UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Vladimir Visbek

Linking Rescue Event Data with Public Data

Master's Thesis (30 ECTS)

Supervisor(s): Siim Karus, PhD

Tartu 2017

Linking Rescue Event Data with Public Data

Abstract:

The workers of the rescue services strive daily to ensure that the people live in a safe environment. Their aim is to achieve the level of safety indicators common to the Nordic countries. It means having fewer accidents, risen preventive awareness and increased co-operation with partners and citizens. The purpose of the thesis was to explore relationships of rescue event data with online data, such as public events and weather indicators, in order to find strong correlations that will allow the Rescue Board to estimate rescue event risk changes. The author applied correlational study, sign test method, time series and logistic regression analyses while studying the data. Results has shown that the weather data is in a strongest relationship with “Fires”, “Helpless animal/bird” and “Oil spills” rescue event types. Also, it appeared that road accidents occur more often on days when parties and celebrations occur as well. Most of these days are in the end of the week. Then, while conducting a logistical regression analysis, it appeared that statistically significant variables from public event and weather datasets predicted the probability of occurrence of rescue event considerably poorly.

Keywords: Logistic Regression, Correlation, Web Scraping

CERCS: P170 Computer science, numerical analysis, systems, control

Päästesündmuste andmete sidumine avalike andmetega

Lühikokkuvõte:

Päästeteenistuste töötajad püüavad iga päev tagada, et inimesed elaksid võimalikult turvalises keskkonnas. Nende eesmärk on jõuda päästealase turvalisuse tasemelt võrdsele positsioonile Põhjamaadega. See tähendaks päästesündmuste langust, tõusnud ohutusalast teadlikkust ja suurenenud koostööd partneritega ja kodanikega. Antud lõputöö eesmärk oli uurida seoseid päästesündmuste andmete ja avalike andmete vahel, nagu avalikud üritused ja ilmaandmed, et leida tugevaid korrelatsioone, mis aitaks Päästeametil paremini hinnata tekkivaid riske. Andmete uurimisel rakendas autor korrelatsioonianalüüsi, Sign testi, aegridade ja logistilise regressiooni metoodikaid. Tulemused näitasid, et ilmaandmetel on kõige tugevam seos päästesündmustega „Tulekahjud“, „Abitus seisundis loom/lind“ ja „Naftasaadustega reostumine“. Samuti tuli välja, et liiklusõnnetusi juhtub rohkem päevadel, kui toimuvad ka peod. Kõige tihedamini juhtub seda nädala lõpus. Lisaks, logistilise regressiooni analüüsi läbiviimisel, tuli välja, et tugevate seostega muutujad avalike ürituste ja ilmastiku andmetest ennustasid päästesündmuste tekkimise tõenäosust üsna nõrgalt.

Võtmesõnad: Logistiline Regressioon, Korrelatsioon, Web Scraping

CERCS: P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll

Table of Contents

1 Introduction 5
1.1 Scope and Limitations 5
1.2 Theoretical framework 6
1.3 Definition of Key Terms 6
1.4 Background 7
1.5 Research questions 8
1.6 Structure of the Study 8
2 Related work 9
3 Methods 11
3.1 Method choice 11
3.2 Correlational study 12
3.3 Sign test 12
3.4 Logistic Regression Analysis 13
4 Description of the data 14
4.1 Rescue event data 14
4.2 Climate data 18
4.3 Public event data 21
4.4 Data preprocessing 22
5 Results 24
5.1 Relationship between weather and rescue events data 24
5.2 Relationship between public events and rescue events 29
5.3 Time series of fires and road accidents 33
5.4 Predicting rescue events occurrence 36
6 Conclusions 42
References 44
License 46

1 Introduction

The Rescue Board is responsible for planning the development of rescue institutions, verifying the preparedness of rescue institutions, directing and coordinating rescue work and fire extinguishing in the case of major accidents and crisis, organizing and carrying out state's fire safety supervision. Among other duties, the Board also organizes and carries out explosive ordnance disposal work. [1]

The workers of the Rescue Board strive daily to ensure that the Estonian people live in a safe environment. However, safety does not always depend only on the Rescue Board. Many other factors could have a great influence on the response times of the rescue services and the general safety, e.g. public events, climate change etc. The awareness and actions of the water and fire accident risk groups are hard to influence, which is why new solutions must be found for risk mitigation through improving the safety of the physical environment. [2]

The purpose of the thesis was to explore relationships of rescue event data with online data. The study utilized a correlation design to analyze rescue event data provided by the Rescue Board to identify the relationships between factors that may illustrate significant associations and bring Rescue Board's attention to online data. The key principles of the correlation approach were to identify associations between rescue events and weather, and between rescue events and public events.

Logistic regression analysis was used to identify the strength of the effect that independent variables have. Also, to understand to what extent the occurrence of different types of rescue events can be predicted if we change the independent variables such as weather indicators and a number of public events.

The Rescue Board aims to achieve the level of safety indicators common to the Nordic countries. All in all, it means having fewer accidents and less damage in the future, risen preventive awareness and increased co-operation with partners and citizens. Prevention work is one of the most important parts in reaching these goals as it shapes the environment in the country, where every person creates and values safety and security. As a result, it helps to reduce the number of accident-related deaths, emergencies, injuries and other damages.

The findings of this study will redound to the benefit of society considering the mission and vision of the Rescue Board. Their efficient work ensures that our environment is safer and healthier. Thus, taking into account the results, this study will help the Rescue Board to achieve that. For the author, the thesis will help to uncover new theories and findings that other researchers were not able to explore yet.

1.1 Scope and Limitations

The scope of the study is confined to the rescue events occurred in Estonia, as is contained in the data sets provided by the Rescue Board. Other online data, climate, and public events were also collected considering the scope of this study.

There were sufficient limitations in using the whole potential of provided rescue events data because it was created and combined by hand, and contained multiple mistakes. For example, although the provided data covered all rescue events in Estonia from 2010 to 2015, due

to the problems with municipality names in the data for the year 2014, this study concentrated only on the period of 2010 to 2013. Also, there were differences in the variables and their names for different years of data.

In addition, there are limitations considering the fact that the rescue events data has locations only on the municipality level, as revealing more specific locations might affect the privacy of peoples' personal information. In the other hand, locations provided in coordinates would have had a much bigger potential and would have increased the specter of methods and analysis techniques that could have been used to analyze the data.

The study was also limited by the lack of online data of Estonian public events. Although the needed datasets were generated using the Web Scarping technique from a trustworthy public event web portal, the resulting data is not initially official.

1.2 Theoretical framework

Weather theory.

It is believed that the temperature is the single most important weather factor affecting fire behavior. Air temperature has an influence on fire because the heat is one of the ignition requirements and it continues the combustion process. In addition, warm forest fuels ignite and burn faster because less heat energy is needed to raise the fuels to their ignition temperature. [3]

Public events theory.

Accident analysis implemented by the Rescue Board has shown that many of the problems result from the people's inability to cope with the problems. Most of the callouts are in some way connected to other people and their behavior. In addition, a lot of fires are human-caused fires, which started from people's carelessness with open fire or smoking. As public events are occasions when people are gathered into big groups and it can be tracked, it was reasonable to study if there is any relationship. [2]

1.3 Definition of Key Terms

Data mining

For the most part, data mining allows us to convert very large and complex data sets, into small, simple and meaningful things. The sheer scale of Big Data has far exceeded human sense-making capabilities and at these scales, relationships are often multi-dimensional or too complex to observe and understand by simply looking at the data. Data mining summarizes and simplifies the data and then allows us to infer things about specific cases; it helps us see the forest without getting lost in the trees. [3]

In addition, data mining includes several main types of pattern detection: anomaly detection, association learning, cluster detection, classification, and regression. [3]

Business Intelligence

Business intelligence is a technology-driven process for analyzing large data sets and presenting actionable information to help users make more informed business decisions. In result business intelligence programs might accelerate and improve decision making; increase operational efficiency. BI systems can also help organizations spot business problems that need to be addressed. [4]

Web Scraping

Web Scraping (also called Web Harvesting, Screen Scraping, Web Data Extraction etc.) is a technique used to extract large amounts of data from websites and then the data is saved to a local file in user's computer or to a database in table format. [5]

Data displayed by websites can usually be viewed only using a web browser. Websites do not offer the functionality to save a copy for personal use. The only option is to copy and paste the data manually, which is a very tedious job if the desired data has a couple of thousands of records. Web Scraping is a technique to automate this process.

1.4 Background

The rescue service of the Republic of Estonia's Rescue Board, the governmental institution within the Ministry of Internal Affairs, is a part of the internal security system. It encompasses rescue institutions located in different parts of the Estonia and having different levels of readiness. The service allows using all of the existing resources and forces in protecting and rescuing people. Its mission is to prevent accidents and to save lives, property and the environment in a fast and professional manner. The vision for 2025 is to reduce accidents and losses to the level seen in Nordic countries. [6, 2, 1]

The Rescue Board was created on May 25, 1992, when the government issued a regulation dismissing the National Fire Service and transferring all their functions and assets to the National Rescue Board. The organization was initially governed locally, but over the course of a couple of decades has grown into a solid organization that provides rescue services through four regional centers: North-Estonian, East-Estonian, South-Estonian and West-Estonian Rescue Centre. [6, 2]

The Rescue Board administers the regional rescue centers, which are state institutions, who organize and carry out rescue works (including firefighting), fire safety supervision, emergency prevention and crisis management. [1]

The Explosive Ordnance Disposal Center is a part of the Rescue Board and is responsible for the activities related to explosive ordnance disposal. Together with Bomb Squads, the Center have also capacities in response to chemical threats. They also handle sniffer dogs and perform life-saving diving. [1]

As of 9th of the November 2016, there are in total 187 rescue stations: 72 professional and 115 voluntary stations. On Figure 1 are shown geographical locations of all stations. Orange ticks indicate professional and green voluntary stations. [7]

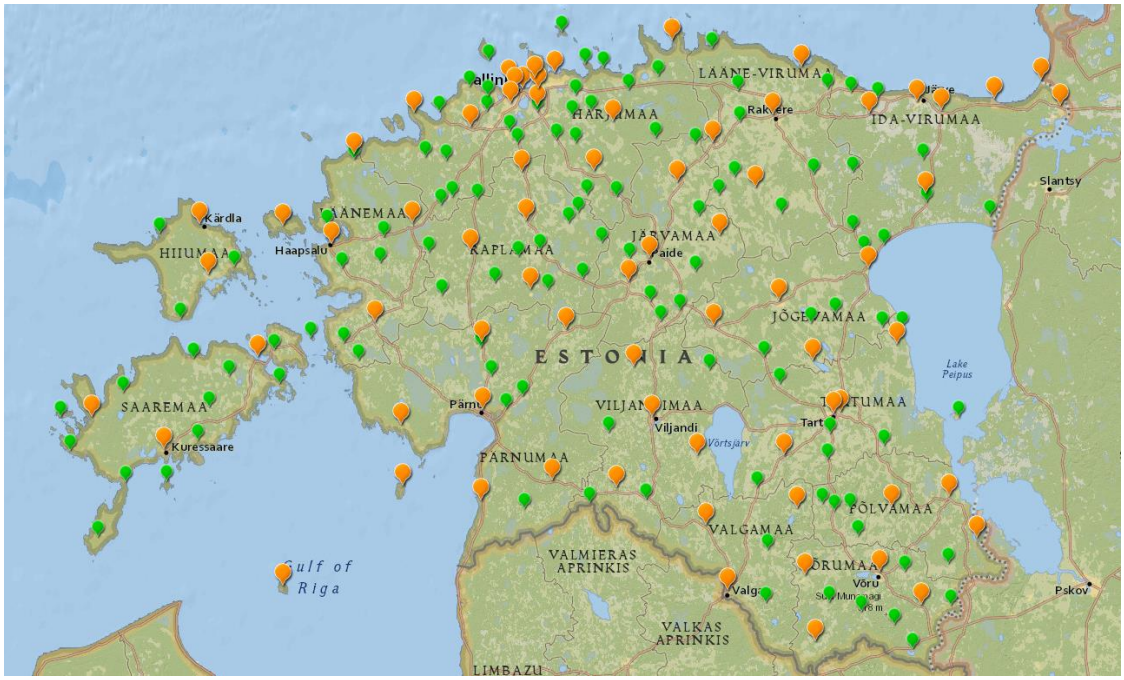

Figure 1. Estonian rescue stations – professional (orange) and voluntary stations (green) [7]

1.5 Research questions

This study addresses three research questions:

1. What relationship exists between different types of rescue events and weather factors?
2. What relationship exists between different types of rescue events and public events?
3. Can the probability of the occurrence of rescue event be predicted based on the climate data and public events variables?

1.6 Structure of the Study

Chapter 2 discusses related work and the findings of these papers. Chapter 3 presents the techniques and tools used while conducting the analysis. Chapter 4 gives an overview of the gathered online data and the rescue events data provided by the Rescue Board. Also, it explains how all these datasets were preprocessed. Chapter 5 shows the results of the research and introduces possible solutions. Finally, chapter 6 gives an overview of the done work, summarizes and discusses the findings, concludes this thesis and outlines future work.

