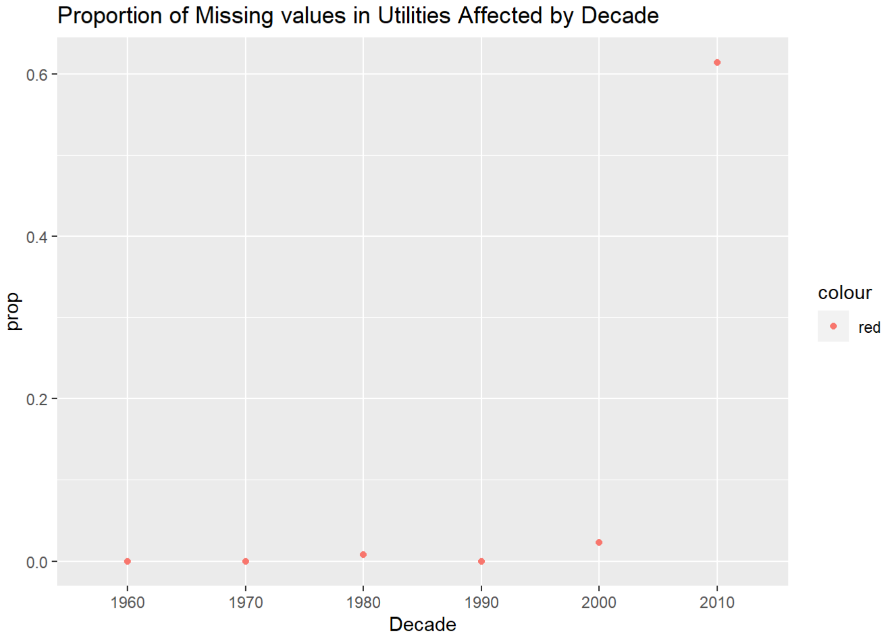
From the missing values we see that the unusual high cost which occured in 1990 was due to less values missing, which confirm the unreliability of this variable. As such we are hesitant to make any claims in regard to this variable and will not conduct any further analysis.

3.8. Utility affected

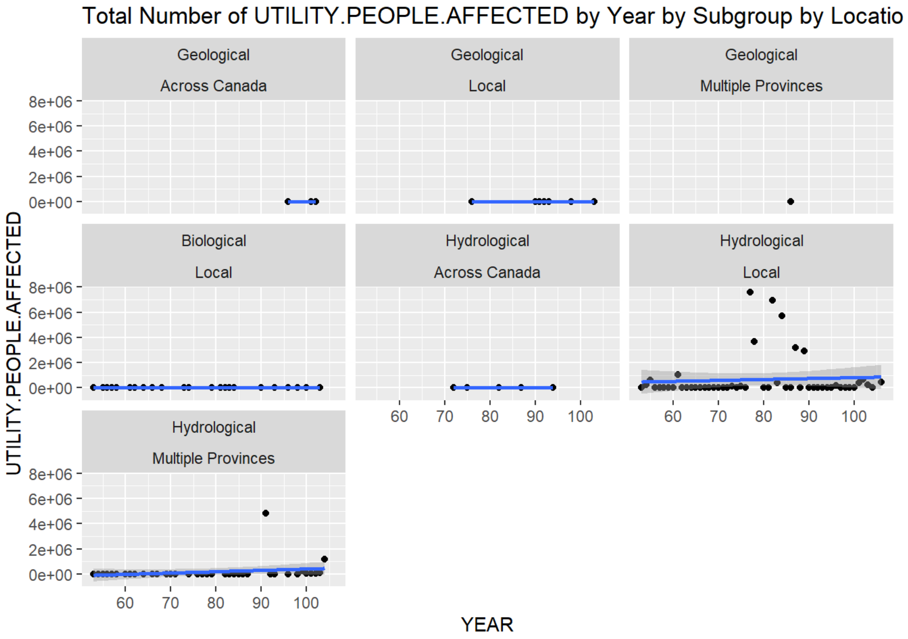
We are moving on to the variable number of people who's utilities have been affected UTILITY.PEOPLE.AFFECTED). As usual we for look at the general patters over the decades.



There is is no utility affected prior to 1950s and also no utilities affected in 1970, which is quite odd. We will only focus on the span of years from 1950-2010, and use Year as the x-axis instead of decade. Again, we look at the missing values to gauge the reliability of the variable.



Variable UTILITY.PEOPLE.AFFECTED has missing values only in the last decade, making this a reliable variable based on the data. For the purposes of exploration we plot UTILITY.PEOPLE.AFFECTED by Subgroup and by location to see if these plots will help us develop a model.



The most significant subplot is Hydrological local events, which has a prominent number of outliers in the mid of YEAR variable. We will try to model these outliers in the next sub section.

