MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

INTERNSHIP REPORT

BUILDING A RISK MAP FOR HURRICANE-FORCE TROPYCAL CYCLONES IN CONTINENTAL PORTUGAL

ANDREA HAUSER

SUPERVISION:

ALEXANDRA BUGALHO DE MOURA
CARLOS EDUARDO DA ROSA

DECEMBER - 2021

Acknowledgments

I would first like to thank my family. They gave me the opportunity to explore new directions in life and seek my own destiny. I am forever indebted to them for the encouragement and love received, this has made the person I am. This journey would not have been possible if not for them, and I dedicate this milestone to them.
I would like to acknowledge professors Alexandra Bugalho de Moura and Carlos Miguel dos Santos Oliveira for the guidance through these months and the numerous virtual-meetings we had. I am also very grateful to my supervisor in Fidelidade Carlos Eduardo da Rosa for the precious hints and ideas he gave me; I benefited enormously from his advice. My sincere thanks to Febio Nuno Alves Levezinho and Rui Alexandre Silva Esteves for the advices received during the meetings of the thursday afternoon and for giving me the opportunity of doing this internship.
I want to thank all my closest friends, the ones I met in Trieste, in Grado and in Lisbon, for the beautiful moments we lived together. Finally I want to thank Paola, for all her love and support.

Contents

Abstract

Tropical cyclones have enormous destructive potential. In 2018 continental Portugal has been affected by hurricane Leslie, the weather-related event having the highest impact ever on the property portfolio of the portuguese insurance company Fidelidade, causing several millions euros of losses. The fear is that, in the near future, the occurrence of this type of events increases in intensity and frequency, as a consequence of the climate change due to the warming of the planet. Quantifying the potential loss to which the property portfolio of Fidelidade could be subject to, helps in approximately determining premiums and capital reserves, as well as in defining the coverage to be provided.

In this work, an approach to model the costs caused by a tropical cyclone extreme event is presented. The model is based on the losses incurred by the property portfolio of Fidelidade due to hurricane Leslie. By using the estimated models, it is possible to produce cost estimates for different scenarios of interest for the company. The estimated models are also used to build a risk map for the councils of continental Portugal.

The results obtained indicate that the councils with the estimated higher average cost ratio are all located along the coast of the country.

Keywords - Property Insurance; Tropical Cyclones; Claims; Cost; Regression Tree; Random Forest; Logistic Regression; Multiple Linear Regression; Risk Map

Resumo

Ciclones tropicais têm um enorme potencial de destruição. Em 2018, Portugal continental foi atingido pelo furacão Leslie, que constituiu o fenómeno metereológico de maior impacto, até à data, no portfólio da companhia de seguros Fidelidade, causando milhões de euros em perdas. De facto, os ciclones tropicais têm um enorme potencial de destruição. A preocupação é que, em breve, a ocorrência deste tipo de fenómenos aumente em intensidade e frequência, como consequência das mudanças climáticas provocadas pelo aquecimento global. Quantificar a potencial perda à qual a companhia Fidelidade pode estar sujeita ajuda a determinar aproximadamente os prémios e provisões, assim como a definir a cobertura a ser providenciada.
Neste trabalho, é apresentada uma abordagem para modelar os custos causados por um ciclone tropical extremo. O modelo é baseado nas perdas provocadas ao portefólio da Fidelidade pelo furacão Leslie. Ao usar os modelos, é possível produzir custos estimados para diferentes cenários de interesse da companhia. Os modelos estimados são também utilizados para construir um mapa de risco para os conselhos de Portugal continental.
Os resultados obtidos indicam que os conselhos com a maior taxa média de custos estimada estão localizados ao longo da costa do país.

Palavras-chave - Seguro Multi-Risco; Ciclones Tropicais; Sinistros; Custos; Árvore de Regressão; Floresta aleatória; Regressão Logística; Regressão Linear Multipla; Mapa de Risco

All models are wrong, but some are useful
George Box

1 Introduction

Many studies have addressed the impact of natural catastrophes on economic losses, which include events like earthquakes, heath waves, hurricanes, floods and so on. As [14] reports, there is evidence that the amount of the losses caused by natural catastrophes has increased every year since 1980. Also [14] reports that the increase is predominantly attributable to weather-related events like storms and floods.
Nevertheless, [14] explains that part of the increase in the losses is caused by socio-economic/demographic factors, such as population growth, ongoing urbanization and increasing property and material values being exposed in hazard-prone areas. Because of such factors influencing the loss trends, it is very difficult to attribute at least part of the effect to global warming, and so to climate change. So, while there is evidence for increases in economic losses related to natural hazards, it is uncertain whether this is due to an increase in the number and intensity of extreme events, or if it can be attributed to socio-economic changes [8].
Natural hazards can represent a serious risk for the insurance sector. For example, only in the first half of 2021, the global insured losses from natural disasters have been of 42 billion dollars, 39% higher than the 21st Century average, which was 30 billion dollars, as reported by [1]. Among these disasters, major storms in western and central Europe in June 2021 caused at least 4.5 billion dollars in insured losses [1].
Hurricanes, which belong to the family of tropical cyclones, are among the most costly natural hazards [16]. As reported by Muncih Re [16], hurricane Katrina, which hit New Orleans in 2005, was the most costly natural disaster of all time for the insurance sector, with losses totalling more than 60 billion dollars. In 2017, the hurricanes Harvey, Irma and Maria caused record insured losses for more than 90 billion dollars within just four weeks. Tropical cyclones can be active for several weeks and can stretch across a large area, while wind speeds can reach more than 250 km/h and in some cases even exceeding 300 km/h [16].
Tropical cyclones occasionally affect Western Europe (1 storm in 1 or 2 years) [12, 15]. Since 1995, The National Oceanic and Atmospheric Administration (NOAA) has documented 6 tropical cyclones affecting continental Portugal.
In 2018 and in 2020, Portugal has been affected by two tropical cyclones called respectively hurricane Leslie and subtropical storm Alpha. The first caused to Fidelidade several millions of losses, while the second one caused approximately 20 times less costs than hurricane Leslie.
The fact of having observed two events of tropical origin in the past three years, leads the insurance company to be concerned in modelling accurately the expected loss that could derive from these type of events. Also, climatology studies based on model simulations, like [11], show an increase of hurricane-force storms of tropical origin over Western Europe during early autumn (Aug-Oct), as a consequence of the greenhouse warming. Thus, being able to quantify the potential losses due to these severe storms to which the portfolio of an insurance company could be subject to is crucial for insurers.
One possible approach to quantify the expected cost is to assess different scenarios to obtain estimates of the average loss expected from events of this type.

One method to quantify the loss is to use the damage function, which is a mathematical relation between the magnitude of a natural hazard and the average damage caused on a specific item (building, person, etc.) or portfolio of items [20].
In [21] the employed damage functions are calibrated against the daily insurance loss data due to storms affecting the residential buildings in Germany from 1997 to 2007. As measure of the intensity of the storm, the daily maximum wind gust data by the German weather service and from the ERA-Interim reanalysis project [1] for the period 1997 to 2007, are employed. In [6], the same insurance loss data and the daily maximum wind gust data from ERA-Interim reanalysis are considered but the damage function is calibrated considering just the significant losses related to large scale winter storms for the period 1997 to 2007. In both approaches damage functions that take as input the wind velocity to estimate the damage caused by the storm are employed. However these events can also bring heavy rains, which can produce floods. Because of this, also the rain amount should be considered as input for the damage function. In [20], alternative damage functions are studied, in order to assess the damage deriving not only from extreme winds, but also from floods and, for life-insurance applications, heath-related deaths.
Applying these methods, requires a solid base of historical losses and meteorological data, which are not always available. Without these data it is difficult, if not impossible, the calibration and evaluation of the damage functions. For this reason, it can be useful to develop an approach to estimate the losses for the company in different scenarios which does not require daily meteorological data or loss data relative to multiple severe past events to calibrate a damage function.
In this work a different approach to estimate the average costs caused by a hurricane-force tropical cyclone over the territory of continental Portugal is applied to the property insurance portfolio of the portuguese company Fidelidade. With this approach the insurer is able to estimate, without using damage functions, the average cost incurred by each policy in the portfolio, in the scenario where the event under study would affect a given part of continental Portugal. Finally, by simulating many different scenarios and averaging the cost estimates obtained, it is possible to define a risk map for continental Portugal.
The results obtained indicate that if hurricane Leslie would have affected the areas of Lisbon or Porto, the company would have incurred in approximately the double of the costs observed by the company in continental Portugal and due to Leslie. Instead, if it would have affected the area of Faro, the expected costs would be of approximately the half. On average, if hurricane Leslie would affect continental Portugal, the company might expect a 1.14 times higher cost than the one observed when Leslie affected Figueira da Foz.
The outline of this thesis is a follows: in Chapter 2 are discussed the difficulties that we can encounter when dealing with climate data and insurance loss data; afterwards the characteristics of the tropical cyclones that affected continental Portugal since 1995 and the loss data employed in this work are briefly described; in Chapter 3 the models used to describe the behaviour of hurricane Leslie are presented, in Chapter 4 3 different scenarios of interest for the insurance company are analyzed

[1] For more infromation about the Era Interim dataset consult https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim

while in Chapter 5 it is described how to employ the model estimated in Chapter 3 inside an algorithm to estimate the average exposure of each council when an event like hurricane Leslie affects continental Portugal. Conclusions are drawn in Chapter 6.

2 Climate Data and Loss Data

A damage function is defined as the mathematical relation between the magnitude of a natural hazard and the average damage caused on a specific item, such as buildings or a person, or portfolio of items [20]. Damage assessment typically relies on damage functions that translate the magnitude of extreme events to a quantifiable damage [20]. Damage functions need meteorological data about the magnitude of the natural hazard under study in order to be estimated.
The high spatial variability of phenomenon like weather-related natural hazards, makes it really difficult to capture the real magnitude of the hazard in each location of interest. This means that the correlation between the magnitude of the event and the damage observed could be weak. In the case of observations coming from meteorological stations, the researcher also needs to handle situations like missing values and possible erroneous observations. There are also situations where meteorological data do not exist or are not publicly available for the area under study.
In the case of our study, the meteorological observations recorded by the Portuguese Institute for Sea and Atmosphere (IPMA) relative to those days of interest for this study, are not easily available and can be obtained only under previous request.
Two sources of climate data which are publicly and easily available, and widely used in the study of losses produced by meteorological events [20, 6, 21], are the ERA-Interim and ERA-5 reanalysis[2] weather data from the ECMWF (European Center for Medium-Range Weather Forecasts). We compared the losses incurred by the property portfolio of the company due to hurricane Leslie, which occurred on 13 October 2018, with the ERA-5 reanalysis data for the same day. The daily maximum wind gust and the total daily precipitation are the quantities considered. The daily maximum wind gust is equal to the maximum value of the hourly 10 meter wind gust and has been computed for each of the 142 locations showed in Figure 1. The total daily precipitation instead is equal to the sum of the 1-hourly total precipitation amount and has been computed for each of the 142 locations showed in Figure 1.
It was observed that the mentioned meteorological variables were not compatible with the amount of losses registered by the company. Indeed the reanalysis values shown in Table 1 are too low to provoke the observed amount of losses. Also, [19] reports that a wind gust of 176 km/h was recorded in Figueira da Foz, and this is not in line with the values reported in Table 1.
Problems related to the calibration of a damage function can also be related to the loss data of the insurance company. Calibrating a damage function requires having enough severe losses due to extreme meteorological events, but the company may not have these data. This might happen if the company is a new player on the market or when the natural hazard under study occurs with low frequency.
In our case, the losses incurred by Fidelidade due to extreme meteorological events were available just from 2012. This is because in that year the company started to label those claims due to specific weather-related hazards, making them distinguishable from the claims due to different reasons.