2 Related work

In 2014, I. Vares identified spatial and temporal patterns of the Rescue Board fire accidents in the period of 2009 to 2013. The author has also analyzed the number of casualties caused by the fire and the relation between fire accidents and temperature in Estonia. I. Vares has stated that there are two main causes of fire: human and nature caused fires. Two most common human-caused fires are caused by carelessness with open fire (candles, matches, campfire etc.), and smoking. Most common nature caused fires are caused by thunderstorm and self-ignition. The author has used quantitative and descriptive data analysis to analyze the patterns and she has also used five data sets: rescue events data, region data, temperature data, population density data and rescue stations distance zone data. [8]

Results have shown, that in the period of 2009-2013, there were 42248 fires and the number of events per year was more or less stable. The most fire accidents occurred in Tallinn and Narva. On average, there were 621 fires in the Tallinn district of Lasnamäe. In comparison, there were only 90 fires in the city of Rakvere. An interesting finding was that the most fires per person have occurred not in the big cities, but in small villages, where the population is low, but the number of fires per each inhabitant was large. Temporal analysis showed that the *fire season* starts with a big leap in April (the number of fires increases compared to March almost for 250%). Then it slowly decreases and stays pretty much on the same level from September until the next April. This big leap is mainly caused by the increase of landscape and forest fires. In addition, the author has found that the most casualties to fire occur at nighttime in February and December. Lastly, I. Vares explained with a table of April and May air temperatures and fires how temperature and fire are related. In the beginning of the April, the number of fires indeed increases together with the temperature. However, these assumptions were made fully on visual observations. [8]

In 2016, J. Horm used the same data provided by the Rescue Board, but with the aim to analyze the data collected between 2010 and 2013 with automatic data mining algorithms. The data was processed using two data mining algorithms from Microsoft SQL Server 2014 Analysis Services software (Microsoft Clustering Algorithm and Microsoft Naïve Bayes Algorithm), in which categories containing similar rescue team callouts were grouped into clusters. [9]

Results have shown that two main county clusters are Tallinn and Ida-Virumaa (northeast Estonia), with more than 80% of rescue events related to fire accidents. Two biggest month clusters were December and August. Almost half of the accidents that occurred in December were nature caused accidents and third were related to fire. However, in August, only 15% were nature caused damages and over half were fires. The biggest cluster for “Fire” showed that for 70% of the time it occurred in the Tallinn county, about 13% of the time in Võrumaa (southeast Estonia) and for 17% of the time in other counties. In addition, almost every second fire is “building fire”. The biggest cluster for “No work performed” has shown that in these cases almost every time the callout type is “Fire alarm” and for 80% of the time, it happens in Tartumaa county. Another interesting cluster for “Landscape fires” has revealed that for 60% of the time it happens in April and for 35% of the time in May. Nature caused damages happened the most in December (43%), and in March and June (20-25%). [9]

The Rescue Board was also interested in trends in the least common callouts (often their properties and patterns stay unnoticed next to callouts that are more common), so the author has also studied a separate dataset with rescue team callouts that occurred less than 3000 times. There were three county clusters that stood up the most: Ida-Virumaa, Tallinn, and

Tartumaa. In the first two “Assistance” and “Helpless person” were the most occurred events. In Tartumaa, together with previous two types, “Gas emergency” and “Water accident” were also popular. [9]

The thesis pointed out a bunch of interesting patterns, trends, and exceptions. In addition, some of the results are similar to the results given in the paper of I. Vares. J. Horm has also stated that there are still a lot of information hidden and suggested to combine the rescue events data set with other additional data sets.

In the United Kingdom in 2009, a research report was published that analyzed the factors affecting the increase in response time of Fire and rescue Services. The aim was to assess the associations between trends in response times and trends in traffic levels, Fire and Rescue Service incident workload and a number of pumping appliances over the period of 1996 to 2006 for England as a whole, each region, FRS family group and individual Fire and Rescue Services. For the past 10 years, the average response time in minutes increased by 18% from 5.5 to 6.5. At the same time, the average mobilization time of the rescue teams didn’t change at all. The possible reasons were identified by consultation with eight Fire and Rescue services: traffic (increase in traffic levels, changes to phasing of traffic lights and illegal parking), new operational policies, changes in local and geographical knowledge of the crews, changes in shift times etc. [10]

Using the quantitative analysis, researchers assessed the following three possible factors: traffic, Fire, and Rescue Service resources and a number of emergency incidents, by comparing them with trends in response times. To identify the greatest associations with the response times, partial correlations and multiple regression analyses were carried out. Results have shown that response times have primarily increased due to rises in traffic levels. The trend was clearly evident from 1999 onwards. The small decrease in the number of pumping appliances and the fall in the number of incidents did not account for the increase in response times. This research report supports the hypothesis that merging rescue events data together with other potentially related data sets might reveal new hidden facts and trends, or at least provide proof or disproof of already stated theories using new statistical methods. [10]

In conclusion, the rescue events data of the Rescue Board has already been previously studied multiple times. However, these papers concentrated only on the rescue events data and its hidden patterns without the aim to analyze the relationships with other online datasets. I. Vares analyzed spatial and temporal patterns and J. Horm categorized rescue events into clusters. So, it was decided to use some other statistical methods and predictive analyses in combination with multiple other potentially related online datasets, as it was similarly performed in the UK in 2009 with the response times of Fire and rescue Services and trends in traffic levels. This is the gap that this thesis intended to focus on and try to fill.

3 Methods

As it was described in Chapter 1, the objective of this paper was to study the relationships between rescue events data and online data, such as public events and weather, and also examine a predictive potential of these data sets. So far we introduced our problem area, purpose, and significance of the study. We explained the motivation behind the selection of our theoretical framework as mean to provide answers to our research questions. We have also discussed the scope, limitations and research questions. Chapter 2 outlined the relevant literature in the context of our study. And in this chapter, we discuss the selection of our research methods and give a brief overview of their usefulness.

3.1 Method choice

Most of the analysis was done using the R open source programming language as it is a software environment for statistical computing and graphics. The R language is widely used among statisticians for developing data analysis. Python programming language was also used, but only for the web scraping the public events data from the Estonian website. [12]

As the whole research evolves and revolves around the research questions, they were the foundation that dictated the choice of the research methods.

Considering that the first two research questions strive to understand the scale of the relationship between two variables, correlation research method seemed to be more than suitable for this job. It is usually used to determine the extent to which two or more variables are related. In addition, there is no need to manipulate the variables but only measure them and look for relations (correlations). However, correlations only describe the relationship, they do not prove cause and effect. [13]

Sign test, which is a statistical method, was also used to test for consistent differences between pairs of observations. In our case, these were mean numbers of rescue events of different types occurred on days with and without specific public events. For each subject of the pair, the sign test determines if one of the pairs (rescue events on days when also public events have occurred) tends to be greater than or less than the other member of the pair (rescue events when there were no public events). This method has also helped to understand how two different data sets are related, and to what extent. [14]

To understand the underlying forces and structure of the mean sequences of the rescue events and public events a method called Time Series was also used. Time Series are an ordered sequence of values of an observed variable at equally spaced time intervals. In our case, these are the weeks of the year. The analysis allows studying the internal structure of the sequences, such as trends, seasonal variation etc. [15]

As one part of the research was to analyze the predictability of the rescue events based on online data sets, a predictive analysis was also performed. Logistic regression was used to describe and explain the relationship between a dependent binary variable, which indicated if at least one rescue event has occurred on a given date, and other independent variables (weather factors and the number of public events). [16]

3.2 Correlational study

Correlation is a statistical technique that shows how strongly two variables are related. The most common type of correlation techniques is a Pearson correlation. The main result of the technique is called correlation coefficient. It ranges from -1.0 to +1.0, with the latter indicating a perfect positive relationship, 0 indicating no relationship, and -1.0 indicating a perfect negative relationship (often called an *inverse* correlation). There is also the second result of each test – statistical significance. It tells how likely it is that the correlations with high coefficient may be due to a chance. [11, 12]

Correlation works only for quantifiable data in which variables are quantities of some sort (not categorical data, such as color and gender). An example of good correlation are height and weight – taller people tend to be heavier than shorter people. [11]

Correlational research is looking for variables that seem to interact with each other. So, when there are two strongly correlated variables and one is changing, it can be assumed how the other will change. In our case, if it appears that some weather factor is in a strong relationship with some type of rescue event, then the Rescue Board can better plan their work when the weather factor is changing. Considering that the correlation can also be negative, it would mean that with the decrease in the value of weather variable there will probably be an increase in the number of accidents.

An important rule should also be never forgotten – correlation does not imply causation. For example, if there is a strong correlation between air temperatures and a number of fire accidents, it does not imply for 100% that the increase in fire accidents is caused by the increase in air temperature.

3.3 Sign test

The Sign test is a non-parametric statistical method that is used to test whether or not two groups are equally sized. It is also called the binomial sign test.

First, positive and negative results are counted for each pair. Then the critical value at the significance level of 0.05 is computed using the formula:

$$\sum_{i=k}^{n} \binom{n}{i} \times p^i \times q^{n-i}$$

where:

- n – number of subjects entered into the analysis
- k – number of positive differences
- p – the probability for positive change under H0 assumption
- q – 1-p, probability for negative change

The null hypothesis is that there is no difference between the observed observations. Also, the test works for left-tailed, right-tailed, and two-tailed tests. In our research, we used a two-tailed test, as we did not know initially at which direction the difference could be. The

alternative hypothesis for the two-tailed is that there is a difference, either greater or less (one direction).

3.4 Logistic Regression Analysis

Logistic regression belongs to the family of generalized linear models. It is a widely used binary classification algorithm and is suitable when the response variable is dichotomous, that is either 1 or 0. In our case, these are 1 if at least one rescue event occurred on a given date, and 0 if not. So, its typical use is prediction binary variable given a set of predictors. The predictors can be categorical, continuous or a mix of both. R makes it easy to fit a logistic regression model. [19]

It can be assumed that all included explanatory variables, the predictors, are always suitable while fitting the model. But usually, it is a part of the analysis to decide which ones should be included. There are two main approaches towards variable selection: automatic methods and all possible regression approach. The latter one considers all possible subset of the pool of predictors and finds the model that best fits the data according to some criteria. These criteria allow us to choose the model with the best score. But considering that we had to fit best logistic regression models for all of the rescue events types for three different cities, it was decided to use automatic search algorithm with forward selection.

When developing models for prediction, the Accuracy is one of the most critical and important metrics that indicates how well the model does in predicting the target variable. If the creation of the model used a training set to learn how and what to predict, then the next process involved using the model estimates to predict values on the testing set. After that, the predicted target variable was compared versus the correct observed values.

Accuracy is not always reliable as it will yield misleading results in the case of unbalanced data (when the number of observations in two classes varies greatly). To assess the performance of the classification a table of confusion (or confusion matrix) and cumulative gain chart were also used. The confusion matrix is a table with four slots for the number of true positives, true negatives, false positives and false negatives. The accuracy and far more other statistics are calculated based on these metrics. E.g. sensitivity (how often does it predict yes if it is actually yes) and specificity (how often does it predict no if it is actually no). A cumulative gain chart represents graphically the improvement that a classification model provides when compared against random guessing and ideal model. [20, 21]

As most of the data after merging rescue events and public events was highly unbalanced, over- and undersampling techniques were used to adjust the distributions of the dataset (the ratio between the different classes represented).

4 Description of the data

In the previous chapter, we discussed the choice and usefulness of our research methods that we apply in our study, with relation to the methods used in relevant literature. Next, we describe in detail the collection and description of all of the used datasets. In addition, we explain in detail how the data was preprocessed for the further analysis.

4.1 Rescue event data

Data collection and description

The representatives of the Rescue Board were contacted through email in order to attain permission to use the data of the rescue events. The data was sent in three .XLSX (Excel) files. First two files included the departure times of the rescue teams. One covered the years 2010 to 2012, and the other one 2013-2015. In total these two files included 256006 records: 37883 in 2010, 38445 in 2011, 34581 in 2012, 42612 in 2013, 56036 in 2014 and 46449 records in 2015. Each of these records is not unique rescue events – they are unique callouts for different rescue stations related to some rescue event. It means that if four different rescue stations were notified of the same rescue event, then there are four records about it, but with different ID numbers. The variables of first two files are shown in Table 1.

Table 1. Description of the data containing the callouts to rescue stations

No.	Name of the variable	Description
1	Callout number	Unique numerical identification digit used to differentiate from the others
2	Callout time and date	The time when the emergency call was received. The format is “DD.MM.YYYY hh:mm:ss”.
3	Callout type	States the type of the rescue event, like building fire or traffic accident.
4	Municipality	Name of the municipality where the accident or damage took place.
5	County	Name of the county where the accident took place.
6	Complex	Indicates if rescue included ambulance (yes or no).
7	Degree of rescue	Indicates the degree of difficulty and needed resources: from 1 to 4, where 1 is easy and 4 is most difficult (schools, high buildings etc.). The value can also be “Consultation” and the rescue team does not leave the station.
8	Designator	Specifies the rescue team and the vehicle type, e.g. Tartu 11 is a rescue team from Tartu on a first main car.
9	Unit	Specifies the type of the rescue team (rescue, voluntary, explosive ordnance disposal etc.).
10	Unit type	Rescue or ambulance.
11	Chief unit	Specifies the region of the rescue.