3.8.1. Modelling

We will be modelling the UTILITY.PEOPLE.AFFECTED variable as a response and variables YEAR as a covariate. The variance of the errors is weighted according to the exponential value of UTILITY.PEOPLE.AFFECTED.

```
model51<-gls(UTILITY.PEOPLE.AFFECTED^(1/3)~YEAR, df.util,
na.action = na.omit, weights=varExp(form=~UTILITY.PEOPLE.AFFECTED))
```

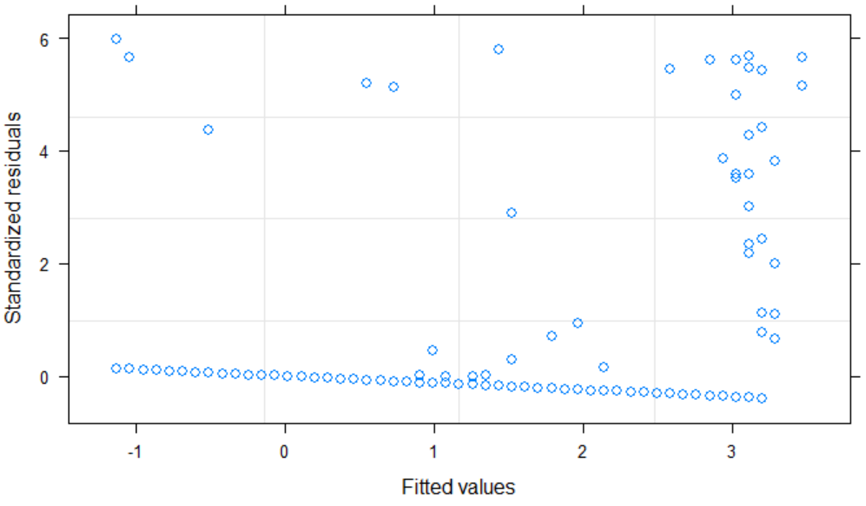
The following is the output of the model

```
Variance function:
 Structure: Exponential of variance covariate
 Formula: ~UTILITY.PEOPLE.AFFECTED
 Parameter estimates:
expon
1e-06
Residual standard error: 8.330429
Degrees of freedom: 627 total; 625 residual
```

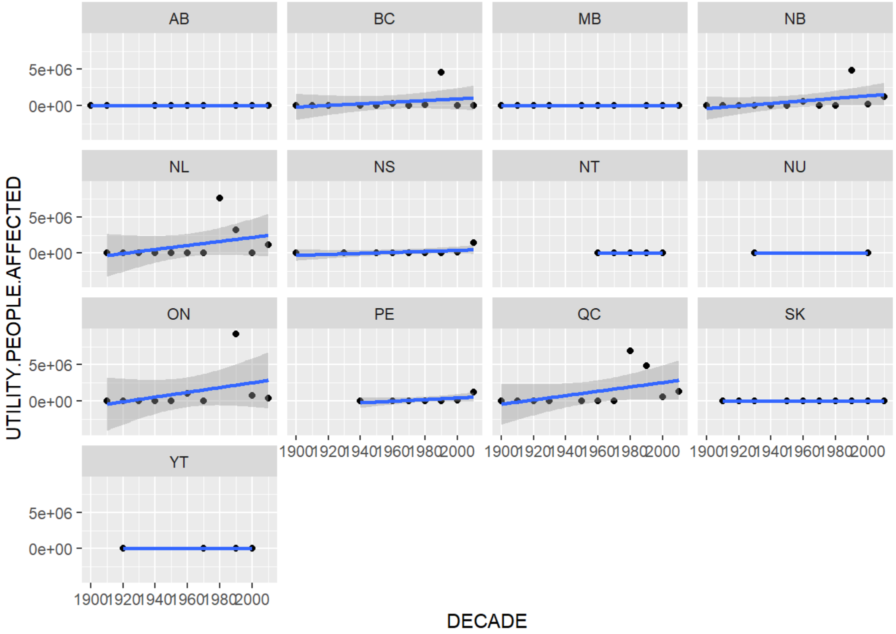
Coefficients:				
	Value	Std.Error	t-value	p-value
(Intercept)	-174.22191	48.83081	-3.567868	4e-04
YEAR	0.08827	0.02452	3.599188	3e-04

Table 7. Model Coefficients

The residual plot for the model is given below



From the graph and the model it is reasonable to say that the number of utilities affected increases with time. Looking at the subplots of provinces.



The overall utility seemed to be unaffected, the most increases come from Quebec and Newfoundland. The rest can be attributed to single outliers.

4. Conclusion

It seems from the above analysis that as the number of natural disasters in Canada continues increasing the impact on the lives of Canadians is not impacted proportionally in terms of death, injuries,utilities affected. However we see an increse in the number of people evacuated, which may suggest that the impact is mitigated mainly through Natural Disaster management practices