[2]For more infromation about the Era 5 dataset consult `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5`

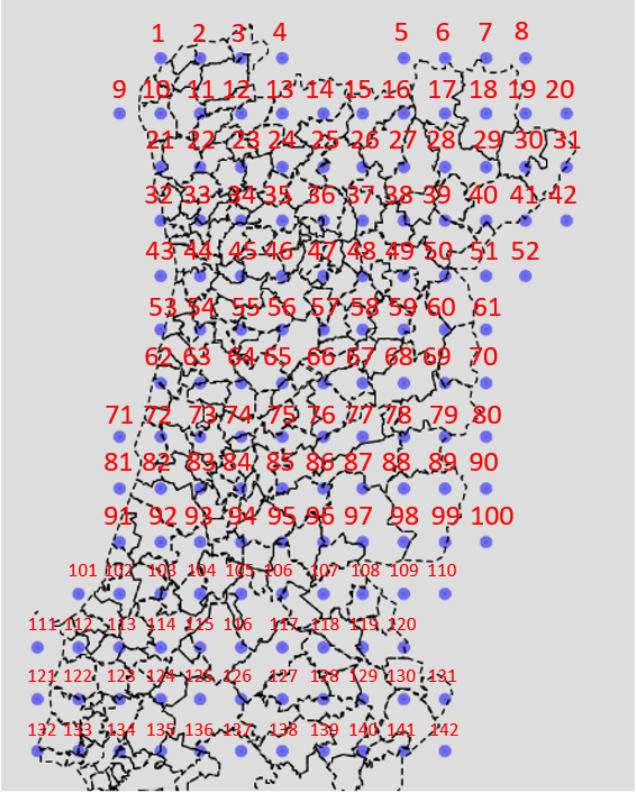

Figure 1: Grid points of the ERA-5 reanalysis dataset

Grid Point Number	10m Maximum Wind Gust (Km/h) Range of values	Total Daily Precipitation (mm) Range of values
1 - 8	49.34 - 60,58	0,71 - 28,40
9 - 20	41.47 - 69.70	0.04 - 25.41
21 - 31	40.06 - 67.23	0.03 - 24.48
32 - 42	41.61 - 69.29	0.02 - 24.62
43 - 52	41.86 - 75.69	0.02 - 27.55
53 - 61	44.26 - 81.56	0.03 - 26.02
62 - 70	39.75 - 84.70	0.28 - 22.99
71 - 80	42.81 - 88.00	0.50 - 24.96
81 - 90	49.15 - 84.82	0.06 - 28.53
91 - 100	53.21 - 80.85	0.00 - 28.89
101 - 110	59.70 - 80.15	0.00 - 28.89
111 - 120	61.52 - 89.99	0.01 - 27.00
121 - 131	60.04 - 85.33	0.00 - 20.49
132 - 142	59.31 - 81.35	0.00 - 10.05

Table 1: Wind and Rain magnitude measures considered and obtained using the ERA-5 reanalysis values for the 13/10/2018

Finally, even if both meteorological and loss data were available, the meteorological data relative to the magnitude of the events under analysis should be in line with

the amount of losses observed. As mentioned, this is not the case with hurricane Leslie, at least for the climate data available.
Because of all reasons mentioned above, we develop an approach, to estimate the costs provoked by an hurricane-force tropical cyclone, that does not require meteorological data related to the magnitude of the event. Given the loss data available, we will use the loss data related to hurricane Leslie.

2.1 Data on Tropical Cyclones

The occurrence of tropical cyclones affecting continental Portugal is documented since 1995 by The National Oceanic and Atmospheric Administration (NOAA). NOAA is the U.S federal agency specialized in the study of the tropical cyclones in the Atlantic and produces every year, since 1995, the Tropical Cyclone Report. It contains comprehensive information on each tropical cyclone, including synoptic history and the post-analysis best track (six-hourly positions and intensities). The tracking charts of those years, since 1995, where at least a tropical cyclone affected continental Portugal, are presented in Appendix A.
The information provided by the following reports of NOAA [18, 7, 19, 2, 3, 5], has been reported in Table 2. In that table the names of the events, the dates of occurrence and the maximum wind velocity registered over continental Portugal for each of the 6 cyclones, are reported.

Name	From - To	Max. wind speed obs. in cont. Portugal (km/h)
Jeanne	21/9 - 30/9/1998	56
Vince	8/10 - 11/10/2005	56
Rafael	12/10 - 17/10/2012	56
Joaquin	27/9 - 7/10/2015	65
Leslie	23/9 - 13/10/2018	176
Alpha	17/9 - 19/10/2020	102

Table 2: Tropical cyclones reaching continental Portugal since 1995

It is interesting to notice that all these 6 events happened between September and October. Indeed tropical cyclones develop over tropical waters and the most active period of the year where the majority of tropical storms and hurricanes develop worldwide is between August and October, as shown in Table 3.
From Table 2 we also notice that not all tropical cyclones reached continental Portugal with extreme wind velocities. In fact, tropical cyclones include depressions, storms and hurricanes. The depressions have maximum sustained surface winds of 61 km/h, the storms between 62 and 119 km/h and the hurricanes of more than 119 km/h. This means that only some types of tropical cyclones can represent a serious risk for the properties insured by the company.
As already mentioned, the company loss data regarding the events in Table 2 prior to 2012 are not available. Anyway, for both tropical cyclone Jeanne and Vince, as

Month	Total Tropical Storms	Total Hurricanes
January	3	2
February	1	0
March	1	1
April	2	0
May	22	4
June	92	33
July	120	55
August	389	245
September	584	404
October	341	205
November	91	59
December	17	6

Table 3: Total number of tropical cyclones registered worldwide by month (1851-2017) [17]

reported respectively by [18] and [7], there were no known casualties or damages reported. For the other 4 events, the total losses incurred by the company are reported in Table 4.

Event	Total cost	Total claims
Rafael	$< 0{,}053x$	$< 0{,}026y$
Joaquin	$< 0{,}104x$	$< 0{,}079y$
Leslie	$> x$	$> y$
Alpha	$\approx 0{,}049x$	$\approx 0{,}059y$

Table 4: Total costs and claims incurred by Fidelidade's property portfolio in Portugal and due to tropical cyclones since 2012, relative to the total cost (x) and total number of claims (y) due to hurricane Leslie

Among the 6 events that affected continental Portugal since 1995, hurricane Leslie was by far the strongest in intensity, as reported in Table 2, and caused the highest total costs to the company, as we can see from Table 4. In fact, just before making its landfall in Figueira da Foz, Leslie was labelled as hurricane. It maintained hurricane force winds also when it made its landfall over Figueira da Foz, as reported by [19].
Since the company is interested in estimating the average costs provoked by a hurricane-force tropical cyclone over the territory of continental Portugal, only the loss data relative to hurricane Leslie has been used. The results in this work are obtained analyzing the property portfolio of the company over the area of continental Portugal for the year of 2018.
We exclude from our analysis the records of the portfolio that don't belong to continental Portugal or for which we don't know the council they belong to. The records not considered represent the 5,22% of the total portfolio of the company in Portugal. The property loss dataset that will be used in the analysis consists of more than 1 million policies. For each policy, information about the council of belonging,

the cost incurred by the company if the policy incurred in a claim, the structural characteristics of the property and many more are available (consult Appendix B for the list of the variables considered).
The variable *cost*, in the case of records with claim incurred, represents the sum of all the expenses that the company had to face for that policy, from the payment of the sum covered by the policy to all the other expenses related to the administration of the claim.
To each property in the dataset, the coordinates of the main city of the council to which the property belongs, are assigned. This assumption, other than favouring the computing speed in the algorithm introduced in Chapter 5, is reasonable since we can assume that the company has most of its exposure, in each council, around the main city.

3 Using Actuarial Loss data to model the storm

The purpose of the models discussed in [21, 6, 20] is to find the damage functions that best fit the observed damage in the area under study. But, as already said, those models require both reliable past meteorological observations and enough insurance losses due to extreme-weather events, in order to calibrate the damage functions. The convenience of those approaches is that with damage functions the researcher can decide the magnitude of the meteorological event he wants to simulate. For example, using damage functions one can arbitrarily choose the speed of the surface wind over different locations of the area under study, and obtain estimates of the relative damage. Subsequently the insurance company, from the estimated damage, estimates the cost for the portfolio.

In our study, using the damage function approach was not possible. As already pointed out in the previous chapter, the climate data from the IPMA's meteorological stations can be obtained only under previous request, while the ERA-5 reanalysis data on wind and rain amounts for the day of 13 October 2018 were not compatible with the observed size of the losses.

The approach used in this work permits to obtain estimates of the expected cost due to an extreme meteorological event affecting a certain region of interest, but does not require the use of damage functions. Because of this, our method models directly the costs incurred by the company using the loss data. We model the costs provoked by a single hazard with a huge impact, like hurricane Leslie, and then apply the model calibrated on that specific event to other regions. In this way, we estimate the costs that the same event would have produced if it would have affected another area.

Since this method does not require any meteorological data, is possible to apply it when the climate data are not available or are not reliable, as discussed in Chapter 2.

To obtain the final model for the expected costs, first we compute the trajectory followed by the cyclone after its landfall. By trajectory we refer to the imaginary line around which the observed claims are distributed. Thus the trajectory is the line that passes through the councils affected by the event, and it is obtained using the least squares method.

It is important to notice that, in those cases where the observed losses do not distribute around an imaginary line, a clear path followed by the storm could be difficult to infer. For this reason, this approach is recommended for events that produce localized damages, where the term "localized" is relative to the size of the area under study. After having obtained the path followed by the event, the claim frequency and the average cost of the policies affected by the hazard are modelled.

With this approach we are able to estimate the expected costs incurred by the insurer in the scenarios where an event of the same type does its landfall in a different point and with a different angle. Also, by simulating many different scenarios and averaging their cost estimates, we are able to define the areas and the classes of policies which have the higher risk, in terms of expected costs, for the company.

3.1 Modeling the storm path

Knowing the distribution of the claims and costs provoked by the event at the level of a unitary region allows a better understanding of the behaviour of the event over the area under analysis. In the case of Portugal, we decide to analyze these quantities at the council level. Two measures are introduced, the cost ratio by council, here denoted by CR_i and defined in Equation (1), and the ratio of affected buildings [13], here denoted by RAB_i and defined in Equation (2). For both the cost ratio and the ratio of affected buildings the subscript i refers to the council.

$$CR_i = \frac{\text{Loss in i}}{\text{Total Amount Insured in i}} \tag{1}$$

$$RAB_i = \frac{\text{Claims in i}}{\text{Total Number of Properties Insured in i}} \tag{2}$$

The CR allows to clearly define the cost level incurred in the different councils. For example, 1 million € of total cost in a relatively small council (in terms of exposure) like Pinhel represents a higher level of destruction rather than 1 million € in the council of Lisbon. Moreover, using a relative measure, minimizes the inflation problem (that is the rise of price levels), making comparable the costs incurred by the company in different epochs. The RAB represents the relative frequency of claims in each council, and also in this case it has the advantage of being a measure comparable in different epochs, when, for instance, the exposure of the company in a certain council has changed. Figure 2 displays the observed CR by council, due to hurricane Leslie, over the territory of continental Portugal. In Figure 2 the CRs of the councils are merged in 8 groups, which have been defined based on the deciles of the CR distribution.

From Figure 2 we can infer the path of hurricane Leslie through the north east of Portugal, and its weakening along this path, reflected by the lower CRs observed as the hurricane moved away from the point of landfall. It is also clear from Figure 2 that the event affected just the Central and North regions, while it did not impact the councils in the Alentejo and the Algarve regions.

It is also clear from Figure 2 that the most affected councils have been the ones facing the coast. Table 5 reports the CR in the 5 most impacted councils. All of them are located in the center of the country and close to the cost.