12	Notification	The time when the rescue team got the notification of the accident. The format is “DD.MM.YYYY hh:mm:ss”.
13	Departure	The time when the rescue team left the rescue station. The format is “DD.MM.YYYY hh:mm:ss”.
14	Cancelled	The time when the departure was canceled (happens usually before departure or before arriving at the accident place). The format is “DD.MM.YYYY hh:mm:ss”.
15	On-site	The time when the rescue team arrives at the accident place. The format is “DD.MM.YYYY hh:mm:ss”.
16	Left	The time when the rescue team left the accident place. The format is “DD.MM.YYYY hh:mm:ss”.
17	Home	The time when the rescue team arrived back to the rescue station. The format is “DD.MM.YYYY hh:mm:ss”.
18	Elimination	The time when the accident elimination took place. The format is “DD.MM.YYYY hh:mm:ss”.
19	Localization	The time when the fire location was determined (used only in case of fires). The format is “DD.MM.YYYY hh:mm:ss”.
20	Rescue event subtype	Specifies the subtype of the rescue event assigned by the rescue team e.g. car accident or fire extinguishing in the building.
21	Building on fire	Specifies what was on fire.
22	Vehicle on fire	Specifies the type of vehicle that was on fire.
23	Device on fire	Specifies what kind of device was on fire.
24	Other on fire	Specifies something that was on fire (garbage, campfire etc.).
25	Landscape fire	Landscape or forest.
26	Work description	Short description of the performed work while eliminating the accident.
27	Special activities	Evacuation, first aid etc.
28	Cause of the accident	Specifies what caused the accident e.g. fireworks, campfire etc.
29	Event status	Specifies the status of the rescue event: wrong, true, false, malicious etc. (only for the year 2015)

As can be seen from the data variables, not all of the fields can always be filled, because some callouts are canceled or some events were not related to fire at all. For example, in the case of drowning accident variables “Building on fire”, “Vehicle on fire” and “Landscape fire” does not contain any value.

The third sent file is a little different compared to the first two. If the first two were more concentrated on the rescue stations and their teams, then this one is more specific on event details. It contains all six years in one file in six separate sheets. There were in total 129607 records: 20791 in 2010, 21124 in 2011, 19238 in 2012, 20265 in 2013, 21716 in 2014 and 26473 in 2014. Table 2 describes the variables of the third file.

Table 2: Description of the data containing rescue events

No.	Name of the variable	Description
1	Callout number	Unique numerical identification digit used to differentiate from the others
2	Callout time and date	The time when the emergency call was received. The format is "DD.MM.YYYY hh:mm:ss".
3	Start of the rescue work	The time when the rescue work has started. The format is "DD.MM.YYYY hh:mm:ss".
4	Time of event localization	The time when the fire location was determined (used only in case of fires). The format is "DD.MM.YYYY hh:mm:ss".
5	Time of event elimination	The time when the accident elimination took place. The format is "DD.MM.YYYY hh:mm:ss".
6	Callout type	States the type of the callout event, like building fire or traffic accident.
7	Callout subtype	States the subtype of the rescue event, like fire extinguishing or traffic emergency.
8	Event type	States the type of the rescue event, like fire extinguishing outside of the building or elimination of nature created damages.
9	County	Name of the county where the accident took place.
10	Municipality	Name of the municipality where the accident or damage took place.
11	Number of fatalities	A number of death occurrences by accident.
12	Localization of the deceased	Time taken to find the deceased people (in minutes).
13	Deceased rescuers	A number of the deceased rescue workers.
14	Injured	A number of injured people.
15	Evacuated	A number of evacuated people.
16	Injured rescuers	A number of the injured rescue workers.
17	Rescued	A number of rescued people.
18	Localization of the rescued	Time taken to find the rescued people (in minutes).
19	Area of burned land	The area damaged to fire (in m^2).

As can be seen from both tables, they have some same variables, like the callout type and identification number. But most of the variables are still different. For example, the first dataset (first two files) is more specific on the rescue stations that got the callout – time of the notification, departure time etc. In the other hand, the second dataset (third file) doesn't have any information of the rescue stations at all but gives more information on the rescue event itself – callout type, a number of injured people etc.

Although the Rescue Board has very accurate locations of the accidents (addresses and even coordinates), due to the privacy of the people's personal information, these datasets include locations only on the municipality and county level. In Estonia, municipalities are of two types: towns or urban municipalities and parishes or rural municipalities. A municipality may contain one or several populated places. In addition, some urban municipalities are divided into districts. Since 12th of the December 2014, there is a total of 213 municipalities: 30 urban and 183 rural (Figure 2). [13]

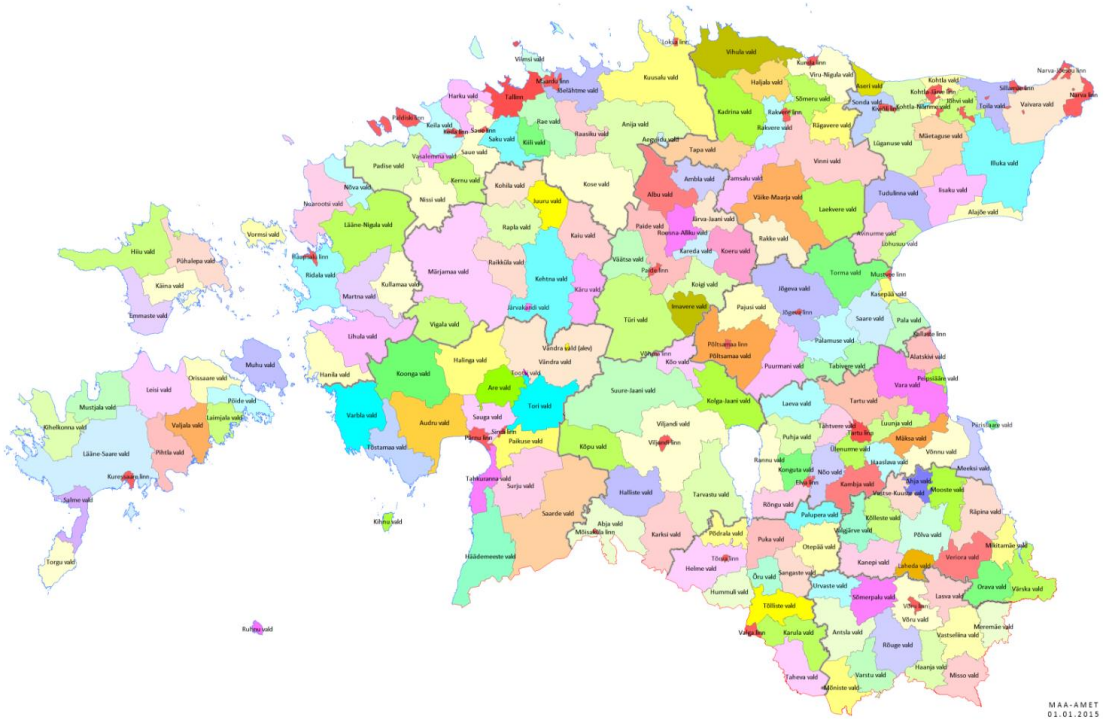

Figure 2. Municipalities of the Estonia [14]

In addition, the provided data (the variables and values) were fully in the Estonian language.

Quality of the rescue events data

While conducting the study the datasets have delivered a lot of problems, as there were many quality issues. Some variables for some years (in both datasets) were missing from other years. For example, in the first dataset, the years 2010 to 2012 had 17 variables, the year 2013 had 19, the year 2014 had 14 and the year 2015 had 30 variables. In the second dataset the years 2010-2013 had 16 variables, but years 2014-2015 had 19. Already only this problem limits significantly the work that can be done with the data, as when analyzing different years, they should all have the same variables.

Another inconsistency with the data is that even if there are same variables for different years, sometimes they still differ by their names. For example, in the first dataset, the names of all variables for years 2011 and 2013 are totally different. The variables stating the callout ID and type in 2011 were called as “Väljakutse aeg” and “Sündmuse liik (SOS)”, but in the year 2013 they are called “VÄLJAKUTSE_NR” and “SÜNDMUS”.

In addition, although both datasets cover the same years, they have separately a different number of records. And the difference is pretty significant – the second dataset, that is more specific to the rescue events, is two times smaller than the first dataset. Furthermore, the analysis has shown that there are some callouts that are included only in the first dataset or in the second one. But at the same time, most of the callouts are included in both.

Furthermore, the rescue events lack more specific description. For example, if the event type is “Nature caused damages”, then the addition of explaining what kind of “nature” caused the damage would have greatly increased the potential of the data.

Lastly, datasets were inconsistent in the way, that their features were not the same across the whole dataset. E.g. in the first dataset, that describes the rescue stations activities, the year 2014 municipality "column" included its correct form only for half of the dataset. The other half was changed for what it seemed to be a more specific location description, like addresses and building details. This made the rescue event data for year 2014 totally unusable.

In conclusion, it seems that these datasets were built by hand and "copy pasted" from some other files. This would explain why they haves such many inconsistencies.

4.2 Climate data

Data collection and description

Two different sources were used to collect the climate data. The first data set was exported from the Statistics Estonia website. SE is a government agency in the area of administration of the Ministry of Finance. Their main task is to provide individuals, business and research circles, international organizations and public institutions with reliable and objective information on the social, economic, demographic and environmental situation and trends in Estonia. The website has a lot of different datasets from population and environment to the economy and social life. As can be seen in Figure 3, when some dataset has been selected it is possible to specify the table features that will be outputted. Also, the output can first be seen on the screen before exporting it in the desired file format. [24]

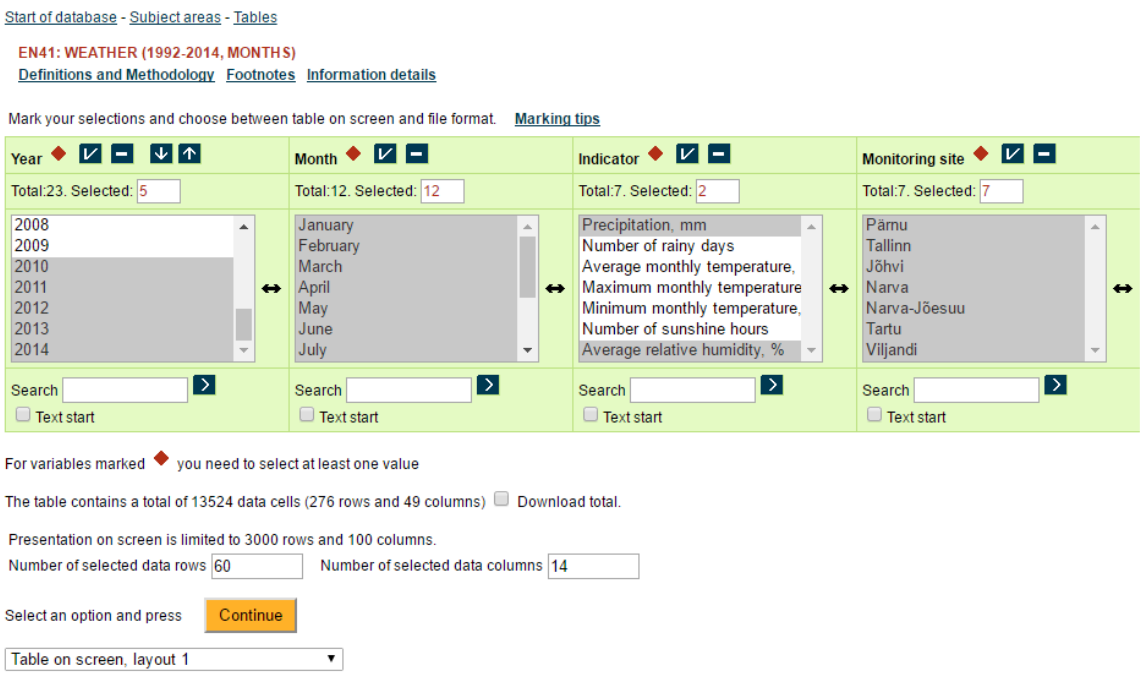

Figure 3. Importing weather data from the Statistics Estonia website

As the data of the rescue events covered the years 2010 to 2015, the chosen weather data covered the years 2010 to 2014 (initially the range is 1994-2014). Also, all other features

were select as well: all months (January to December), all weather indicators and all offered cities. There were in total seven climate specific variables: precipitation (mm); a number of rainy days; average, maximum and minimum temperature (all in Celsius); a number of sunshine hours; Sand average relative humidity (percentage). In addition, there were seven cities that could be selected from, Pärnu, Tallinn, Jõhvi, Narva, Narva-Jõesuu, Tartu, and Viljandi. Although, Narva city did not have any values for the selected years and variables.

Another source for the climate data was the website of the National Center for Environmental Information. NCEI is responsible for providing access to one of the most significant archives on Earth, with comprehensive atmospheric, oceanic, and geophysical data. NCEI develops global and national datasets, which are utilized to maximize the use of our natural and climatic resources while also minimizing the risks caused by climate variability and weather extremes. [25]

The datasets listed on the website include daily summaries, global monthly and yearly summaries, normals monthly, daily, hourly etc. The most useful dataset for this research was of the daily summaries and it comes specifically from Global Historical Climate Network. The daily data was developed to meet the needs of climate analysis and monitoring studies that require data at a sub-monthly time resolution, which is perfect for this study.

The data is accessed through Climate Data Online search interface. The user should specify weather observation type, date range, what to search for (stations, countries, cities etc.) and a search term. The data is then added to the cart where the desired formatting and additional output options can be specified. The output file can include total precipitation, depth of snow on ground and average, maximum and minimum temperatures as the weather indicators. The dataset is then sent to the entered email address after the request is submitted.

For this paper, the search was completed for whole Estonia with all available data types. CSV (Comma Separated Value) file was selected as an output format as it is the most suitable for the later analysis. Besides the weather data, the dataset also included the following variables: station (code), station name, elevation, latitude, longitude, and date. There were in total 14 weather stations: Vilsandi, Virtsu, Võru, Kihnu, Tallinn, Viljandi, Türi, Kuusiku, Kunda, Tartu, Lääne-Nigula, Väike-Maarja, Ruhnu and Valga. In addition, the dataset included 30476 observations.