Council	Region	CR
Montemor o Velho	Center	0,0047268
Soure	Center	0,0033196
Figueira da Foz	Center	0,0032859
Cantanhede	Center	0,0014495
Pombal	Center	0,0014305

Table 5: CR of the 5 councils in continental Portugal most affected by hurricane Leslie

To estimate the trajectory followed by the hurricane, we plotted in Figure 3 the coordinates of the main cities of those councils which reported damages due to

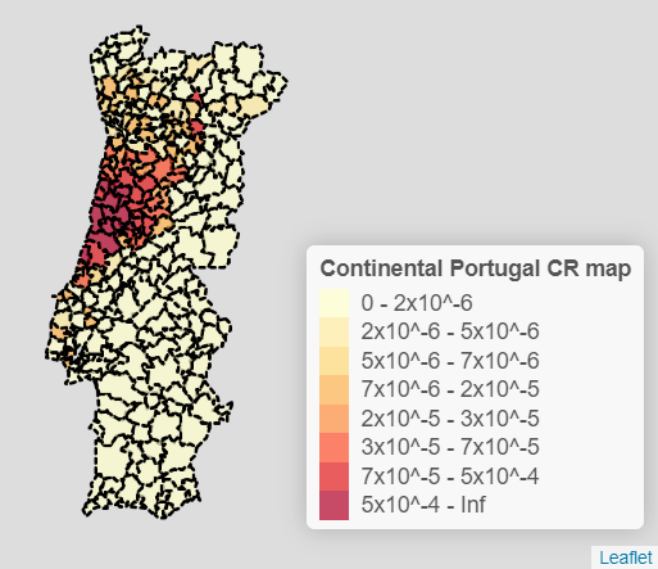

Figure 2: Distribution of the observed CRs due to hurricane Leslie over continental Portugal

hurricane Leslie. We consider the main city of each council as representative of the location of most of the claims in the council, since it's reasonable to assume that the company has most of its exposure around the main city of the council.

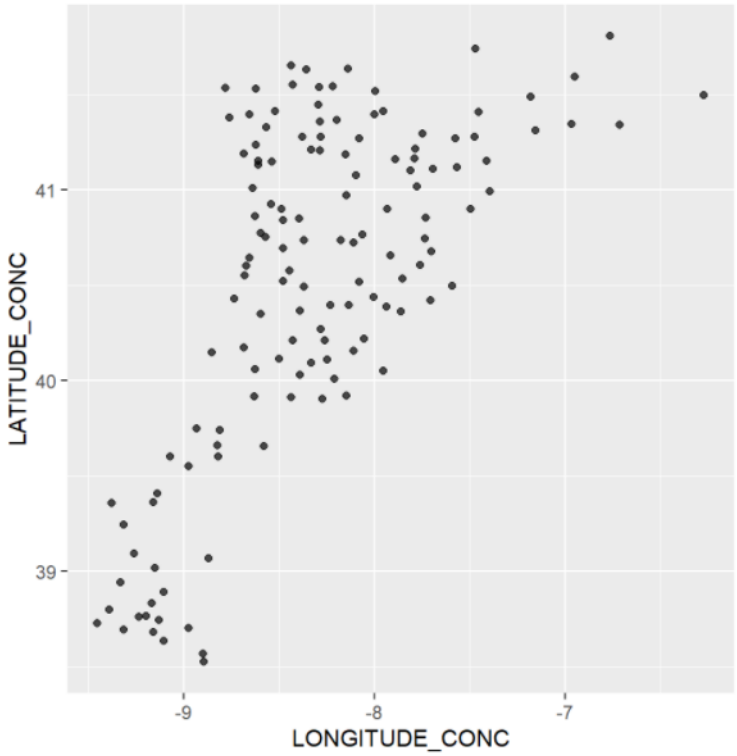

Figure 3: Coordinates of the main cities of those councils that reported at least one claim due to hurricane Leslie

The least square method constrained to the condition in Equation (3), has been used to obtain the red line showed in Figure 4, which represents the trajectory of the hurricane. The constraint guarantees that the trajectory passes through the council of Figueira da Foz. The choice of imposing the trajectory passing through Figueira da Foz is based on [19], which reports that hurricane Leslie made its landfall in this council. Figueira da Foz is also where the highest wind gust has been registered the 13 October 2018 [19]. Based on this, we assume that the main city of Figueira da Foz council, which is Figueira da Foz, is the landfall point of hurricane Leslie.

$$\begin{gathered}\text{Minimize} \quad \sum_{i=1}^{n}(LATITUDE_i - (\alpha + \beta LONGITUDE_i))^2 \\ \text{constrained by:} \\ \alpha = LATITUDE_{Fig.daFoz} - \beta LONGITUDE_{Fig.daFoz}\end{gathered} \tag{3}$$

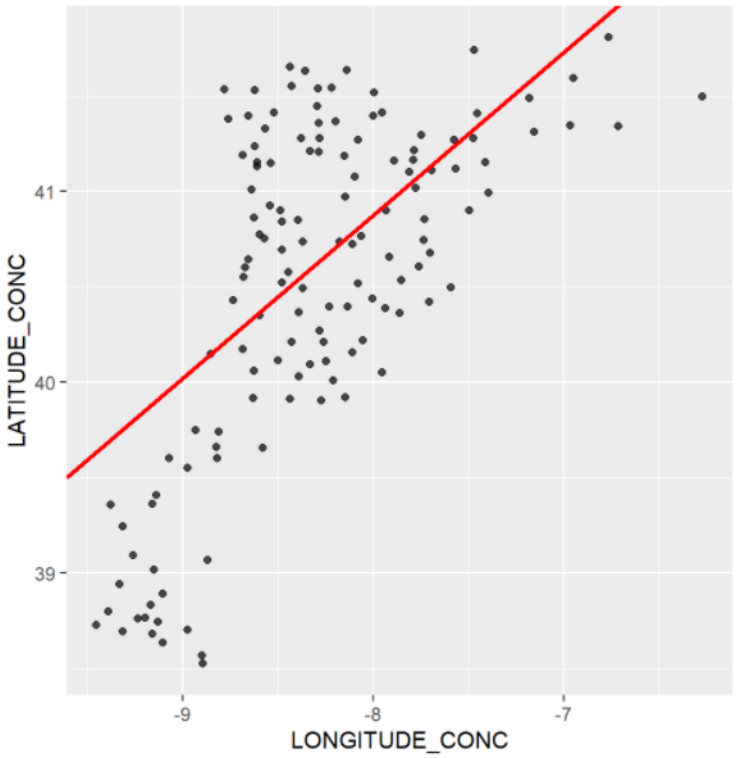

Figure 4: Coordinates of the main cities of those councils that reported at least one claim due to hurricane Leslie and inferred trajectory of the hurricane

The part of the trajectory line going from Figueira da Foz until the furthest council in the north-east of the country that reported claims, which is Bragança, has a length of approximately 260 km. This means that the hurricane caused losses to the company for at least 260 km travelling inland.
We do not have information available about the damages caused outside the borders of continental Portugal, but, for prudence reasons, we decide to assume that

the event could produce damages for 300 km before dissipating. Because of this, in the simulations performed in Chapters 4 and 5, the trajectory length is always assumed equal to 300 km.
After having obtained the trajectory, two variables, called D_1 and D_2, are computed. D_1 represents the distance of the object insured from the point where the storm makes its landfall (also called "entrance point"). Recall that in the case of hurricane Leslie, the entrance point is Figueira da Foz. D_2 represents the perpendicular distance of the object insured from the trajectory.
The computation of the variable D_1 is justified by the decrease of the observed CR by council along the trajectory line, starting from the landfall point, as showed in Figure 2. The computation of the variable D_2 is instead justified by the fact that the councils affected are closely distributed around the trajectory line, as showed in Figure 4.
At this point, we would like to have a descriptive model that permits us to understand how the variables D_1 and D_2 relate to the observed RAB and CR of the councils. A regression tree model is able to set simple rules to describe the observed RAB and CR for different values of D_1 and D_2. The tree-based methods consider a partition of the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one [9].
We use the CART method for tree-based regression and classification which we describe in the following [9]. Let us consider a regression problem with continuous response variable Y and inputs X_1 and X_2, each taking values in the unit interval. We first split the space into two regions, and model the response by the mean of Y in each region. We choose the variable and split-point to achieve the best fit. Then one, or both, of these regions are split into two more regions. This process is continued, until some stopping rule is applied.

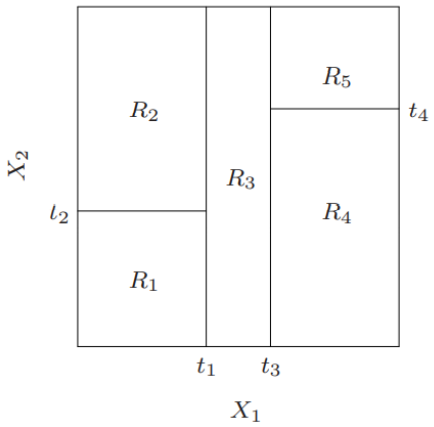

Figure 5: Example of a partition of a two-dimensional feature space performed by CART method

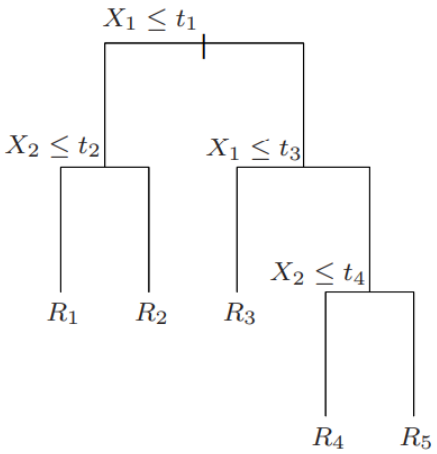

Figure 6: Tree corresponding to the partition in Figure 5

For example, in Figure 5, we first split the region at $X_1 = t_1$. Then the region $X_1 \leq$

t_1 is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$. Finally, the region $X_1 > t_3$ is split at $X_2 = t_4$. The result of this process is a partition of the space of (X_1,X_2) into the five regions R_1, R_2, . . . , R_5 shown in Figure 5. This partition of the space can also be represented through a tree graph, as in Figure 6.
The corresponding regression model predicts Y with a constant c_m in region R_m, that is:

$$Y = \hat{f}(X_1, X_2) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_{\mathrm{m}}\} \tag{4}$$

Let us see now how to grow a regression tree. The purpose is that the algorithm automatically provides the splitting variables and split points, and also what topology (shape) the tree should have. Consider p inputs and a response, for each of N observations: that is, (x_i,y_i) for i = 1,2,....,N, with x_i = (x_{i1},x_{i2},....,x_{ip}). Suppose first that we have a partition into M regions R_1, R_2, . . . , R_M, and we model the response as a constant c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I\{x \in R_{\mathrm{m}}\} \tag{5}$$

If we adopt as minimization criterion the sum of squares $\sum(y_i - f(x_i))^2$, it is easy to see that the best $\hat{c}_m$ is just the average of y_i in region R_m:

$$\hat{c}_m = ave(y_i | x_i \in R_{\mathrm{m}}) \tag{6}$$

To find the best binary partition in terms of minimizing the sum of squares, we start by considering all the data. Then, we consider a splitting variable, j, and split point, s, and define the pair of half-planes:

$$R_1(j, s) = \{X | X_{\mathrm{j}} \leq \mathrm{s}\} \text{ and } R_2(j, s) = \{X | X_{\mathrm{j}} > \mathrm{s}\} \tag{7}$$

Then, we seek the splitting variable j and split point s that solve

$$\min_{j,s} = [\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2] \tag{8}$$

For any choice j and s, the inner minimization is solved by

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j, s)) \text{ and } \hat{c}_2 = ave(y_i | x_i \in R_2(j, s)) \tag{9}$$

For each splitting variable j, the determination of the split point s can be done very quickly and, hence, by scanning through all of the inputs, the determination of the best pair (j, s) is feasible.
Having found the best split, we partition the data into the two resulting regions and repeat the same splitting process on each of the two regions, and so on.
The question that arises is: how large should we grow the tree?. Clearly a very large tree might overfit the data, while a small tree might not capture the important structure. The preferred strategy to obtain the optimal tree size is to grow a large tree T_0, stopping the splitting process only when some minimum node size (say 5)

is reached. A node is comprised by the elements of Y which belong to the subset of the feature space defined by the corresponding splits of the input variables. Then this large tree is pruned using cost-complexity pruning, which we describe in the following.
We define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0, that is by collapsing any number of its internal (non-terminal) nodes. We index terminal nodes by m, with node m representing region R_{m}. Let $|T|$ denote the number of terminal nodes in T. Letting

$$N_{\mathrm{m}} = \#\{x_i \in R_{\mathrm{m}}\},\ \hat{c}_{\mathrm{m}} = \frac{1}{N_{\mathrm{m}}} \sum_{x_i \in R_m} y_i,\ Q_m(T) = \frac{1}{N_{\mathrm{m}}} \sum_{x_i \in R_m} (y_i - \hat{c}_{\mathrm{m}})^2 \tag{10}$$

we define the cost complexity criterion

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_{\mathrm{m}} Q_m(T) + \alpha|T| \tag{11}$$