Quality of the weather data

The problem with the first weather dataset (from the Statistics Estonia website) is that its data was only on a monthly resolution. So, it was pretty small if we consider that for the range of 5 years (2010-2014) it made only 60 observations. This was the reason, why another source with a better time resolution was considered in the first place. But on the other hand, it was still usable as it contained interesting and unique weather indicators.

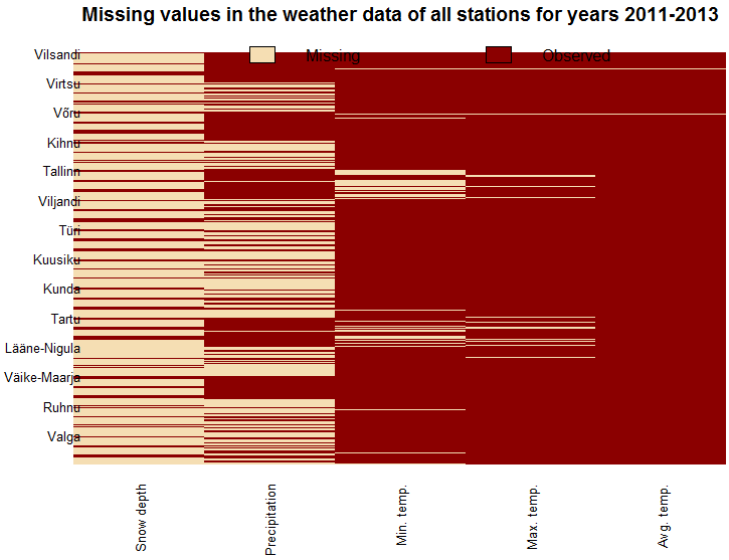

Figure 4. Missing values in the weather data for all stations for years 2011-2013

As for the second weather dataset (from the National Center for Environmental Information) it had also had some quality issues. Figure 4 shows which weather indicators have missing values. The y-axis represents all weather stations, and the space between them is the range of 2011 to 2013. If some date has a missing value for some weather variable, then a yellow colored line is put there. So, the more yellow color means more missing values. In the other hand, if the value is not missing, then the red colored line is put there. As can be seen in the figure, variables depth of snow on the ground and precipitation include a significant number of missing values. But, there is precipitation data for Vilsandi, Võru, Tallinn, Tartu and Väike-Maarja, as there are red boxes that cover most of the time ranges.

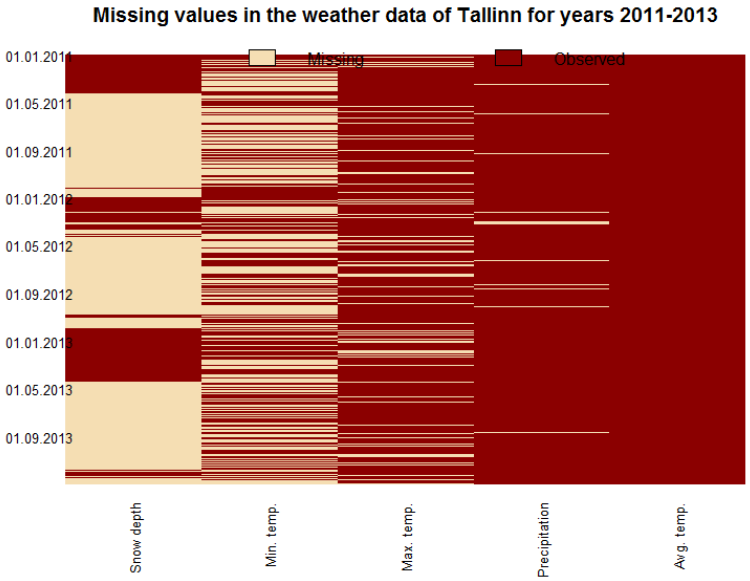

Figure 5. Missing values in the weather data of Tallinn station for years 2011-2013

On Figure 5, that shows only the missing values for Tallinn, can be seen that there is indeed a lot of missing values for snow depth, as snow appears only at winter. The time range of three years on the y-axis indicates, that the data for depth of snow starts each year a little bit before the January and lasts each year almost until the May. Also, as it was assumed based on Figure 4, there is a lot of data of precipitation. But compared to the data for all stations, there is very little data of minimum temperature. In addition, there are is a number of missing values of maximum temperature, but it is not significantly big. So, considering the stated above, only the maximum temperature, average temperature and precipitation were used in the analysis of the data, and their missing values were adjusted based on the monthly averages.

4.3 Public event data

Data collection and description

When considering what other public data can be used to compare with rescue events the public events was one of the most interesting ones, as nobody has really ever analyzed them together. Currently, there are no any public sources for complete datasets of Estonian public events, so the only choice was to use Web Scraping on one of the biggest and most popular Estonian cultural portals Culture.ee. The web portal has an event calendar, which publicizes the cultural events taking place in larger and smaller places of Estonia in all their variety. [26]

The archive of the portal has events previously published in Culture.ee. Search is possible among events that took place in Estonia by time, topic, place and organizer. The Web Scraping was implemented in Python with Beautiful Soup and *urllib.request* libraries. [27]

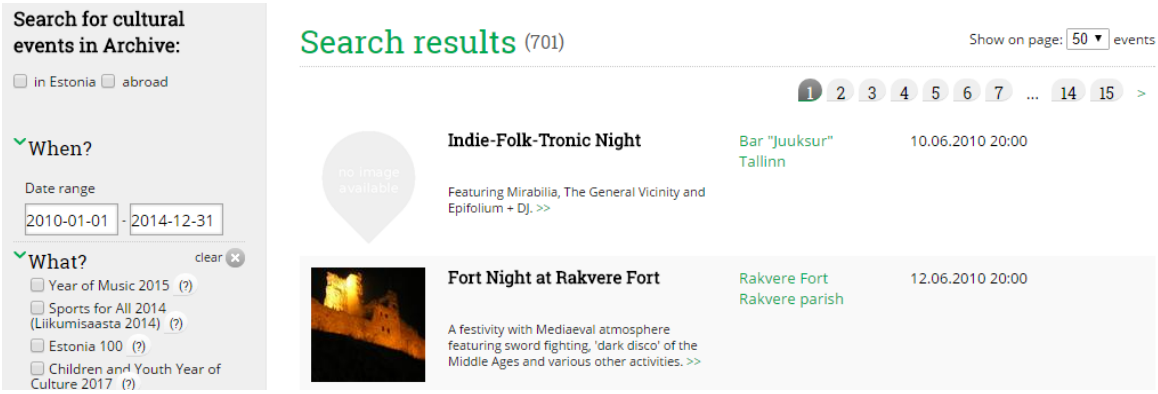

Figure 6. Public events on the Culture.ee website

On Figure 6 can be seen how data on the website was presented. The desired information was in an HTML table with three columns. The first column was the name of the public event, the second column described the location and the third one stated the date of the event. Beautiful Soup library allowed to find the "<tr>" tag on the page and extract desired information into the Excel file.

As a result, multiple datasets were created – all public events, only parties and celebrations, only music events and festivals etc. All of them had the variables Year, Location, Date, City and Event name.

Quality of the public events data

Although the public event data was created by the thesis author, it still has had some quality issues as well. The problem that limited the research the most is that a significant number of public events started only in May of 2010. In the period of January to April include there were only two public events. It means that on the website there is missing data for these dates.

Another big issue with the data presented on the website is that sometimes it was shown in ranges. Many times it was in a good format, e.g. "29.01.2011 22:00". But there were events, for example, with the date "03.02.2011 – 07.02.2011" and it could have meant either that this event occurred two times – on third and seventh of the February, or it occurred more than two times on days in the presented date range. And there was no way to get a more specific date of occurrence of the event.

4.4 Data preprocessing

Preparations for rescue events data

There were multiple cases during the analysis when different types of rescue events needed to be extracted to be analyzed separately. These types were chosen by the number of their occurrence – most popular rescue events types. These were fires (included all types of fires beside fire alarms), road accidents, nature caused damages and injuries (weakness, helpless person, pain in the chest, stroke, stomach, limb etc.). Only for the sign test analysis a rescue event type "Helpless person" was also additionally created. Furthermore, for correlational analysis between rescue events and weather indicators, other less popular types were studied as well. These were helpless animal/bird, others, assistance, fire alarm and oil spills.

Data preprocessing for correlational study of rescue event and weather data

To assess the depth of the relationship between weather and rescue events, a correlational study was performed with the number of rescue events versus the weather variables of the dataset exported from the Statistics Estonia website. As the rescue data for the year 2014 was unusable and the weather data ranged from 2010 to 2014, the range 2010-2013 was used for this study. Firstly, all of the rescue events for this range were combined into one dataset. Then, considering that the weather data was on the monthly time resolution, the rescue events data was converted into the same resolution with the additional variable representing the number of occurred rescue events in the given month. Next, this new dataset was merged by months and years with the weather data. As a result, the final data frame included 48 rows (12 months for 4 years) with seven weather variables, year, month and a number of rescue events. In addition, Tallinn and Tartu were chosen from the initial seven cities, as they are the two biggest cities in Estonia and they are located considerably far from each other (in terms of the country size). These were specified when subsetting from initial datasets.

Data preprocessing for statistical analysis with Sign test and time series

For the Sign test, two data sets were used: rescue events data and the data of parties and celebrations. Three years, 2010 to 2013, were studied separately and each of them was analyzed in three-time resolutions: by months, by weekdays and by weeks. The aim was to find two sequences of the means of rescue events in these resolutions for days with parties and without. So, for each observation in the time resolution (in a loop through each month, weekday and week) a merged dataset by date was created. Both were prepared before that. Combined data frame included three columns: Date, Number of public events and Number of rescue events. Then a mean number of all rescue events was calculated for dates when at least one public event has occurred. Another mean was for rescue events on days without any public events. The last one was also adjusted by the number of days when neither rescue and public events have occurred. As result, in the case of monthly time resolution, there were two sequences of length 12 and 54 in the case of weeks. These were later transferred to Excel file for better visual representation and more comfortable analysis.

For testing and time series plots, the chosen sequences were taken back to R.

Data preprocessing for logistic regression analysis of rescue events, public events and weather data

For the logistic regression analysis rescue events, public events and weather datasets were used. All three datasets were merged into one data frame by the Date variable. Rescue events data was converted into data set with two variables: Date and Number of rescue events occurred on that day. Public events data was transformed in a similar way – two columns: Date and Number of public events.

The weather dataset from the National Center for Environmental Information website included observations on a daily time resolution, so it was very suitable for this analysis. The initial dataset included precipitation, depth of the snow, maximum, minimum and average temperature as indicators of the weather conditions. However, considering that there were too many missing values in the depth of snow and minimum temperature variables, they were excluded from the analysis (Figure 5). In addition, considering that the dataset included a suitable amount of precipitation data only from the five stations, only two were chosen that are located near the biggest cities of Estonia: Tallinn and Tartu. Also, initially all temperatures were in Fahrenheit, so they were converted into Celsius.

If there was some date missing from one initial dataset, but present in another, then it was still included in resulting merged data set, with the missing value of the corresponding variable. For example, if there were no rescue events and public events on the 10th of the April, then only weather data was filled. The value of the public events variable was then changed to 0, which means no public events occurred on this date. So, as the weather data included all dates of the chosen years the new data frame also included all these days. Considering that the logistic regression is a method used to predict categorical variable, an additional column was created with two different possible values. It was either 0 or 1 for each day, based on the column “Number of rescue events”. If there was at least one rescue event, then the value was set to 1, and if the value was missing, then 0. The initial column showing the number of rescue events was then removed, as it wasn’t needed in the further analysis.

5 Results

In the previous chapter, we gave an overview over the datasets used in this study. Described their collection processes, explained limitations, and how they were preprocessed. In Chapter 5 the focus is on the results and important findings of this research supported by the results from the papers discussed in the relevant literature.

5.1 Relationship between weather and rescue events data

All years have been studied separately and combined together. Some years showed different results. For example, for the Tallinn dataset, the correlation between fire rescue events and precipitation in 2010 was 0.01, but in 2013 it was -0.26. Also, the correlation for relative humidity in 2010 was 0.55, which is already a moderate uphill relationship, but on the other hand in 2013 it was -0.49 and it means a moderate downhill relationship. So, it was decided to make conclusions only based on the correlation results of all years combined as it will consider the results from each year and indicate more accurate relationships.

Table 3. Correlations of weather indicators and rescue events of Tallinn in 2010-2013

	All events	Road accidents	Nature caused	Injuries and accidents	Fire accidents	Fire alarms	Other events
Precipitation	0,08	0,25	0,21	-0,01	-0,16	0,30	0,26
Number of rainy days	-0,54	0,03	0,38	-0,08	-0,61	0,32	-0,50
Average temperature	0,68	0,23	0,01	-0,02	0,49	0,03	0,78
Maximum temperature	0,77	0,18	-0,09	-0,04	0,63	-0,01	0,81
Minimum temperature	0,67	0,23	0,12	-0,03	0,46	0,09	0,72
Number of sunshine hours	0,76	0,14	-0,25	0,02	0,71	-0,27	0,79
Average relative humidity	-0,69	0,00	0,27	-0,02	-0,74	0,40	-0,63

Table 3 shows how different types of rescue events correlate with the weather data. The table is colored based on the values – stronger green or red color means stronger relationship. Strong green color indicates positive relationship and strong red shows a negative relationship. Yellow color means that there is no or very weak linear relationship between the variables.