The idea is to find, for each α, the subtree $T_\alpha \subseteq T_0$ to minimize $C_\alpha(T)$. If the target is a classification outcome taking values $1, 2, \ldots, K$, the only changes needed in the tree algorithm pertain to the criteria for splitting nodes and pruning the tree. For the regression we consider the squared-error node impurity measure $Q_m(T)$ defined in Equation (10). However this is not suitable for classification. For this reason, let us consider, in a node m representing a region R_m with N_m observations, the next quantity

$$\hat{p}_{mk} = \frac{1}{N_{\mathrm{m}}} \sum_{x_i \in R_m} I(y_i = k), \tag{12}$$

which represents the proportion of class k observations in node m. We classify the observations in node m to class $k(m) = \arg\max_k \hat{p}_{mk}$, which is the class that has most observations in node m.
A measure of $Q_m(T)$ node impurity used in the CART algorithm is the Gini index:

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} \tag{13}$$

Figures 7 and 8 report the two regression trees obtained for, respectively, the RAB and the CR of the affected councils (the ones which reported at least 1 claim) against the variables D_1 and D_2. To obtain both trees, the complexity parameter α has been set to -1, to ensure that the trees are fully grown, and 20 has been set as the smallest number of observations that are allowed in a terminal node.
In this case, the CART algorithm has divided the feature space in 5 regions, and the values predicted for each region, together with the percentage of councils belonging to that region, are reported in the blue squares of Figures 7 and 8 which represent the final nodes of the trees.
As already said, the purpose of these regression trees is not to predict, but to describe the effect of the variables D_1 and D_2 on the RAB and CR observed.

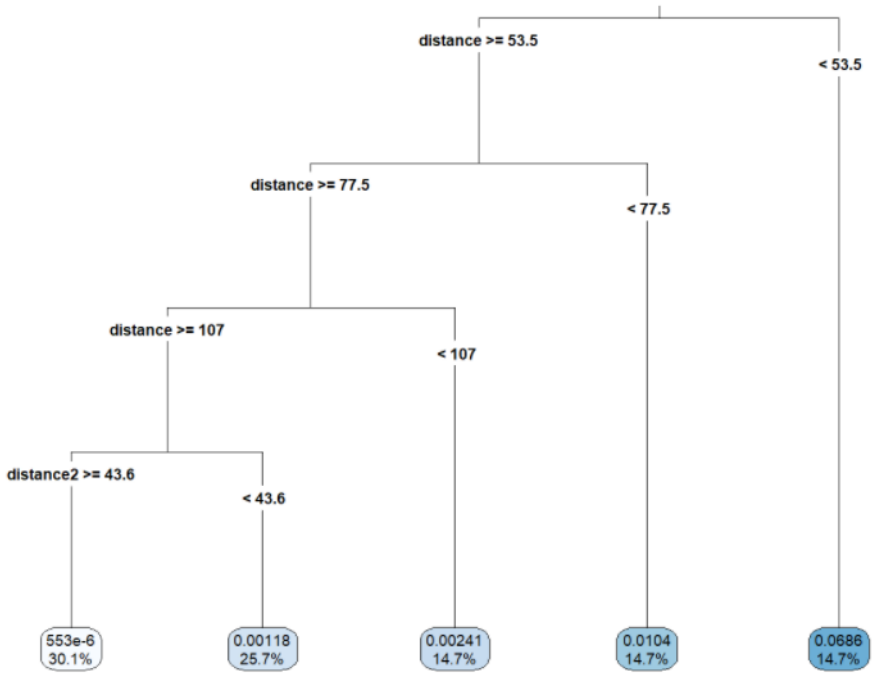

Figure 7: Tree corresponding to the partition of the feature space defined by the variables D_1 and D_2 performed by the CART method applied to the observed RAB of the councils hit by hurricane Leslie

The terminal nodes of the two regression trees in Figures 7 and 8, are obtained by partitioning the feature space defined by the two variables D_1 and D_2 as described in the CART algorithm. We use the splitting rules employed by that algorithm, that are the conditions by which the intermediate nodes are split, to build two factor variables called *intensity* and *intensity2* which are described in Appendix B. The splitting rules are the numerical conditions displayed in Figures 7 and 8. In this way *intensity* and *intensity2* capture the combined effect of D_1 and D_2 on the different magnitudes of the CR and the RAB observed. These variables will be used later in this chapter and in Chapter 4 as predictors in the models and also as a tool to analyze the results of the simulated scenarios.

3.2 Modelling the claims and costs of the affected councils

As Figure 2 shows, hurricane Leslie affected mainly the central and north part of the country. In order to obtain useful information about the characteristics of those policies which reported a claim when affected by the hurricane, we consider only those councils which have actually been affected.
The first model we introduce, has thus the purpose to predict, based on the trajectory of the hurricane, which councils will be affected and which will not. For this task, variables D_1 and D_2 can be used as input variables to predict a binary outcome, 1 or 0, if, respectively, the council is affected or not.

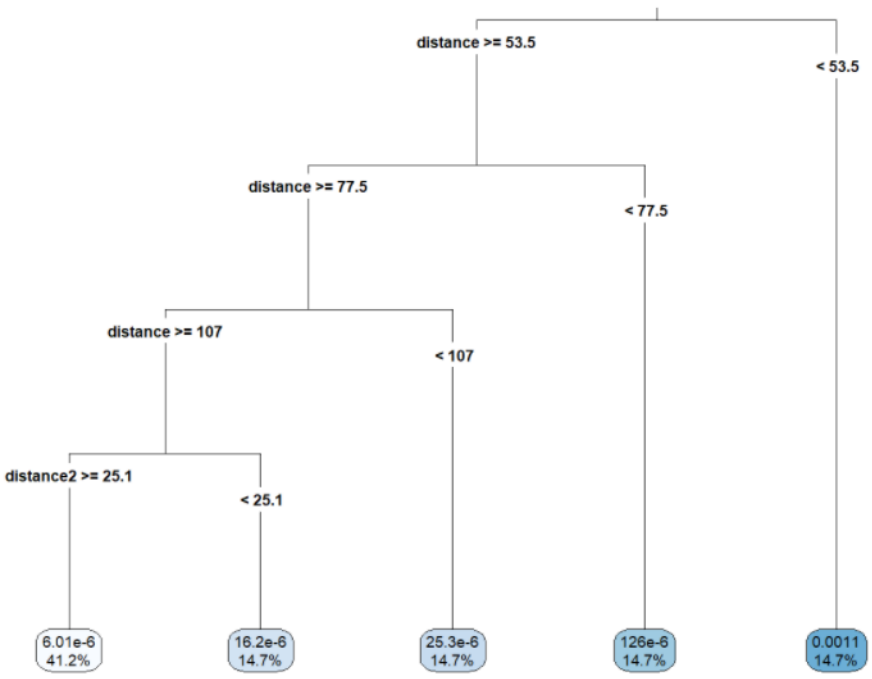

Figure 8: Tree corresponding to the partition of the feature space defined by the variables D_1 and D_2 performed by the CART method applied to the observed CR of the councils hit by hurricane Leslie

Since the purpose of the model is to predict rather than explain, a machine learning approach is preferred over a regression based approach. In this work, a Random Forest model for classification is used. After having introduced the concept of regression and classification tree algorithm, let us introduce the concept of bagging. Suppose we fit a model to our training data, which is a sample of the data that we use to fit our model. The training data are $Z = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ and we use them to obtain the prediction $\hat{f}(x)$ at input x. The bootstrap aggregation, also called bagging, averages this prediction over a collection of bootstrap samples. A bootstrap sample is a sample of the same size of the training data (also called training set) obtained by sampling with replacement from the training data. By the bootstrap aggregation we can thereby reduce the variance of the prediction, as shown in [9]. For each bootstrap sample Z^{*b}, b = 1, 2, . . . , B, we fit our model obtaining the prediction $\hat{f}^{*b}(x)$. The bagging estimate is defined by:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{14}$$

Random forests [4] is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. Algorithm 1 shows how the Random Forest work.

Algorithm 1 Random Forest for Regression or Classification.

1. For $b = 1$ to B:
(a) Draw a bootstrap sample Z^* of size N from the training data.
(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size N_{min} is reached.
i. Select m variables at random from the p variables.
ii. Pick the best variable/split-point among the m.
iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x\}_1^B$

3.2.1 Modelling the affected councils

The dataset we use for estimating which councils will be affected, is comprised by the 278 councils of continental Portugal. For each council three features are reported: the distance from the landfall point D_1, the distance from the trajectory D_2, and the number 1 or 0 respectively if the council has been affected or not by the hurricane. It is important to highlight that among the 278 councils in continental Portugal, the company had losses caused by Leslie in 136 of them. This means that our dataset is not imbalanced and so we don't have to be concerned about this affecting the estimates of the model. Since the size of this dataset is considered to be relatively small, to evaluate the predictive ability of the model, a 10-fold cross validation method is employed.
In a k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. In this experiment, to produce a final estimate of the performance, the mean and variance of the sensitivity [3] (15) and specificity [4] (16) indexes have been computed. The results are reported in Table 6.

$$\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}} \tag{15}$$

[3] The sensitivity is also denoted true positive rate
[4] The specificity is also denoted true negative rate

$$\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}} \tag{16}$$

	Random Forest
E(Sensitivity)	0,8593
Var(Sensitivity)	0,0049
E(Specificity)	0,8121
Var(Specificity)	0,0139

Table 6: Mean and variance of the sensitivity and specificity indexes computed for the Random Forest model on 10 different folds

The performance of the model is good, we obtain an average of 86% of true positive rate and an average of 81% of true negative rate. Later in this work, we will refer to the the Random Forest model calibrated on the dataset comprised by the 278 records as "RF Model".

3.2.2 Modelling the claim frequency

We want to model now the average claim frequency for those policies that are affected by hurricane Leslie. For calibrating purposes, we take into consideration only the policies which belong to the councils affected. Calibrating the model on the whole portfolio would lead to biased estimates. For instance, let us consider that the policies regarding properties in the area affected by hurricane Leslie, are mainly single-family houses, while in the non-affected parts of the country the majority of the policies correspond to apartments. Then, it is probable that our model would assign to the single-family houses a higher probability of having a claim than to the apartments, but only because the latter category has not being subjected to the magnitude of the meteorological event. For this reason it is important to calibrate the model considering only the area actually affected by the hazard.
Since the company is interested in having a functional relation between the predicted probabilities of claims and the characteristics of the policies, a regression based approach is employed. The logistic regression [10] is used to model the probability of the binary event, i.e 0 = claim does not happen or 1 = claim happens. Two logistic regression models, defined in Equations (17) and (18), have been considered as possible candidates. The most significant explanatory variables, from a statistical view point, that are included in the models are characteristics related to the insured property itself, like for instance the type of property and the year of construction. Characteristics related to the location of the property, like the altitude and the territorial type [5] have also been included. For a complete description of all the variables employed refer to Appendix B.
The main difference between the two regression models (17) and (18), is how the

[5] for a detailed description of the variables *Forest Area*, *Bush Area* and *Urban Area* please refer to Appendix B

effect of D_1 and D_2 on the independent variable is expressed. In regression (17), the factor variable *intensity* is used. While in regression (18), the variables D_1 and D_2 are included in the model as numeric variables and an interaction term is also considered.

$$\begin{aligned} log(\frac{p}{1-p}) = \beta_0 + \beta_1 Type\ of\ Property + \beta_2 Year\ of\ Construction \\ + \beta_3 Framing\ of\ the\ Housing + \beta_4 Type\ of\ Housing + \beta_5 Altitude \\ + \beta_6 Type\ of\ Floor + \beta_7 Forest\ Area + \beta_8 Bush\ Area + \beta_9 intensity \end{aligned} \quad (17)$$

$$\begin{aligned} log(\frac{p}{1-p}) = \beta_0 + \beta_1 Type\ of\ Property + \beta_2 Year\ of\ Construction \\ + \beta_3 Framing\ of\ the\ Housing + \beta_4 Type\ of\ Housing + \beta_5 Altitude \\ + \beta_6 Type\ of\ Floor + \beta_7 Forest\ Area + \beta_8 Bush\ Area + \beta_9 D_1 + \beta_{10} D_2 + \beta_{11} D_1 * D_2 \end{aligned} \quad (18)$$

To validate the adjustment of the two models, the 10-fold cross validation method is performed for each model using the dataset composed by the policies that belong to the councils affected by hurricane Leslie. The dataset is composed by more than 800.000 policies and, in order to compare the performance of the two regressions (17) and (18), we first decided to transform the estimated probabilities into outcomes 0 or 1. In this case 2 methods have been compared. The first, that we will call Method 1, consists in setting a cutoff value in such way that the number of estimated claims in the training set is as close as possible to the number of observed claims. The second method, called Method 2, consists in sampling outcomes 0 or 1 for each record, from a Bernoulli distribution using the probabilities estimated through the logistic regression. The analysis of the two models and the two cut-off methods, is performed on the test set by assessing two metrics: i) the weighted correlation (W.Corr) between the predicted and the observed number of claims in each risk class, weighted by the number of unit risks belonging to the class, and ii) the root mean squared error (RMSE) between predicted and observed values in each risk class. The risk classes are composed by the intersection of the levels of the explanatory variables which models (17) and (18) have in common. The variables are: *Type of Property*, *Year of Construction*, *Framing of the Housing*, *Type of the Housing*, *Altitude*, *Type of Floor*, *Forest Area* and *Bush Area*. Table 7 gives an example of how a risk class is made.

Type of Prop.	Year of Constr.	.....	Unit Risks	Observe	Predict
Content	]1992;2002]	.....	4783	120	101

Table 7: Example of a risk class

The models are evaluated based on their ability to predict the number of claims for a certain risk class, and not on their ability to predict if an individual policy will

have a claim or not. In an insurance context, we are interested in predicting the average claim frequency for a class rather than for an individual policy.
The results obtained from the 10-fold cross validations applied to the combinations of the two different models and methods are reported in Table 8.

	Reg.(17) M. 1	Reg.(17) M. 2	Reg.(18) M. 1	Reg.(18) M. 2
E(W.Corr)	0,6478	0,8483	0,7162	0,8703
SD(W.Corr)	0,0383	0,0260	0,0228	0,0226
E(RMSE)	4,8790	1,1211	3,0240	1,0707
SD(RMSE)	0,1873	0,0815	0,1396	0,0613

Table 8: Mean and standard deviation of the weighted correlation and RMSE indexes of the regressions in Equations (17) and (18) and cut-off methods (M.) 1 and 2 on 10 different folds

In addition, in order to offer a visual comparison of the two methods and models to the reader, we randomly divided the dataset into training set and test set, in a proportion of 80% and 20%. Figures 9 to 12, display the predicted versus the observed values in the test set for the risk classes obtained using each of the two cut-off methods and models described before.

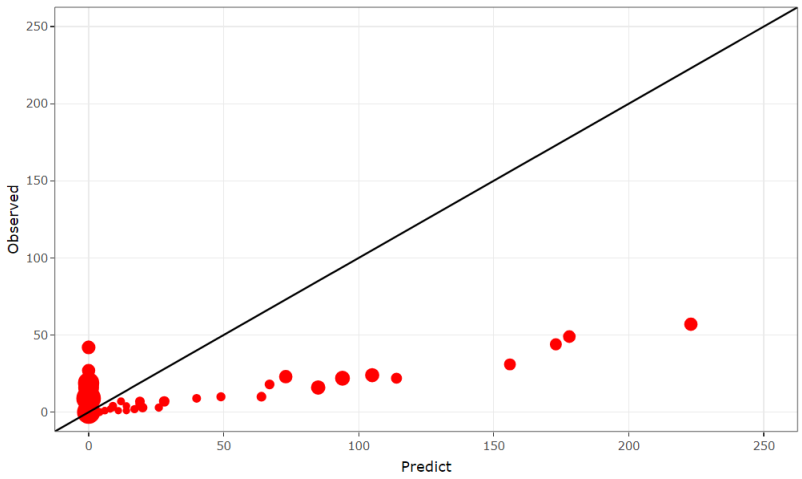

Figure 9: Number of claims predicted by regression in Equation (17) and cut-off Method 1 for different risk classes versus the observed values

From a visual inspection it is clear that the cut-off Method 2 leads to better predictions than the cut-off Method 1. The coefficient of correlation and the RMSE among the risk classes are reported in Table 8, for the 2 regression models and the 2 cut-off methods.
Results in Table 8 confirm that for both regression models in Equations (17) and (18), Method 2 performs better than Method 1. However the model in Equation (18) with cut-off Method 2 has a slightly higher weighted coefficient of correlation and slightly lower RMSE than the model in (17). The levels of the variable *intensity* indicate the extent of the areas over which the hurricane has a certain probability of causing a claim. This information is useful for the analysis that is performed later

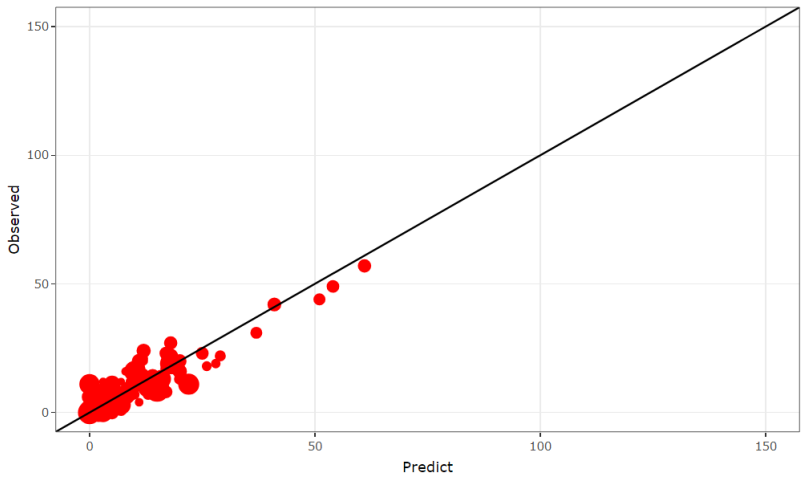

Figure 10: Number of claims predicted by regression in Equation (17) and cut-off Method 2 for different risk classes versus the observed values

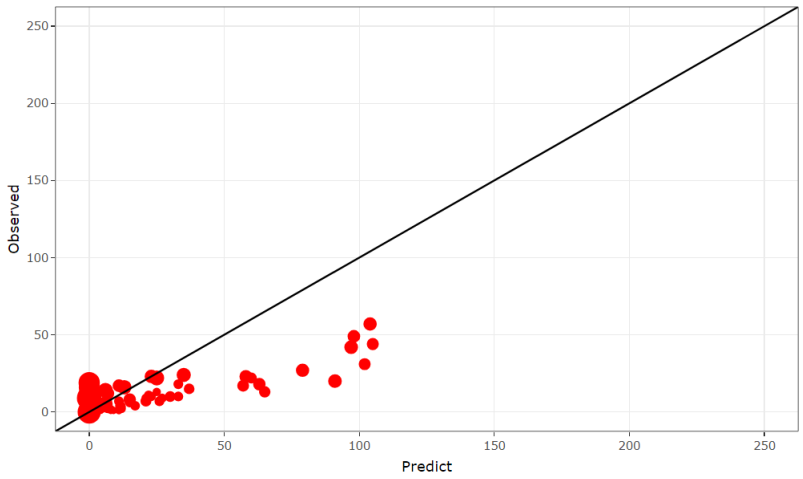

Figure 11: Number of claims predicted by regression in Equation (18) and cut-off Method 1 for different risk classes versus the observed values

in this work. Also, the effect of the variables D_1 and D_2, in regression model (17), is captured in just 1 factor variable, instead of 3 numeric variables. In regression model (18), the interaction between the variables D_1 and D_2 is represented with the product of both variables. Although the interaction term is statistically significant, its interpretation is not immediate. For all the previous reasons, in the following of this work, we consider regression model (17) and the cut-off Method 2 to predict the claim frequency.

Now, to build the final model to be employed in the simulations of Chapters 4 and 5, we calibrate regression (17) on the whole dataset, since the more data we use, the more likely it is to generalise well. Table 9 refers to regression (17) calibrated on the whole dataset of policies in those councils affected by hurricane Leslie.

The coefficients estimated for regression (17) are all statistically significant at a

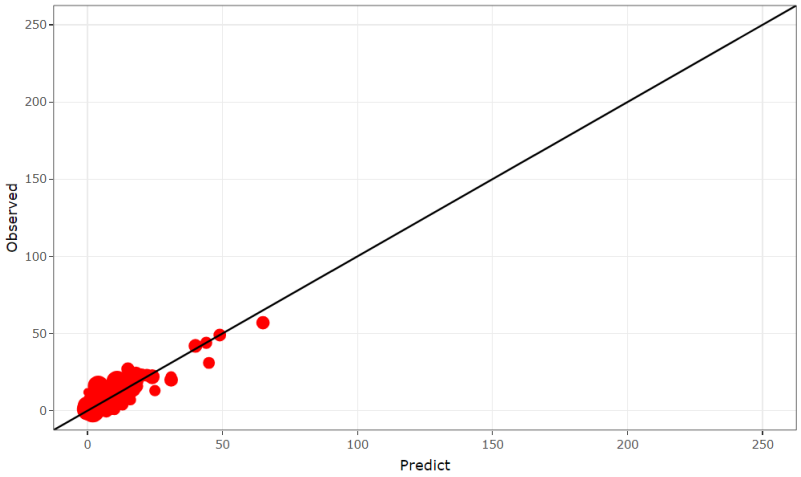

Figure 12: Number of claims predicted by regression in Equation (18) and cut-off Method 2 for different risk classes versus the observed values

Coefficient	Estimate	Std. Error	z value	p value	Signif. code
(Intercept)	-2.03009	0.04181	-48.553	< 2e-16	***
T.o.P Content	-2.35047	0.04111	-57.180	< 2e-16	***
Y.o.C LEV2	0.09722	0.03597	2.703	0.006875	**
Y.o.C LEV3	0.13160	0.03513	3.746	0.000180	***
Y.o.C LEV4	0.26766	0.03445	7.771	7.82e-15	***
Framing semi-det	-0.43324	0.03462	-12.514	< 2e-16	***
Framing other	-0.10034	0.04487	-2.236	0.025326	*
Type single-fam	0.67367	0.03321	20.285	< 2e-16	***
Type other	0.34408	0.05098	6.749	1.49e-11	***
Altitude LEV2	-0.34458	0.03101	-11.112	< 2e-16	***
Altitude LEV3	-0.61717	0.05267	-11.719	< 2e-16	***
intensity LEV2	-2.19481	0.04816	-45.577	< 2e-16	***
intensity LEV3	-3.37780	0.07508	-44.992	< 2e-16	***
intensity LEV4	-4.33357	0.11695	-37.055	< 2e-16	***
intensity LEV5	-5.29454	0.06855	-77.241	< 2e-16	***
T.o.F LEV2	0.25213	0.07144	3.529	0.000416	***
T.o.F ND	0.18343	0.03350	5.475	4.38e-08	***
Bush Area LEV2	-0.28077	0.02480	-11.324	< 2e-16	***
Forest Area LEV2	-0.55605	0.02629	-21.152	< 2e-16	***

Table 9: Summary of the coefficients estimated for regression (17)

significance level of 0.05 [6]. We can analyze the marginal effect of any explanatory variable of regression (17) on the odds of having a claim, by simply using the formula defined in Equation (19). If we consider a logistic regression model with independent variable Y and a vector of explanatory variables X, then the marginal effect of

[6] for the significance codes, refer to Table 22 in Appendix B

the categorical variable X_k on the $odds(Y = 1|X, X_k = 1)$ against the $odds(Y = 1|X, X_k = 0)$ is:

$$\frac{odds(Y = 1|X, X_k = 1)}{odds(Y = 1|X, X_k = 0)} = e^{\beta_k} \tag{19}$$

By applying Equation (19) to the results in Table 9, we can see that the odds of having a claim are 96% higher for the single-family houses than for the apartments. Another interesting result is that the odds of having a claim are almost 90% lower for the policies that are located in the area defined by the second level of the variable *intensity*, with respect to those located at level one. This means that the properties located at a distance inferior to 54 km from the landfall point of the hurricane have 90% higher odds of having a claim compared to those at a distance comprised between 54 and 78 km. Also the coefficients estimated for the levels of the variable *intensity*, tend to decrease from the first level to the fifth meaning that the propensity of the hurricane to cause claims decreases by moving away from the landfall point and the center of the trajectory.