It can be seen, that “Road accidents” and “Injuries” have no or very weak relationship. The variable “All events”, which is all rescue events combined, has the strongest relationships and almost with all of the weather indicators. It only has no relationship with the precipitation. It seems that it partly inherits its strong correlation from the “Fire accidents”, as it has very similar results. I. Vares already stated in her thesis, that there is a relation between fires and temperatures, and this correlation approves that.

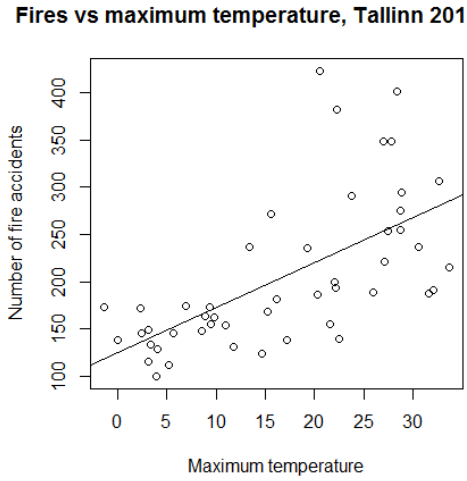

Figure 7. Fires vs maximum temperature, Tallinn 2010-2013

On Figure 7 the number of fire accidents is plotted against the maximum temperature. The line shows a strong uphill linear (positive) relationship. This correlation can also be seen from histograms on Figure 8. The left histogram shows a monthly distribution of "Fire accidents" and the right one "Maximum temperature". As was already stated in I. Vares thesis, there is a steep leap in April, when the number of fire accidents increases drastically and stays the highest until the end of the summer. The same is with the maximum temperatures. So, it can be assumed that on warmer days, people deal with open fires more often and it leads to the growth in the number of fire accidents. In addition, accordingly to the weather theory, on higher temperatures, nature fuels ignite and burn faster because less heat energy is needed to raise the fuels to their ignition temperature. But fire alarms are in some relationship with precipitation, humidity, rain and sunshine hours. The most obvious thing that can be assumed, is that more fire alarms occur on days when fires are less likely to occur, so that the workers of the rescue stations can conduct practices.

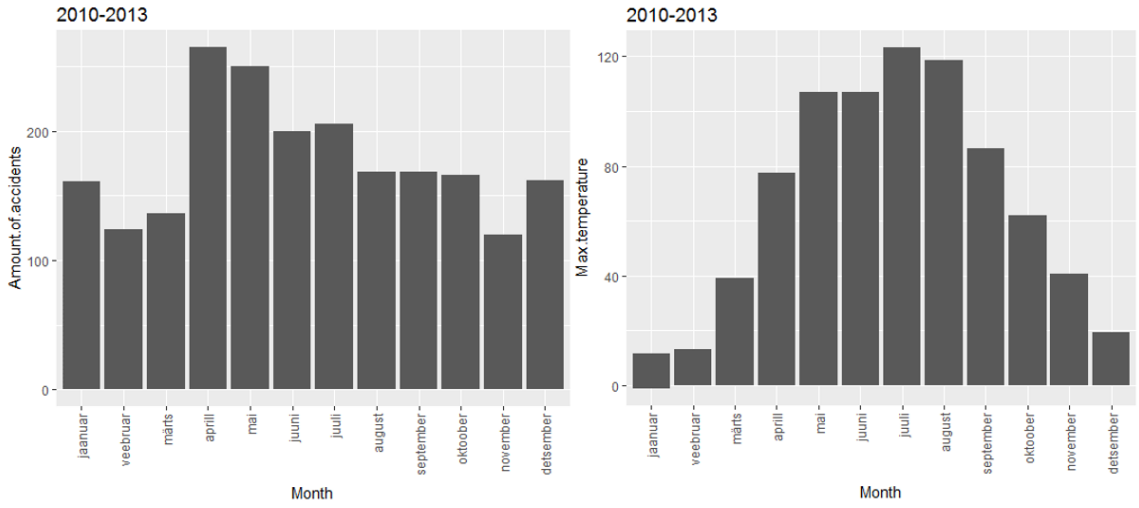

Figure 8. Monthly distribution of fires (left) and max. temperature (right) for the years 2010-2013

In the other hand, fire accidents have a strong downhill (negative) relationship with the number of rainy days and relative humidity, which is actually not a surprise. The humidity level is lowest on warmer days, which is when more of the fires occur. Same with the rainy days.

Table 4. Correlations between weather indicators and different fires of Tallinn

	Forest and land fires	Vehicle on fire	Fires in-side	Fires out-side
Precipitation	-0,23	0,17	0,12	0,16
Number of rainy days	-0,41	0,10	0,07	-0,18
Average temperature	0,20	0,13	0,16	0,47
Maximum temperature	0,35	0,05	0,06	0,40
Minimum temperature	0,19	0,13	0,16	0,45
Number of sunshine hours	0,45	-0,02	0,02	0,36
Average relative humid-ity	-0,56	0,02	-0,03	-0,30

As “Fire accidents” have the strongest correlations with the weather indicators, it was wise to take a closer look at different fire types and their relationship with the weather. Table 4 shows that there are four different fire types: “Forest and land fires”, “Vehicle fires”, “Fires inside” and “Fires outside”. The colors mean the same thing as in Table 3. It appears that fires related to vehicles and buildings have no linear relationship with the weather data. There is a very weak relationship between vehicles on fire and precipitation, and it can be assumed, that on days with higher precipitation there are a little more car crashes that result in getting on fire than usually. In the other hand, fires occurring outside or in forests have much stronger relationships. High temperature and longer sunshine can be considered as the main cause of these fires. But at the same time, rain and high humidity are factors that have the opposite effect.

Table 5. Correlations between weather indicators and other rescue events of Tallinn, 2010-2013

	Helpless ani-mal/bird	Others	Assis-tance	Oil spills	Less popular events
Precipitation	0,35	-0,10	0,26	0,00	-0,09915
Number of rainy days	-0,43	-0,06	0,10	0,55	-0,2739
Average temperature	0,71	-0,10	0,17	0,75	0,327841
Maximum temperature	0,76	-0,03	0,08	0,78	0,291745
Minimum temperature	0,63	-0,08	0,17	0,73	0,342549
Number of sunshine hours	0,73	0,03	0,02	0,73	0,292833
Average relative humid-ity	-0,56	0,01	-0,03	-0,64	-0,28791

Table 3 has also shown that there are strong correlations between weather factors and other rescue event types (last column named "Other events") that were not included in the table. So, another table was created (Table 5) where next four additional most occurred rescue events are presented. As can be seen, rescue events called "Helpless animal/bird" and "Oil spills" have very strong correlations with most of the weather factors. As temperatures increase and humidity decrease on warmer months, it can be assumed that the case with the animals and birds is that they are also more active in these months and as a result events when they need help occur more often. In addition, it appears that the weather has no relationship with the rescue events when people need help or assistance. In the given table the column "Others" indicates rescue events that the Rescue Board has named as "Others" itself, and seems that in terms of this correlational analysis they are not interesting. Lastly, the last column called "Less popular events" combines all other rescue events types that were not included in Table 3 and 5.

Table 6. Correlations between weather indicators and rescue events of Tartu, 2010-2013

	All events	Road accidents	Nature caused	Injuries and accidents	Fire accidents	Fire alarm	Less popular events
Precipitation	0,20	-0,03	0,53	-0,12	-0,08	0,02	0,28
Number of rainy days	-0,35	-0,06	0,18	-0,03	-0,50	0,08	-0,38
Average temperature	0,59	-0,03	0,24	0,16	0,39	0,01	0,72
Maximum temperature	0,67	0,00	0,25	0,09	0,50	-0,02	0,79
Minimum temperature	0,54	-0,08	0,28	0,14	0,36	0,06	0,70
Number of sunshine hours	0,61	0,08	0,01	0,02	0,60	-0,12	0,70
Average relative humidity	-0,47	-0,10	0,04	-0,09	-0,62	0,20	-0,57

Table 6 shows correlations between the number of different rescue events types and weather factors in Tartu. It can be seen that columns representing all events, injuries, fire accidents, less popular events have pretty much the same colors as in the table of Tallinn. So, we can assume that the location does not matter and the strong linear relationship remains the same everywhere. The column for road accidents is a little less green, but in the case of Tallinn, the relationships were still too weak to make any assumptions. But the results for nature caused damages are a little bit different, which is actually surprising. Although the distance between Tartu and Tallinn is about 170 km, in general, they should have similar weather conditions. But it seems that capital's close location to the sea affects the correlation with nature caused damages a little differently. In addition, fire alarms have also a little different trend in Tartu than in Tallinn – the correlation is much weaker.

Table 7. Correlations between weather indicators and other rescue events of Tartu, 2010-2013

	Helpless animal/bird	Others	Assistance	Oil spills	Less popular events
Precipitation	0,25	-0,17	0,44	0,05	0,20
Number of rainy days	-0,49	-0,07	0,14	-0,36	-0,16
Average temperature	0,62	-0,17	0,27	0,62	0,58
Maximum temperature	0,69	-0,09	0,17	0,66	0,62
Minimum temperature	0,54	-0,16	0,32	0,60	0,58
Number of sunshine hours	0,73	0,00	0,01	0,56	0,46
Average relative humidity	-0,59	0,00	0,03	-0,50	-0,39

Analysis of other less popular rescue events in Tartu (helpless animal/bird, others, assistance and oil spills) has shown only a little difference compared to the results of Tallinn (Table 7). For example, assistance correlation with precipitation and minimum temperature has increased to 0.44 and 0.32. So, it can be assumed that on colder days with higher precipitation, there are more people in the helpless condition. Although, there is no relationship with the average relative humidity. Another, but bigger difference, is between oil spills and a number of rainy days. In Tallinn, the correlation was 0.55, but in Tartu, it is -0.36.

In conclusion, using the correlational research method we have approved some findings made by I. Vares. There is a steep leap in the number of fire-related accidents in April, which stays high until the end of the summer. We have also demonstrated that the landscape, forest and outside fires affect this growth the most. In addition, fires have a strong negative relationship with relative humidity and the number of rainy days in the month. It means that on days with the rain or high humidity there is moderately less chance of fire occurrences. And this relationship is not dependent on location. Furthermore, road accidents and people injuries have no or very weak relationships with the weather indicators. But it seems that the relationship of nature caused damages is affected by the location, as it showed a little bit different results for two different cities, while other rescue event types showed pretty much the same results. Lastly, it appeared that there is a strong correlation between weather factors, and rescue events called “Helpless animal/bird” and “Oil spills”. Both of these have positive relationships with temperatures, precipitation and number of sunshine hours, and negative with average relative humidity and number of rainy days. Although, in Tartu, oil spills are in a moderate uphill relationship with the number of rainy days.

5.2 Relationship between public events and rescue events

To assess the relationship between public events and rescue events a statistical method called Sign test was implemented. The data of public events was extracted from Culture.ee website. Here the goal was to test for consistent differences between two pairs of observations. The first set contained the average number of different types of rescue events occurred together with specific public events – parties and celebrations. And the other represented the average number of rescue events of the same type, but on days without any parties and celebrations. Also, tests were performed in three different time resolutions: months, weeks and weekdays.

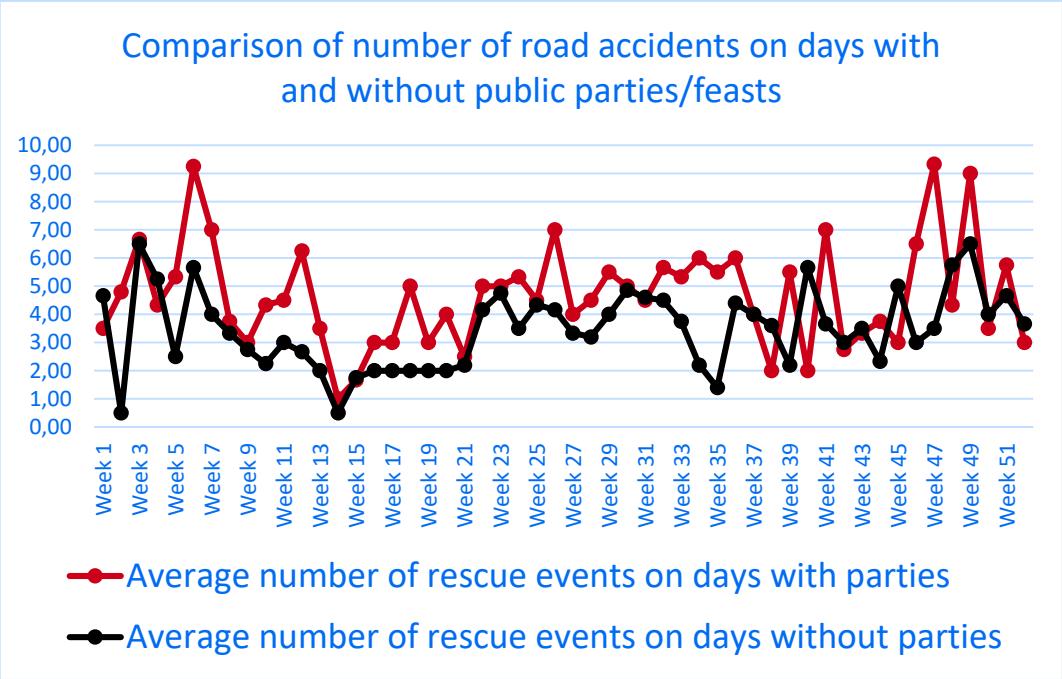

Figure 9. Comparison of mean number of road accidents on days with and without parties/celebrations in 2011

As an example, Figure 9 shows two sets of numbers representing the average number of road the accidents on each week of 2011. The red line indicates the mean number of accidents that occurred only on these days of the corresponding week when some parties or celebrations have occurred too. In the other hand, the black line shows the average number of rescue events that took place on other days, when there were no public events. E.g. on the first week there were on average about five road accidents on days with parties, and a little less than four on days without parties. If there were some days when neither of them occurred, then the second number was adjusted based on how many days of the week had at least one of these events.