3.2.3 Modelling the claim severity

Next, we aim at modelling the average cost for a claim. The number of policies in continental Portugal that incurred in a claim due to hurricane Leslie, and thus represented a cost for the company, was approximately the 1 % of the portfolio of the company in the councils affected by the event. If we consider the cost distribution relative to the whole portfolio over the affected councils, the distribution is highly right skewed, with most of its mass concentrated in 0. If instead we consider the cost distribution relative to those policies which incurred in a claim, the distribution is also highly right skewed, but there is no probability mass concentrated in 0. In the latter case, we were able to log-transform the cost distribution and observed that the log-cost distribution was well-approximated by a normal distribution. This permits us to utilize a multiple linear regression model [10] to predict the average log-cost. Two multiple linear regression models, defined in Equations (20) and (21), have been considered. The explanatory variables employed in both models are characteristics of the property insured, like the type of property, the amount of capital insured and the type of housing, as well as characteristics of the area where the property is located. For the complete description of the variables refer to Appendix B.

The main difference between the two multiple linear regression models considered is how the effect of the variables D_1 and D_2 is included. Similarly to what was done in regression models (17) and (18) for the claim frequency, in the regression model (20) the factor variable *intensity2* is used, while in the regression model (21), the variables D_1 and D_2 are included in the model as numeric variables, and an interaction term is also added.

$$log(\text{Cost} \mid \text{Cost} > 0) = \beta_0 + \beta_1 \, Type \ of \ Property + \beta_2 \, Capital \ Insured + \beta_3 \, Type \ of \ Housing + \beta_4 \, Urban \ Area + \beta_5 \, intensity2 \quad (20)$$

$$log(\text{Cost} \mid \text{Cost} > 0) = \beta_0 + \beta_1 \, Type \ of \ Property + \beta_2 \, Capital \ Insured + \beta_3 \, Type \ of \ Housing + \beta_4 \, Urban \ Area + \beta_5 D_1 + \beta_6 D_2 + \beta_7 D_1 * D_2 \quad (21)$$

Since the dataset comprised by the policies which had a claim only has 8.487 records, we evaluate the prediction ability of the two models (20) and (21) on a 10-fold cross validation. The performance of the two models is evaluated assessing two metrics: the weighted correlation between the total cost predicted and the one observed in each risk class, weighted by the number of unit risks belonging to the class, and the RMSE between predicted and observed values in each risk class. The risk classes are formed by the intersection of the levels of the explanatory variables which models (20) and (21) have in common. The variables are: *Type of Property*, *Capital Insured*, *Type of the Housing* and *Urban Area*. Also in this case we evaluate the prediction capacity of the models for a risk class, instead of doing it for the expected cost of an individual policy. Modelling the logarithm of the cost poses the problem of transforming the estimated expected log-cost back to expected cost. To do so, the Duan's smearing factor (D_{Smear}) [10] is estimated, using the residuals of the regression, and employed, as shown in Equation (22), to estimate the expected cost.

$$E(y|x) = e^{E(log(y|x))} D_{Smear} \quad (22)$$

The mean and standard deviation of the weighted coefficient of correlation, and the RMSE for the total cost predicted versus the one observed for each risk class, are reported in Table 10 for the 2 considered models.

	Regression (20)	Regression (21)
E(Corr)	0,9841	0,9845
SD(Corr)	0,0173	0,0173
E(RMSE)	10370,34	10231,79
SD(RMSE)	6597,22	6584,81

Table 10: Mean and standard deviation of the weighted coefficient of correlation and RMSE indexes computed for the regressions in Equations (20) and (21) on 10 different folds

The weighted coefficient of correlation of regression (21) is 0,04% higher than that of regression (20), while the average RMSE is 1,34% lower than regression (20). Regression (21) performs slightly better but regression (20) has the advantage of being easier to interpret than regression (21). The levels of the variable *intensity2*

indicate the extent of the areas over which the hurricane has the power of causing higher costs, and this information is useful for the analysis that is made later on. Also, the effect of the variables D_1 and D_2, in regression (20), is captured in only 1 factor variable instead of 3 numeric variables, like in regression (21). In regression (21), the interaction between the variables D_1 and D_2 is represented by the product of both variables. Although the interaction term is statistically significant, its interpretation is less intuitive. Hence, regression (20) is the model we use to predict the expected cost incurred by a policy, given that the policy has a claim.
Now, to build the final model, we calibrate regression (20) on the dataset comprised by those policies which reported a claim and represented a cost for the company. The estimated coefficients are reported in Table 11.

Coefficients	Estimate	Std. Error	z value	p value	signif. code
(Intercept)	5.94432	0.03175	187.214	< 2e-16	***
T.o.P Content	-0.43345	0.04606	-9.411	< 2e-16	***
Cap. Ins LEV2	0.12030	0.03327	3.616	0.000301	***
Cap. Ins LEV3	0.30771	0.03441	8.942	< 2e-16	***
Cap. Ins LEV4	0.53860	0.03487	15.447	< 2e-16	***
Type single-fam	0.67329	0.02694	24.993	< 2e-16	***
Type other	0.48910	0.03829	12.774	< 2e-16	***
intensity LEV2	-0.25639	0.04639	-5.527	3.35e-08	***
intensity LEV3	-0.17290	0.07146	-2.419	0.015564	*
intensity LEV4	-0.59452	0.13755	-4.322	1.56e-05	***
intensity LEV5	-0.40922	0.06332	-6.463	1.08e-10	***
Urban Area LEV2	-0.20916	0.03763	-5.559	2.79e-08	***
Urban Area LEV3	-0.35671	0.02711	-13.159	< 2e-16	***

Table 11: Summary of the coefficients estimated for regression (20)

We can compute the percentual change in the cost due to the marginal effect of any explanatory variable by using the expression in (23). If we consider a multiple linear regression model with independent variable $log(Y)$ and a vector of explanatory variables X, then the percentage change on Y due to the factorial variable x_k is:

$$100(e^{\beta_k} - 1) \tag{23}$$

By applying the expression in (23) to the results in Table 11, it is interesting to see that the average cost for a single-family home is 96% higher than for an apartment. Another interesting result is that the average cost decreases by 23% between the policies located in the area defined by the first level of the variable *intensity2* and the ones located in the area defined by the second level of the same variable. This means that the properties located at a distance inferior to 54 km from the landfall point of the hurricane have, on average, 23% higher costs compared to the ones located at a distance comprised between 54 and 78 km. From Table 11 we can also see how the average cost tends to decrease as the concentration in the council of urban areas increases.

3.2.4 Modelling the claims and their costs

Finally we aim to evaluate the prediction ability of regressions (17) and (20) combined. To evaluate the models, the dataset relative to the portfolio of the company in the councils affected by hurricane Leslie is divided in training and test sets. The evaluation of the performance of the two models on the test set, for the number of claims and their severity, is again performed by assessing two metrics: the weighted correlation between the predicted and observed total costs, in each risk class, weighted by the number of unit risks belonging to the class, and the RMSE between predicted and observed values in the risk classes. The risk classes in the test set are composed by the intersection of the levels of the explanatory variables of the regression models (17) and (20) which are: *Type of Property*, *Year of Construction*, *Capital Insured*, *Framing of the Housing*, *Type of the Housing*, *Altitude*, *Type of Floor*, *Forest Area*, *Bush Area*, *Urban Area*, *intensity* and *intensity2*. The dataset is divided in training set and test set in a proportion of 60% and 40%, in order avoid having too few policies belonging to each risk class in the test set. We calibrate regression (17) on the training set and then regression (20) on the subset of the training set composed by all the policies with a claim and that represented a positive cost for the company. After having estimated the two models, we apply regression (17) and the cut-off Method 2 to the test set, to predict which policies will have a claim and, for each of these policies, we apply regression (20) to predict the average cost. Figure 13 displays the predicted versus observed cost for each risk class, while Table 12 reports the weighted coefficient of correlation between predicted and observed costs and also the RMSE.

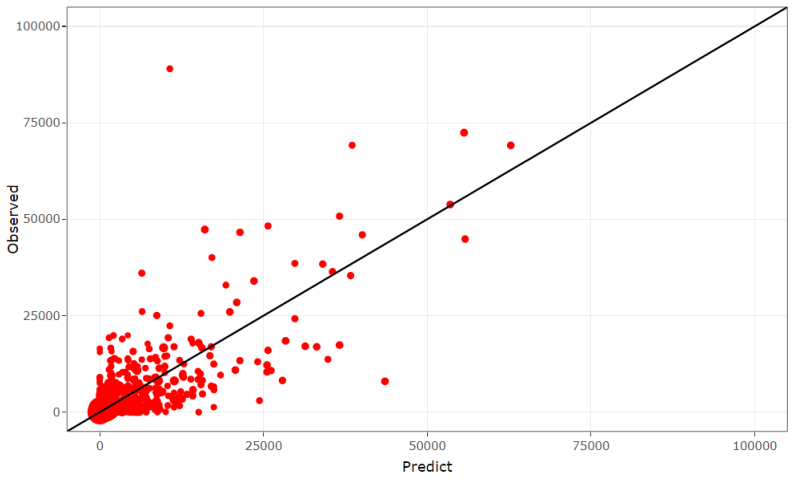

Figure 13: Costs predicted by the regressions in Equation (17) and (20) for different risk classes versus the observed values

The performance of the 2 regressions combined is good, with a high correlation between the predicted and observed values for each risk class.

Model	Weighted Correlation	RMSE
Regressions (17) and (20)	0,76	3894,91

Table 12: Weighted coefficient of correlation and RMSE indexes computed for the total costs predicted by the regressions in Equation (17) and (20) for different risk classes

4 Case scenarios in continental Portugal

In this chapter we estimate the expected cost for the company in the simulated scenarios where, a hurricane-force storm like Leslie, affects continental Portugal but in a different part of the country and with a different trajectory. These simulated scenarios are used afterwards to build a risk map for events of this type in continental Portugal.

First, we have to trace the trajectory of the storm. Since we assume that Leslie caused damages inland along 300 km, the length of the trajectory over land is fixed at 300 km. The user can choose the angle of the trajectory and the landfall point over continental Portugal. The initial point of the trajectory of the hurricanes will always be set in those councils that have a direct exposition to the ocean. This assumption is due to the fact that, as shown in Appendix A, tropical cyclones turn northeastward when moving from the tropics to the midlatitudes. This implies that all the "coast councils" that go from Vila Real de Santo Antonio in the district of Faro, along the south and west coast of Portugal, until the council of Caminha, in the district of Viana do Castelo, are possible points of entrance for the simulated hurricane. It is also important to remark that, for the sake of computational simplicity, we assume that the exact point where the hurricane landfalls has the coordinates of the main city of the council itself. Also, for the same reason, the coordinates of each policy are assumed to be the coordinates of the main city of the council where the policy belongs to. This implies that each policy belonging to a certain council has the same values of D_1 and D_2. Once the trajectory is drawn, the variables D_1 and D_2, and subsequently *intensity* and *intensity2*, are obtained for each policy. The RF Model is firstly employed to simulate which councils are affected by the hurricane, based on the values of the variables D_1 and D_2. After the affected councils have been selected, regression (17) is applied to the policies which belong to the affected councils, to estimate the probability of incurring in a claim during the extreme event. Then the cut-off Method 2 is applied to the probabilities estimated by regression (17) to simulate the occurrence, or not, of a claim. After this, regression (20) is applied to the subset comprised by the records which are predicted having a claim, to estimate the average log-cost incurred by the company. This quantity is then transformed in average cost using the Duan's smearing factor, as described in Equation (22), which has been estimated using the residuals of regression (20) calibrated on the dataset comprised by those policies that had a claim due to Leslie.