The sign test was performed for six different types of rescue events: road accidents, nature caused damages, injuries, fire related events, people in the helpless state and all rescue events combined. As was already mentioned before, it occurred that the rescue event data for year 2014 was unusable. Considering the fact, that public data started from the middle of the year, it was decided to perform this statistical method only on years 2011 to 2013. In

addition, the tests were implemented for all rescue events and public events in the whole country combined, as there were very few parties and celebrations in cities and municipalities taken separately.

The first thing that stuck out pretty clearly is that creating the observation of weekdays was a pointless thing to do, as each sequence consisted only of seven observations (seven different weekdays), which is a very small number for a sign test and it gave bad results for each of the rescue event types on each year.

```
> binom.test(38, 51, 0.5, alternative = "greater")

        Exact binomial test

data:  38 and 51
number of successes = 38, number of trials = 51, p-value = 0.0003105
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.625584 1.000000
sample estimates:
probability of success
              0.745098
```

Figure 10. Sign test result for alternative hypothesis of fires

The sign test gave the best results for the fire-related rescue events. The null hypothesis was that there is no difference between fires and public parties/celebrations. But in year 2011in a weekly resolution there were 38 positive and 13 negative differences, which gives a probability of no difference p=0.0003105. So, the null hypothesis is rejected at a significance level of p=0.05. Furthermore, an alternative hypothesis stating that if there is a party or celebration, the number of fires will not increase, was also rejected with the p-value 0.0003105 (Figure 10). Both hypotheses were also rejected in a monthly time resolution. The year 2012 showed the same results. In the other hand, the year 2013 did not have such good results in a weekly time resolution, but the hypothesis was still rejected in the monthly sequences. So, we can conclude that there is a relationship between fire related rescue events and such public events like parties and celebrations.

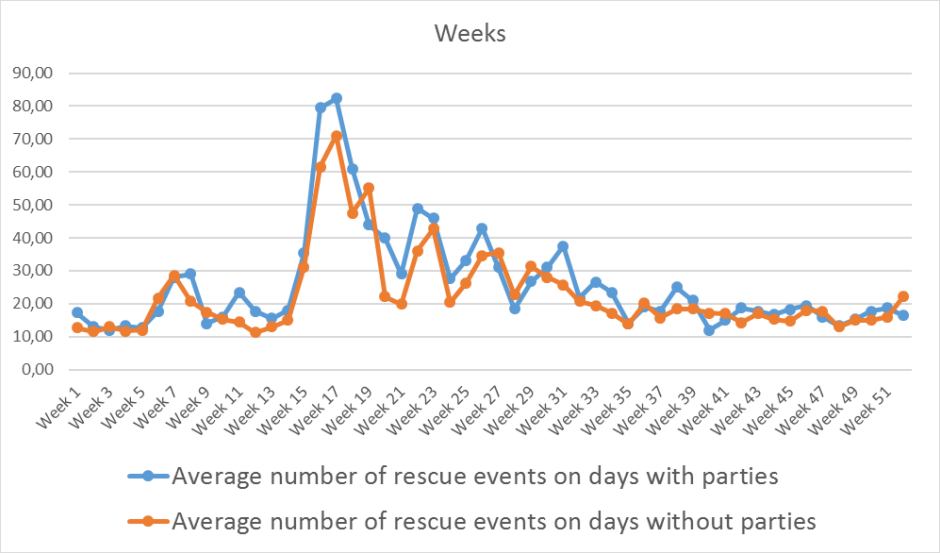

Figure 11. Comparison of mean numbers of fires on days with and without parties (2011)

It can also be seen in Figure 11, where the blue line, which indicates the average number of fires on days with parties in 2011, is almost always higher than the red line. This is why there are much more positive differences in sign test. But also, another thing that the graph shows, is that both blue and the red lines are very close to each other. In addition, they have a very strong positive linear relationship – 0.89. Figure 12 shows how these both sequences look when plotted against each other.

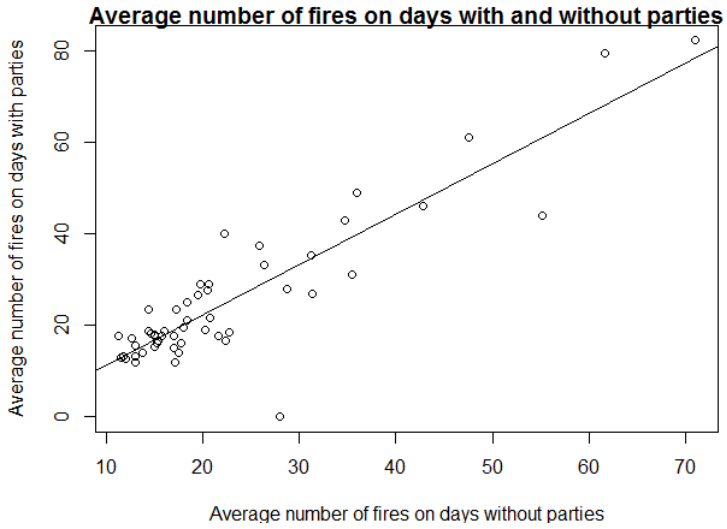

Figure 12. Averagenumber of fires on days with and without parties

Strong correlation and a linear relationship mean that although the blue line is a little bit higher and there are more positive differences, both sequences are very close to each other. It can be clearly seen in Figure 11 – the orange line is always very near to the blue line. And this means that there is always almost the same average number of fires both on days with parties/celebrations and without. So, there are no any consistent differences between

these observations. In years 2012 and 2013 the correlations are 0.59 and 0.72 respectively, which are not as high as the first one but can still be considered as strong ones.

Another rescue event type with good sign test results, but with a much smaller correlation value were road accidents. Figure 9 shows its weekly sequences, and it can be seen that lines are not as close as in the case with the fires, while the red line representing the average number of road accidents on days with parties is still visually higher than the black line. There were 38 positive and 13 negative differences with a p-value 0.0003105 rejecting the null hypothesis and proving that there is a significant relationship between road accidents and public parties in 2011. The null and greater alternative hypotheses were also rejected based on the data of 2012. In 2013 only the alternative hypothesis was rejected at a significance level of p=0.1. The null hypothesis was not rejected, as the p-value was 0.15.

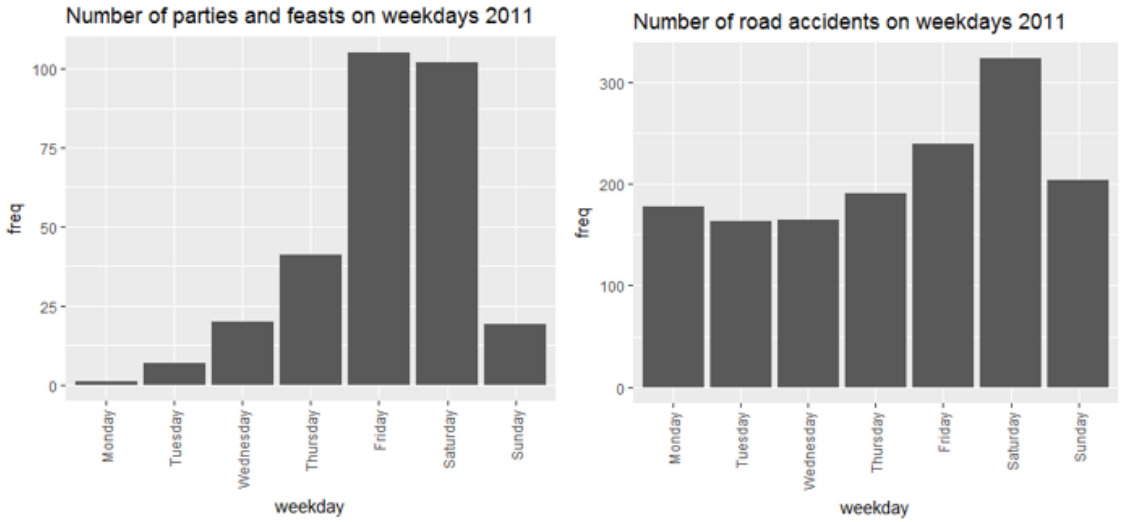

Figure 13. Distribution of parties (left) and road accidents (right) on weekdays (2011)

Two histograms on Figure 13 may explain the results of the Sign tests for road accidents. The left one shows the number of parties and celebrations on all weekdays of 2011. People have more free time at the end of the weekend, so most of the parties occur on Fridays and Saturdays. Sunday has less public events because it is the last day before the start of a new week. For some reason, there is a similar tendency with the road accidents. There are also more of them on Friday and Saturday, and less on Sunday. So, basically, the sequence of average numbers of road accidents occurred on days with the parties is higher, because these days with the parties are always the last few days of the week and there are always more road accidents.

```
Call:
lm(formula = paa ~ Weekday:pidu, data = newdata)

Residuals:
   Min     1Q Median     3Q    Max
-5.906 -1.740 -0.740  1.260 19.162

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.73998    0.09899  37.783   <2e-16 ***
WeekdayMonday:pidu   -0.08097    0.31339  -0.258   0.7962
WeekdayTuesday:pidu  -0.15130    0.12167  -1.244   0.2139
WeekdayWednesday:pidu  0.15869    0.21375   0.742   0.4580
WeekdayThursday:pidu   0.01021    0.14035   0.073   0.9420
WeekdayFriday:pidu     0.03021    0.03949   0.765   0.4445
WeekdaySaturday:pidu   0.09845    0.04080   2.413   0.0160 *
WeekdaySunday:pidu     0.21404    0.12334   1.735   0.0829 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.81 on 1088 degrees of freedom
Multiple R-squared:  0.01005,   Adjusted R-squared:  0.003678
F-statistic: 1.577 on 7 and 1088 DF,  p-value: 0.1381
```

Figure 14. Logistic regression model

In addition, Figure 14 shows a logistic regression model, where weekday as a factor together with the number of public events (parties and celebrations) is a predictor for binomial variable stating if a road accident will occur or not. And here we can see as well that Saturday and Sunday are statistically significant.

Lastly, rescue event type “people in helpless state” has shown some other interesting results. The first half of the year 2011 did not have these events at all (seems like one more problem with the data) – they start from the week 25, so this year gave bad results. But both years 2012 and 2013 included it fully from the first week to the last. The Sign test for year 2012 has shown a p-value 0.097 for a greater alternative hypothesis, so it can be rejected with a significance at 0.1. But the next year rejected a hypothesis which stated that if there is a party or celebration, the number of rescue events will not decrease. There were only 9 positive and 38 negative differences in the sequences.

In conclusion, the sign test has shown that there is a difference between fire accidents and public events, such as parties and celebrations. Furthermore, the rejection of alternative hypotheses has meant that on days with these public events there are more fire related rescue events. But the fact that both lines on Figure 11 follow the same trend and are always close to each other, and that these sequences of numbers of fires on days with and without public events are in a strong linear positive correlation, means that there is only a small difference between these variables. Although there are more fires on days with parties and celebrations, the actual difference is very small. The Sign test has also shown a relationship between road accidents and public events, and it appeared that it is probably due to the fact that both of these occur more often in the end of the week. Furthermore, the tests for rescue events when people were in a helpless state indicated that in 2012 there are more of these accidents on days with parties and celebrations. But in 2013 the situation was the opposite – fewer rescue events on days with public events. So, in this case, more data for other years should be analyzed to make any conclusion. Finally, all rescue events combined, nature caused damages and people injuries did not show any relationships with parties and celebrations.

5.3 Time series of fires and road accidents

From the previous analysis, we have found pairs of observations that have shown some relationship. These were road accidents and fires with parties and celebrations. Another thing that can also be done with their sequences of numbers of rescue events on days with

and without public days is to plot their time series and analyze for trends and seasonal patterns.

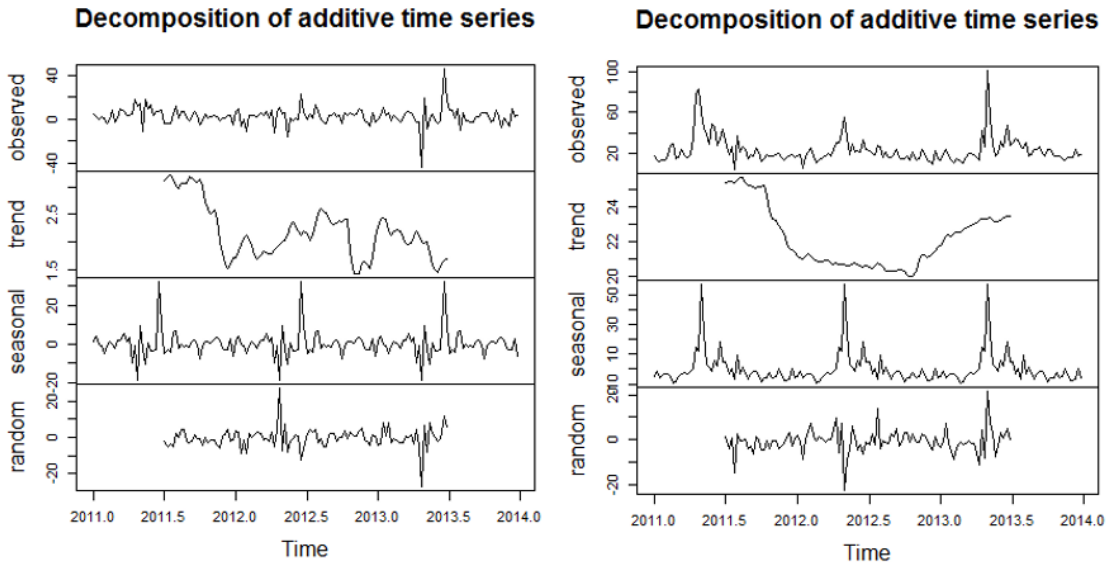

Figure 15. Time series of fires

On Figure 15 are shown two time series for fires and public events from 2011 to 2013. The left one is a sequence that represents the subtraction of numbers of fire accidents occurred on days with parties/celebrations and a number of same events without any parties. As there were more rescue events on days with public events, the sequence is mostly positive. The top graph shows the original time series. Somewhere in the middle of the year 2013, there are two biggest leaps down and up. First, there was some week when there were a lot of fires on days without public events, and shortly after that, there was a week with a lot of fires on days with parties. Something similar, but with much smaller changes, happened in the middle of 2012. Next component shows an estimated trend of the difference, and we can see a decrease until the start of 2012. Then it increases until a gap that lasts for a few months and decreases again. Here it would be better to have more years to observe to make any accurate assumptions. From the third graph, we can see that there is indeed a seasonal pattern in the middle of each year.

Time series graphs on the right represent the sequence of a number of fire accidents occurred on days with parties and celebrations. Here we can see how this number changes over time. The estimated trend component shows that from 2011 to the beginning of 2013 the number of fires on days with public events has constantly decreased. After that, it started increasing. From the seasonal graph, we can see again the pattern of fire increases in April.

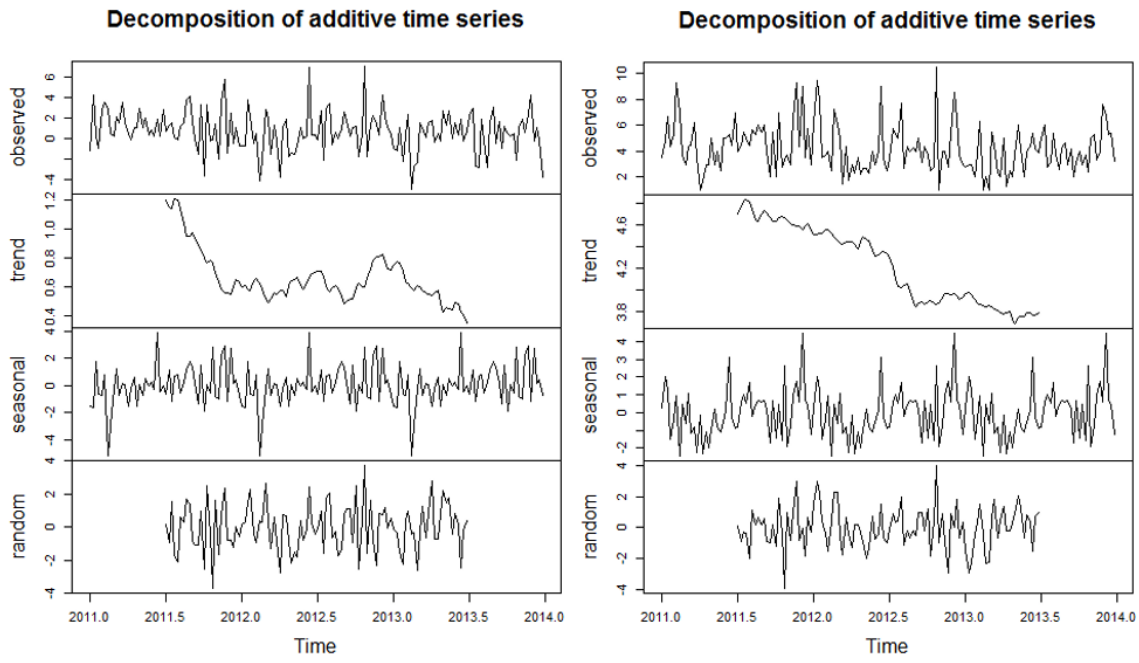

Figure 16. Time series of road accidents

Figure 16 is similar to the previous one, but here are the time series of road accidents and public events. Left on is the subtraction of numbers of road accidents occurred on days with parties and celebrations, and numbers of the same rescue events occurred on days without these public events. Similarly to fires, there is a decrease in rescue events on days with public events up until the beginning of the year 2012. Then it stays at about the same level for a year, increases a bit and starts decreasing again. The seasonal graph shows that at the beginning of each year there are some weeks when more road accidents occur on days without public events, but then somewhere around the May and in the end of the year on days with parties.

The right decomposition represents only the number of road accidents on days with public events. Here the most interesting component is estimated trend time series. Throughout the whole period of 2010-2013, the number of rescue events always decreases. The seasonal graph has too many fluctuations, so it is difficult to make any conclusions.

In conclusion, the seasonal graph of fire time series has shown again the increase of fires in April. There seems to be also some time in the summer (around June) when the number of fires on days with parties is much bigger than on days without. The number of fires has also decreased from 2011 to the beginning of 2013, but then it started to increase again. The time series of road accidents have shown that there is a trending decrease in the number of these rescue events on days with public events. Although these time series show some interesting trends and conclusion can be made, it would be good to plot time series for bigger time ranges.

5.4 Predicting rescue events occurrence

The purpose of the analysis was to try to predict if the rescue event of different types would occur based on the weather and public events data. The method was used for five types of rescue events: road accidents, nature caused damage, injuries, fire related accidents and all types of rescue events combined. It was also run for Tallinn and Tartu.

```
Public.events     Precipitation      Avg.temperature    Max.temperature   Rescue.event.occurs
Min.    : 0.000   Min.    :0.00000   Min.    :-26.670   Min.    :-20.56   0:329
1st Qu.: 2.000   1st Qu.:0.00000   1st Qu.: -0.560   1st Qu.:  3.33   1:767
Median : 5.000   Median :0.01000   Median :  6.670   Median : 11.11
Mean   : 5.997   Mean   :0.06563   Mean   :  6.426   Mean   : 11.07
3rd Qu.: 9.000   3rd Qu.:0.08000   3rd Qu.: 14.440   3rd Qu.: 20.00
Max.   :37.000   Max.   :1.30000   Max.   : 26.110   Max.   : 32.22
```

Figure 17. Summary of merged (fires, weather, public events) Tartu data

As an example, Figure 17 shows a summary of the merged data set from rescue events, weather and public events data sets for fire related accidents happened in Tartu city in 2011 to 2013. As can be seen, throughout these three years there are days with zero and thirty-seven public events at most, but on average there were about six events per day. The coldest day is with a temperature -26 degrees Celsius. As Estonia is not a very hot country, the maximum temperature for these three years was only 32 degrees. But on average the temperature is still close to six. Also, the distribution of days with and without fire related rescue events is not ideal but considerably good – 329 days with and 767 days without fire events.

The data set was then split into training and testing datasets, 70 and 30 percent respectively. So, the resulting training data set included 776 random records and the other one 320 records. The training set was used to fit the model which was later tested by using the testing set.

Table 8. Distribution of classes for Tallinn data

	All events	Road accidents	Nature caused	Injuries and accidents	Fires
Days with rescue events	1096	421	929	855	1089
Days without rescue events	0	675	167	241	7

Table 9. Distribution of classes for Tartu data

	All events	Road accidents	Nature caused	Injuries and accidents	Fires
Days with rescue events	1065	210	57	256	767
Days without rescue events	31	886	1039	840	329

While conducting the analysis it appeared that the classes are quite unbalanced. Tables 8 and 9 show the distributions of how many days are with and without each type of rescue events in Tallinn and Tartu. The problem with Tallinn is that there are no or very little days without fires and all rescue events combined. So, these were left out of the analysis.

In addition, for both cities, injuries and accidents did not have any statistically significant variables. We can also recall that both correlational study and sign test analysis haven't shown any relationship with this type of rescue events as well.

```
Call:
glm(formula = Rescue.event.occurs ~ Public.events + Max.temperature +
    Precipitation + Avg.temperature, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9041  -1.3513   0.7414   0.8509   1.5202

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.48803    0.18528   2.634 0.008440 **
Public.events   -0.01290    0.01607  -0.802 0.422402
Max.temperature  0.11288    0.02830   3.989 6.63e-05 ***
Precipitation   -1.32759    0.53725  -2.471 0.013470 *
Avg.temperature -0.10751    0.03053  -3.521 0.000429 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1001.5  on 821  degrees of freedom
Residual deviance:  975.1  on 817  degrees of freedom
AIC: 985.1

Number of Fisher Scoring iterations: 4
```

Figure 18. Summary of the model predicting fires in Tallinn

On Figure 18 is shown a summary of the fitted model with all four initial variables as predictors for the fire rescue events. First of all, public events are not statistically significant, as its p-value is only about 0.4. As for the statistically significant variables, maximum temperature has the lowest p-value, 6.63×10^{-5}, suggesting a strong association of the maximum temperature with the probability of fire related rescue event occurrence. Another significant variable is the average temperature with the p-value 0.000429, which is also much lower than 0.05. The negative coefficient for the precipitation predictor suggests that all other variables being equal, the fire accident is likely to occur when precipitation is lower.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Rescue.event.occurs

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             775     941.53
Public.events   1   0.0993       774     941.43 0.7526769
Precipitation   1   5.7105       773     935.72 0.0168642 *
Avg.temperature 1   2.2148       772     933.51 0.1366937
Max.temperature 1  11.5467       771     921.96 0.0006787 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 19. Analysis of deviance table

Then we performed an Analysis of Variance, to assess variables on how they improve the model (Figure 19). The difference between the residual deviance and the null deviance shows how our model is doing against the model with only the intercept (a null model). The wider this gap, the better. From the figure, we can see the drop in deviance when adding each variable one at a time. Again, adding maximum temperature, average temperature and precipitation reduce the residual deviance. The public events variable seems to improve the model less, so it can be considered to be removed from the model. Based on these results we can make an assumption, that the number of public events is not significant when we try to predict the occurrence of the fire-related rescue event.

```
Call:
glm(formula = Rescue.event.occurs ~ Precipitation + Max.temperature +
    Avg.temperature, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0219  -1.3899   0.7505   0.8383   1.4001

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.51303    0.15011   3.418 0.000632 ***
Precipitation   -1.21692    0.55375  -2.198 0.027978 *
Max.temperature  0.09366    0.02792   3.355 0.000795 ***
Avg.temperature -0.08533    0.03008  -2.837 0.004559 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 941.53  on 775  degrees of freedom
Residual deviance: 922.10  on 772  degrees of freedom
AIC: 930.1

Number of Fisher Scoring iterations: 4
```

Figure 20. Summary of the model predicting fires in Tartu

As the main purpose of logistic regression analysis is to find the best model, there are situations, when the set of explanatory variables to be included is not predetermined and selecting them becomes part of the analysis. Considering that the public events variable had a low p-value and did not improve the model, it was removed. So, an automatic Forward variable selection method was used, where the model with only the intercept was selected as a null model and the model with all variables against each other as full model. Figure 20 shows the summary of the best model outputted from the variable selection. It includes variables: average temperature, maximum temperature, and precipitation. Here, all variables are statistically significant, as they all have low p-values. Also, we can see that this model is definitely better than the one on Figure 18 because there is a drop in residual deviance and AIC value.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 267  50
         1   1   2

               Accuracy : 0.8406
                 95% CI : (0.7958, 0.879)
    No Information Rate : 0.8375
    P-Value [Acc > NIR] : 0.4766

                  Kappa : 0.056
 Mcnemar's Test P-Value : 1.801e-11

            Sensitivity : 0.99627
            Specificity : 0.03846
         Pos Pred Value : 0.84227
         Neg Pred Value : 0.66667
             Prevalence : 0.83750
         Detection Rate : 0.83437
   Detection Prevalence : 0.99062
      Balanced Accuracy : 0.51737

       'Positive' Class : 0
```

Figure 21. Confusion matrix and statistics of nature caused damages in Tallinn

The imbalance of classes has also caused a so called accuracy paradox. Figure 21 shows a confusion matrix and statistics for logistic model predicting the probability of occurrence of nature caused damages in Tallinn based on its most significant values: precipitation and average temperature. We can see that the accuracy is pretty high – 0.84. The problem is that the model looks at the data and decides that the best thing to do is to always predict most popular class and achieve high accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 230  41
         1  38  11

               Accuracy : 0.7531
                 95% CI : (0.7021, 0.7994)
    No Information Rate : 0.8375
    P-Value [Acc > NIR] : 1.000

                  Kappa : 0.0714
 Mcnemar's Test P-Value : 0.822

            Sensitivity : 0.8582
            Specificity : 0.2115
         Pos Pred Value : 0.8487
         Neg Pred Value : 0.2245
             Prevalence : 0.8375
         Detection Rate : 0.7188
   Detection Prevalence : 0.8469
      Balanced Accuracy : 0.5349

       'Positive' Class : 0
```

Figure 22. Confusion matrix and statistics of nature caused damages after optimization

One solution to this problem is to optimize the model by over- or under-sampling the training data. Figure 22 shows the confusion matrix after applying both methods. The accuracy has decreased, but at the same time, the specificity has increased.