4.1 Lisbon and Porto Metropolitan Areas and Algarve Region

In the first simulation, Cascais is choosen as entrance point. The choice is due to the fact that it is one of the councils that belongs to the Lisbon Metropolitan Area, which is a region of great exposure for the company and it is directly exposed to the ocean, and so to the cyclones.
It is assumed a linear trajectory with 45 degrees. The results obtained for this simulation, which are reported in Table 13 are: number of estimated claims and total costs respectively 3,22 and 2,08 times higher than in Figueira da Foz. The distribution on the map of the estimated CR for each council, together with the trajectory of the hurricane, are shown in Figure 14.

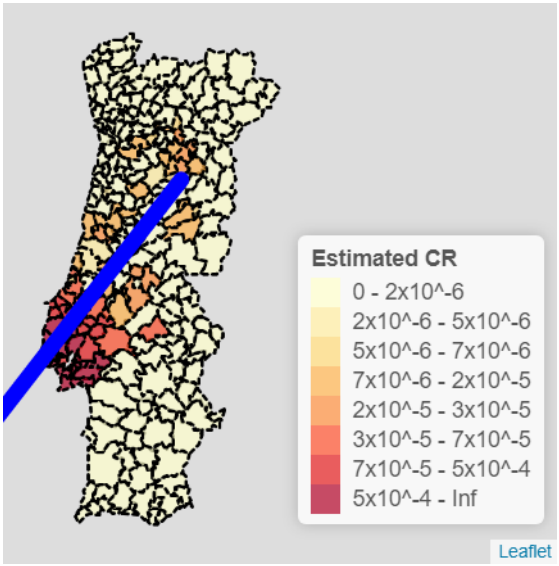

Figure 14: CR map of continental Portugal obtained for the simulated scenario of hurricane Leslie entering in Cascais with a trajectory of 45 degrees

It is now necessary to understand if the estimates obtained are reasonable. It is reasonable that, in this scenario, our model estimates higher number of claims and higher total costs, compared to the ones due to hurricane Leslie, since the Lisbon Metropolitan Area is the largest urban area of the country. As reported in Tables 14 and 15, respectively, the number of properties insured by the company in an area of 54 km around the point of entrance of the storm is 3.9 times higher in Cascais than in Figueira da Foz. Also, the amount of capital insured in the same area is 3.8 times higher in Cascais. We will refer often to the value of 54 km. This value corresponds to the first level of the variables *intensity* and *intensity2*. In regression (17), the coefficient estimated for the first level of *intensity* is the highest among all the other levels of this variable and the same is valid for regression (20) and the variable *intensity2*, as shown in Tables 9 and 11. This means that, those policies located at a distance inferior to 54 km from the point of landfall of the hurricane, have the higher probability of having a claim and also the higher average cost for the company.

As shown in Table 13, the mean cost per claim, here denoted by MCC, and defined in Equation (24), in this scenario is 0,65 times the one observed in Figueira da Foz.

$$\text{MCC} = \frac{\text{Total Cost}}{\text{Total Number of Claims}} \tag{24}$$

A lower mean cost per claim in this scenario, when compared with Figueira da Foz, is justified by the different characteristics of the portfolio of the company over the areas affected by the hurricane. Among the predictors of regressions (17) and (20) that are characteristics of the property insured, the *Type of Housing* is the variable with the largest estimated coefficients. As seen in Chapter 3, regression (17) estimates that a single-family house has 90% higher odds of incurring in a claim than an apartment, and regression (20) estimates that the average cost for a single-family house is 96% higher than for an apartment. Table 16 presents the percentage of single-family houses, apartments and other types of housing in the area of 54 km around the entrance point of the hurricane. The concentration of single-family houses around Cascais is 37% lower than that of Figueira da Foz, while the concentration of apartments is 36% higher. The lower MCC can be partially explained by the different concentration of single-family houses in the two areas.
For the second simulation, Porto is choosen as entrance point of the hurricane in Portugal. This choice is due to the fact that the Metropolitan Area of Porto is the second largest in the country, regarding population, after Lisbon, representing a major area of exposure for the company in the north of the country. In this simulation the degree of the trajectory is assumed to be of 330 degrees. As reported in Table 13, the results obtained for this simulation are: number of estimated claims and total costs respectively 1,99 and 2,62 times higher than in Figueira da Foz. The estimated CR distribution over continental Portugal, together with the trajectory of the hurricane, are shown in Figure 15.

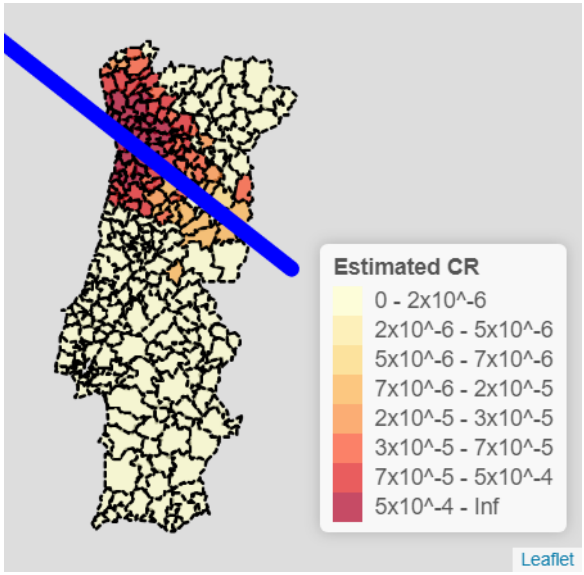

Figure 15: CR map of continental Portugal obtained for the simulated scenario of hurricane Leslie entering in Porto with a trajectory of 330 degrees

It is reasonable that the simulated number of claims and total costs are higher than

in Figueira da Foz, since the Porto Metropolitan Area, is the second largest urban area of the country. As reported in Tables 14 and 15, respectively, the number of properties insured by the company in an area of 54 km around the entrance point of the storm is 3.1 times higher in Porto than in Figueira da Foz. Also, the amount of capital insured in the same area is 2.9 times higher in Porto region than in Figueira da Foz. As shown in Table 13, the MCC is 0,76 times the one observed in Figueira da Foz.
Although Porto has less properties and capital insured than Cascais, the number of claims and the total costs estimated are very similar. This is explained by the different concentration of apartments and single-family houses in these two areas. Table 16 reports the concentration of the categories of the variable *Type of Housing* for the policies in the portfolio located in an area of 54 Km around Porto. The concentration of single-family houses is 25,5% higher around Porto than around Cascais.
In the last case scenario, Faro is choosen as entrance point of the storm in Portugal. The choice is due to the fact that Faro is the largest city in Algarve, representing a major area of exposure for the company in the south of the country. In this simulation the degree of the trajectory is assumed to be of 60 degrees. As reported in Table 13, the results obtained for this simulation are the following: number of estimated claims and total costs respectively 0,51 and 0,45 times the ones observed in Figueira da Foz.
The estimated CR distribution and the trajectory of the hurricane, are shown in Figure 16.

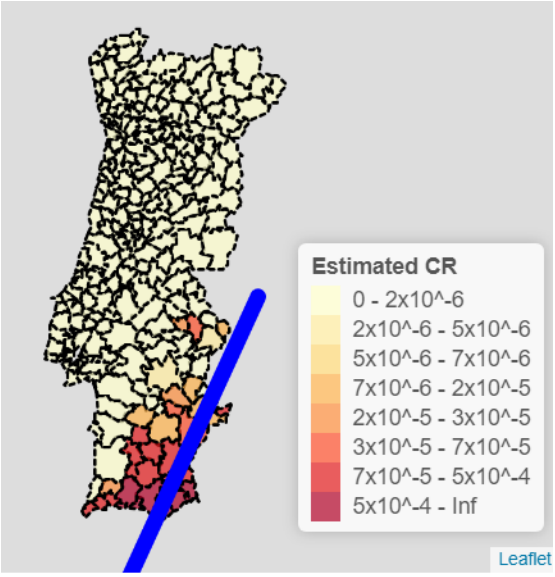

Figure 16: CR map of continental Portugal obtained for the simulated scenario of hurricane Leslie entering in Faro with a trajectory of 60 degrees

It is reasonable that the simulated number of claims and total costs are now lower than in Figueira da Foz. As reported in Tables 14 and 15, respectively, the number of properties insured by the company, and also the amount of capital insured in an area of 54 km around Faro, is approximately half of that in Figueira da Foz. As

shown in Table 13, the MCC is 0,87 times the one observed in Figueira da Foz. The fact that the MCC is similar in Faro and Figueira da Foz is expectable since one can notice that, the number of properties insured and capital insured in Faro, are almost proportional to the ones in Figueira da Foz. The difference in the MCC in these two scenarios can be explained by the fact that the concentration of single-family houses is 20% lower around Faro than around Figueira da Foz.

Landfall Point	Total Cost (€)	Total Numb.of Claims	MCC (€)
Figueira da Foz	x	y	z
Cascais	2,08x	3,22y	0,65z
Porto	1,99x	2,62y	0,76z
Faro	0,45x	0,51y	0,87z

Table 13: Total cost, Total Number of Claims and MCC for the cases of Figueira da Foz (benchmark), and the simulated scenarios of Cascais, Porto and Faro

Landfall Point	$\frac{\text{Numb. Prop. Insured in i}}{\text{Numb. Prop. Insured in Fig.da Foz}}$
Figueira da Foz	1
Cascais	3,9
Porto	3,1
Faro	0,55

Table 14: Table of the ratios between the number of properties insured in a radius of 54 km around the landfall point i and the number of properties insured in the same area around Figueira da Foz

Landfall Point	$\frac{\text{Amount Capital Insured in i}}{\text{Amount Capital Insured in Fig.da Foz}}$
Figueira da Foz	1
Cascais	3,8
Porto	2,9
Faro	0,54

Table 15: Table of the ratios between the total amount of capital insured in a radius of 54 km around the landfall point i and the total amount of capital insured in the same area around Figueira da Foz

Landfall Point	Type of Housing		
	apartment	single-family house	other
Figueira da Foz	33,95 %	51,98 %	14,07 %
Cascais	69,99 %	14,94 %	15,06 %
Porto	43,96 %	40,46 %	15,58 %
Faro	56,33 %	31,52 %	12,15 %

Table 16: Concentration of the variable *Type of Housing* for those policies located in an area inferior to 54 km around the landfall point

5 Assessing the average loss over continental Portugal

The interest of the insurance company is to mutualize the loss deriving from extreme events, like hurricane-force tropical cyclones, among the members of the portfolio. Nevertheless, if the company had to perform the mutualization based on its past experience, the properties insured in the area of Coimbra and Figueira da Foz (the ones mainly affected by hurricane Leslie in 2018) would be penalized by an excessive premium. The random nature of meteorological extreme events, encourages the insurer to look for solutions that mutualize the expected loss among other locations that could be affected in the future. The simulations in Chapter 4, used to produce estimates for the expected cost under different scenarios, can be repeated a large number of times selecting each time (i) a different point of entrance, and (ii) a different trajectory for the hazard. Afterwards all the estimates produced can be averaged to obtain the average expected loss in each council in continental Portugal. At each iteration of the process, the entrance point and the degree of the trajectory are randomly sampled from a probability distribution arbitrarily assigned. As already said in the previous chapter, any council on the coast can be a possible point of entrance of the hurricane.
The coast of Portugal is divided in 4 zones and to each zone is assigned a probability of being the entrance point of the hurricane. The first zone goes from Vila Real de Santo Antonio to Vila do Bispo, representing the coast of the Algarve region. The second zone goes from Aljezur to Alcacer to Sal, representing the coast of Alentejo region. The third zone goes from Setubal to Vagos, representing the coast of the center of Portugal. The fourth zone goes from Ilhavo to Caminha, representing the coast of the north region. Then, a probability distribution is assigned over the councils belonging to each of the four zones. The degree of the trajectory is also randomly sampled after the algorithm has randomly selected both the zone and the council. Since it is assumed that the hurricane moves eastward, the possible range of degrees is set between 270 and 90 degrees, if it affects the west coast, and between 0 and 90 degrees if affects the south coast.
The iterative process described before is shown in the Algorithm 2.
By using this strategy, at each iteration we can save the estimates obtained for each scenario and eventually average all of them to compute important quantities, such as the total cost distribution, the average CR for each council and the risk premium for the policies belonging to the risk classes we are interested to analyze.