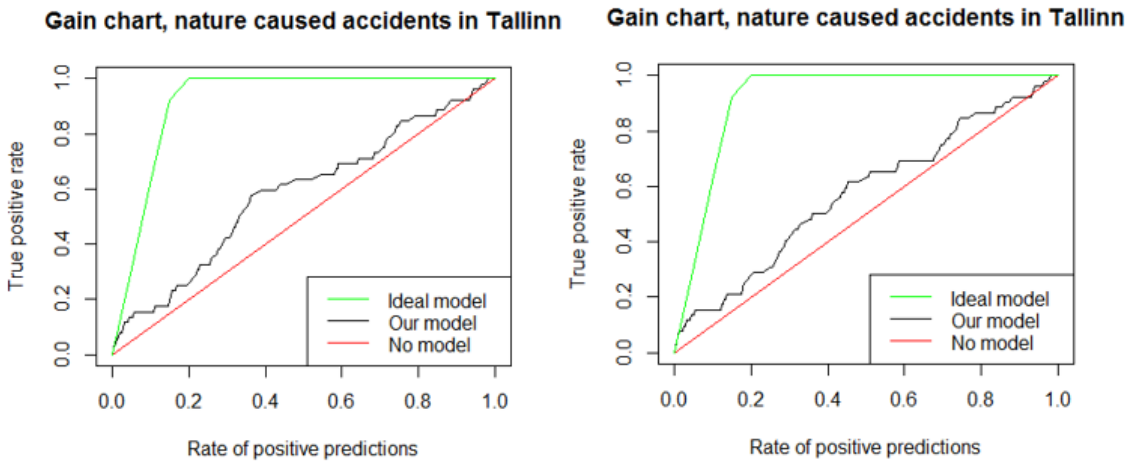

Figure 23. Cumulative gain charts of nature caused accidents in Tallinn

Figure 23 shows two charts of cumulative gains. The left one is the cumulative gain of the first model. The chart has three lines: green represents the ideal model, which predicts accurately 100% of the time, the red is the worst case or the result of random guessing, and the black line is our model. We can see that it does very well for about 15-18% when it first increases together with the ideal case. So, we could predict correctly about 15-18% of all nature caused damages in Tallinn. After that, the model predicts almost randomly. The chart on the right is model after optimization. There is a small difference in the middle of the graph, but it didn't improve the model significantly.

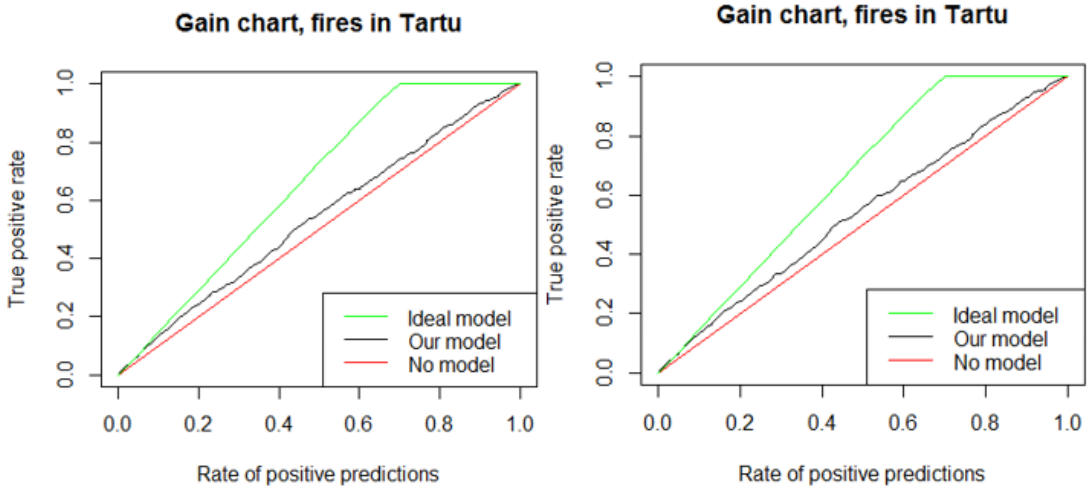

Figure 24. Cumulative gain charts of fires in Tartu

Figure 24 represents fire related rescue events in Tartu. The left one is again without any optimizations to the model. But we can already see that in the beginning, it is pretty close to the ideal model. The right one, which is optimized, got only very small improvements.

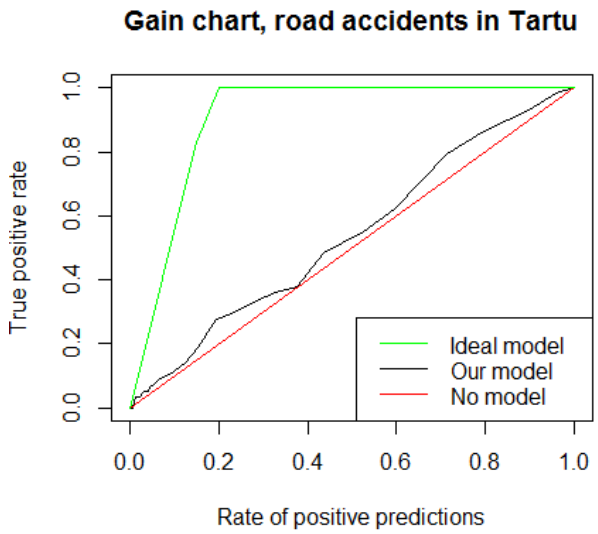

Figure 25. Cumulative gain chart of road accidents in Tartu

These previously described models (Figure 23 and 24) are best ones out of all others, although they are not actually very good too. Figure 25 shows one of the other bad models, which didn't improve after optimization as well. The black line representing our model is very close to red line and is far from the ideal model. The accuracy without optimization was 0.82 and sensitivity close to 1, while specificity was 0. Optimization dropped accuracy to 0.49 and sensitivity to 0.48 but increased specificity to 0.55. Statistically most significant and the only predictor was the variable "Public events", which is surprising considering that road accidents should be in at least moderate relationship with the temperature indicators. It was like that with the road accidents in Tallinn, but the resulting model was still bad even after improvements.

In conclusion, although the logistic regression models had some statistically significant predictors, like maximum and average temperature in case of fires in Tartu or precipitation and average temperature for nature caused damages in Tallinn, their predictive power was not strong. Based on the results we could predict with confidence about 16% of nature caused rescue events in Tallinn, and about 19% of fires in Tartu.

6 Conclusions

The correlational study has first shown us that analyzing the years separately might give us not quite accurate results. It was because some correlations differed greatly on different years. So, it was decided to make conclusions only based on the results gathered from the analysis of all years combined. The rescue event type "Injuries" did not show any correlations with weather factors at all. Furthermore, it did not show any relationship with anything in all used statistical methods. In the other hand, fires were in strong uphill linear relationship with temperatures and sunshine hours, and also strong downhill relationship with the number of rainy days and humidity. This approves the conclusion made by I. Vares and J. Horm in their theses. Also, the results for "Fire alarm" has showed that that there are more alarms on days with higher humidity and smaller sunshine hours which is the time when less fire occur. So, it can be assumed, that rescue workers perform fire alarms trainings and practice on days when it is less likely that the fire will occur. After analyzing the fires in more depth it was found that the fire type that correlates the most with the weather indicators is "Fire outside". This rescue event occurs more often on warmer days. The rescue event type that showed the strongest relationship with weather was "Helpless animal/bird", and it was in strong positive correlation with temperatures and sunshine hours. Here we can assume that there are more of these rescue events on warmer and dryer days because this is the time when the animals and birds are the most active. Also, the oil spills were strongly positively correlated with temperatures and rainy days. In Tartu almost all of the rescue events gave the same results, but there were some other interesting results as well. For example, the oil spills in Tartu were in the moderate negative relationship. Lastly, in both cities, the event type "Assistance" has shown a little higher than weak correlation with precipitation. Based on that we can assume that on days with higher precipitation people need help and assistance more often.

The sign test showed strongest relationship between fires and parties. But the further analysis has shown, that the number of fires on both days with and without parties follow the same pattern, as if there is always a linear increase or decrease. It was also proven by plotting these sequences against each other and finding their correlation. The value of 0.84 can be considered as very strong, and it says that in case of public events it doesn't matter if there is a party on no, the number of fires stays on the same level. But at the same time the sign test has shown us that the number of road accidents depends on party occurrences. Later it appeared that both parties and road accidents occur more often in the of the week, and this is the reason of the relationship. Plotting their distributions on weekdays and fitting a regression model approved this theory. There were also some interesting findings for "People in helpless state", but there were only two years of data for this, so a bigger time range of data is needed to make any conclusions.

From the time series of the mean sequences created while conducting the sign test analysis we found some other interesting observations. The seasonal component of fires showed us again the increase in the number of fires in April. In addition, the number of fires was decreasing from 2011 to the beginning of 2013, but then started increasing. Here, a bigger time range is also needed, so the results would be more descriptive and accurate.

In logistical regression analysis the main problem was that the classes were highly unbalanced and the model made always predictions towards the bigger class. Furthermore, although some of the variables were statistically significant, even after over- and under-sampling (optimization of the unbalanced data), the model did not improve. The best model from all of these was the model predicting fires in Tartu, and from the cumulative gain chart

showed that for about 18-19% the model was close to ideal model. Also, the model for nature caused damages in Tallinn was near perfect for about 15% of the data. So, we can tell for sure that there is potential, but probably the rescue events data for a bigger time range is needed, or some other additional datasets that will help weather and public events to make better predictions.

In conclusion, we can see that using additional datasets, such as weather and public events data, has definitely shown us some new interesting findings about the rescue events. There is even some potential in using predictive analytics. So, for the Rescue Board it is definitely encouraging to use other data as it will help with better understanding of their data and better management of the rescue work.

References

[1] "Rescue Board," 5 12 2016. [Online]. Available: https://www.eesti.ee/eng/kodakondsus/turvalisus/paasteamet. [Accessed 10 10 2016].

[2] Rescue Board Yearbook 2015, Tallinn: the Rescue Board, 2016.

[3] "Wildfires, weather & climate," [Online]. Available: https://www2.ucar.edu/news/backgrounders/wildfires-weather-climate. [Accessed 18 May 2017].

[4] A. Furnas, "Everything You Wanted to Know About Data Mining but Were Afraid to Ask," The Atlantic Monthly Group, 3 April 2012. [Online]. Available: http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/. [Accessed 11 November 2016].

[5] M. Rouse, "business intelligence (BI)," TechTarget, [Online]. Available: http://searchdatamanagement.techtarget.com/definition/business-intelligence. [Accessed 10 November 2016].

[6] "What is Web Scraping?," [Online]. Available: https://www.webharvy.com/articles/what-is-web-scraping.html. [Accessed 18 May 2017].

[7] Rescue Board Yearbook 2014, Tallinn: the Rescue Board, 2015.

[8] "Map of the rescue stations," Estonian Rescue Board, 9 September 2015. [Online]. Available: https://rescue.ee/et/kodanikule/komandodekaart/. [Accessed 10 November 2016].

[9] I. Vares, Spatial and temporal pattern of The Estonian Rescue Board fire accidents in periood 2009-2013, Tartu: University of Tartu, 2014.

[10] J. Horm, Rescue event categorization based on Estonian Rescue Services data from 2010–2013, Tartu: University of Tartu, 2016.

[11] "Review of Fire and Rescue Service response times," Communities and Local Government, London, 2009.

[12] "R (programming language)," [Online]. Available: https://en.wikipedia.org/wiki/R_(programming_language). [Accessed 18 May 2017].

[13] "Correlation and dependence," [Online]. Available: https://en.wikipedia.org/wiki/Correlation_and_dependence. [Accessed 18 May 2017].

[14] "Sign Test," [Online]. Available: https://www.statisticssolutions.com/non-parametric-analysis-sign-test/. [Accessed 18 May 2017].

[15] "Time series," [Online]. Available: https://en.wikipedia.org/wiki/Time_series. [Accessed 18 May 2017].

[16] "What is Logistic Regression?," [Online]. Available: http://www.statisticssolutions.com/what-is-logistic-regression/. [Accessed 18 May 2017].

[17] "Correlation," Creative Research Systems, [Online]. Available: http://www.surveysystem.com/correlation.htm. [Accessed 10 November 2016].

[18] "Analysis of fire and rescue service performance and outcomes with reference to population socio-demographics," Communities and Local Government, London, 2008.

[19] M. Alice, "How to perform a Logistic Regression in R," 13 September 2015. [Online]. Available: https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/. [Accessed 18 May 2017].

[20] "Confusion matrix," [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix. [Accessed 18 May 2017].

[21] "Lift Chart (Analysis Services - Data Mining)," 2 March 2016. [Online]. Available: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining. [Accessed 18 May 2017].

[22] "Municipalities," Estonian Ministry of Finance, 1 January 2015. [Online]. Available: http://www.fin.ee/kov. [Accessed 10 November 2016].

[23] "Administrative and minucipality division," Estonian Land Board, 23 September 2016. [Online]. Available: http://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Haldus-ja-asustusjaotus-p119.html. [Accessed 10 November 2016].

[24] "Statistics Estonia," [Online]. Available: http://www.stat.ee/en. [Accessed 18 May 2017].

[25] "National centers for environmental information," [Online]. Available: https://www.ncdc.noaa.gov/about. [Accessed 18 May 2017].

[26] "What is culture.ee?," [Online]. Available: http://www.culture.ee/en/kultuurikalendrist/. [Accessed 18 May 2017].

[27] "Beautiful Soup Documentation," [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed 18 May 2017].

License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Vladimir Visbek**,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

Linking Rescue Event Data with Public Data,

(title of thesis)

supervised by Siim Karus,

(supervisor's name)

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **19.05.2017**