Algorithm 2

1. For i = 1 to N:
(a) sample a Zone;
(b) sample a Council;
(c) sample a degree of the trajectory;
(d) draw the trajectory and compute D_1, D_2, *intensity* and *intensity2*;
(e) use RF Model, regression (17) and regression (20) to estimate respectively which councils will be affected, which policies will have a claim and its cost.

A risk class is formed by the intersection of the levels of the variables that define the characteristics of the policies. To estimate the average risk premium for a policy belonging to a certain risk class i, as defined in Equation (25), it is only necessary to divide the average cost predicted for the class, which is equal to the sum of the total costs predicted for the class i in each simulated scenario j divided by the number of simulations N, by the number of policies belonging to the class.

5.1 Building a risk map

The information provided by the tracking charts in Appendix A is not enough to infer which part of the coast of continental Portugal has higher risk of being hitted by a tropical cyclone. But it helps in choosing which probability distribution assign. From the inspection of the tracks of the 6 tropical cyclones, 5 of them affected the west coast of Portugal and 1 the south coast. Also, out of the 5 events which landfall in the west coast, just Joaquin, in 2015, affected the Alentejo region. The other 4 events, instead, landfall in the part of the coast defined by the zones 3 and 4. Based on this, the probabilities assigned to the 4 zones in which the coast of Portugal is divided, are reported in Table 17. The probabilities in Table 17 are conditional probabilities since they represent the probability of being hit by the hurricane given that continental Portugal is affected by it.

Zone	Description of the zone	Probability
1	from Vila Real de Santo Antonio to Vila do Bispo	1/6
2	from Aljezur to Alcacer to Sal	1/6
3	from Setubal to Vagos	2/6
4	from Ilhavo to Caminha	2/6

Table 17: Probabilities assigned to the 4 zones in which the coast of continental Portugal is divided

Then to each of the councils belonging to one of the four zones, the same probability of being the point of entrance of the hurricane is assigned. This choice is due to the lack of any specific information that could lead to assign more weight to some council, rather then another. Among the 6 tropical cyclones that affected continental Portugal since 1995, the part of the trajectory which is over land of 4 of these cyclones, points northward. So, we can consider that the cyclone's trajectory, in those cases, had a positive degree. Based on this, we want to assign more weight

to the positives trajectories. We decided to sample the angle of the trajectory from a uniform distribution between 0 and 90 degrees, if the point of entrance of the storm is a council in zone 1 (Algarve coast), and from a triangular distribution between 270 and 90 degrees with mode 45 degrees, if the point of entrance of the storm is a council in zones 2, 3 or 4 (west coast). The triangular distribution has been choosen because it is a bounded distribution and allows to easily set its mode. The value of 45 degrees has been choosen to give more probability to the positive angle trajectories, as mentioned before. The following results are relative to 1000 iterations of Algorithm 2.
Table 18 displays how many times each zone has been sampled in the process.

Zone	Frequency
1	0,18
2	0,16
3	0,33
4	0,33

Table 18: Frequency with which each of the zones has been selected by the algorithm in the simulation

Table 19 displays the distribution of the total costs. It is interesting to notice that in 30% of the scenarios, a hurricane with the power of Leslie would cause a total loss smaller or equal to half of the total loss that Leslie caused to the company. Also, in 56% of the simulated scenarios, a hurricane of this magnitude would have caused a higher loss for the company. Also from the simulation, the estimated average cost for a hurricane of the magnitude of Leslie affecting continental Portugal, is 1,14 times higher than the one due to Leslie in Figueira da Foz, while the maximum cost estimated is 2,32 times higher.

Total Cost (TC)	Percentage
$TC < 0{,}5$	30 %
$0{,}5 \leq TC < 1$	14,1 %
$1 \leq TC < 1{,}5$	21,8 %
$1{,}5 \leq TC < 2$	20,4 %
$TC \geq 2$	13,7 %

Table 19: Distribution of the Total Cost, relative to the total cost due to Leslie, for 1000 different simulated scenarios over continental Portugal

$$\hat{Average\ Cost_i} = \frac{\sum_{j=1}^{N} \hat{Total\ Cost_{ij}}}{N}, \quad RiskPremium_i = \frac{\hat{Average\ Cost_i}}{\#policies\ in\ class\ i} \tag{25}$$

Table 20 reports the average risk premium for the policies belonging to the classes obtained intersecting the different categories of the variables *Type of Property*, *Year of Construction*, *Capital Insured*, *Type of Framing*, *Type of Housing* and *Region*[7].

[7]defined in Appendix B

T.o.P	Y.o.C	Cap. Ins.	Type	T.o.F	Region	Risk prem.
Building	]2002;2018]	> 165.000	single-fam	res.cluster	ALGARVE	58,93
Building	]2002;2018]	> 165.000	single-fam	other	ALGARVE	55,15
Building	]1992;2002]	> 165.000	single-fam	res.cluster	ALGARVE	53,75
Building	]1982;1992]	> 165.000	single-fam	res.cluster	ALGARVE	53,11
Building	]1992;2002]	> 165.000	single-fam	other	ALGARVE	51,01

Table 20: The 5 highest estimated risk premiums for those risk classes of policies given by the intersection of the levels of the variables *Type of Property*, *Year of Construction*, *Capital Insured*, *Type of Framing*, *Type of Housing* and *Region*

As we would expect from the coefficients estimated for regressions (17) and (20) reported in Tables 9 and 11, the higher risk premiums are estimated for the single-family houses with capital insured higher than 165.000 €. Also, from the simulation, the higher risk premiums are obtained for those policies located in the Algarve region. This result can be explained by the high concentration of single-family houses in that area. Figure 17 reports the distribution of the average CR estimated for each council, while Table 21 reports the 5 councils with the highest estimated average CRs.

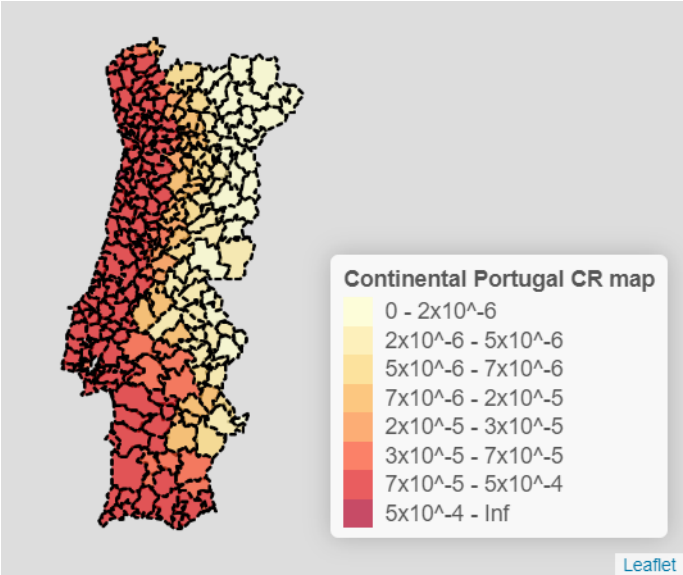

Figure 17: Distribution of the CR by council over continental Portugal due to a hurricane-force tropical cyclone

As expected, the councils with the higher CRs are the ones close to the coast. Also those in the central region of continental Portugal are the ones with higher CRs. This result can partially be explained by the higher probability assigned to the zones 3 and 4 of being affected and also to the high concentration of single-family houses in those regions.

Council	Region	Average CR
Murtosa	Center	0.0005621
Estarreja	Center	0.0004930
Vagos	Center	0.0004647
Montemor-o-Velho	Center	0.0004134
Ovar	Center	0.0003515

Table 21: Councils of continental Portugal with the highest estimated CR due to a hurricane-force tropical cyclone

6 Conclusions

The lack of long series of data regarding insurance losses due to weather-related events and of meteorological variables regarding the magnitude of these hazards, makes the estimation of the losses due to this type of events a difficult challenge for the insurer.

This work describes an alternative approach to model the damages caused by an extreme meteorological hazard which does not require the use of climate data or the use of the losses caused by other extreme events. Nevertheless we can not estimate the expected costs related to an event with different magnitudes of the one observed. Also, this approach requires the damages produced by the natural hazard to be relatively localized over the area under analysis, and distributed around an imaginary line, called trajectory.

Based on the loss data due to hurricane Leslie, we used a logistic regression to model the probability for a policy of having a claim and a multiple linear regression to model the average cost when a claim happens. The results obtained from the analysis of the coefficients of the two regressions, give interesting results which indicate to the insurer which policies carry the higher risk when affected by this type of events.

The approach used in this work to model the costs provoked by hurricane Leslie is easily applicable to any other location in continental Portugal, through simulation, and enables to quickly estimate the expected losses in many different scenarios. This allows the company to build a risk map for the occurrence of extreme tropical cyclones in continental Portugal. We provide an algorithm to do that.

The results obtained allowed to define the total cost distribution, the average risk premium for the policies belonging to the risk classes of interest, and the average cost ratio for the councils of continental Portugal.

Since the estimated models have been calibrated on the losses provoked by a single event, in order to have more reliable results, the information obtained from the losses due to other extreme meteorological events, should be considered by the insurer in his final evaluations.

A Tracking Charts

NOAA Tropical cyclones tracking charts of the years where, at least a tropical cyclone, affected continental Portugal (Images provided by the NOAA/ESRL Physical Sciences Laboratory, Boulder Colorado from their web site at https://www.psd.noaa.gov/):

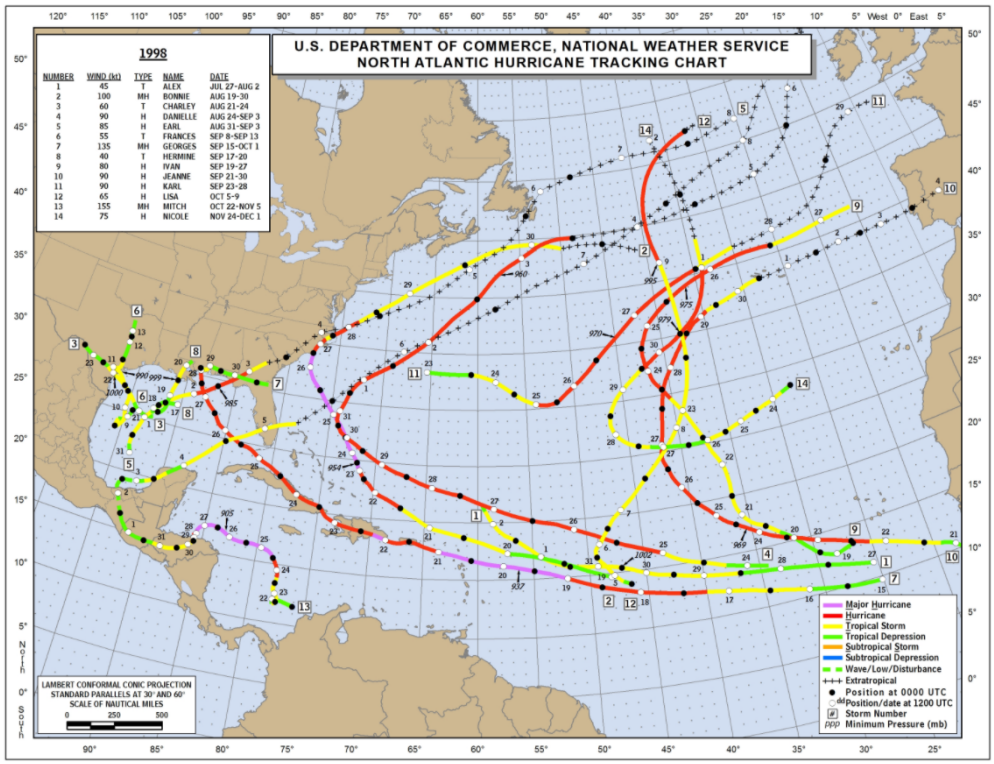

Figure 18: 1998 North Atlantic Hurricane Season Track Map

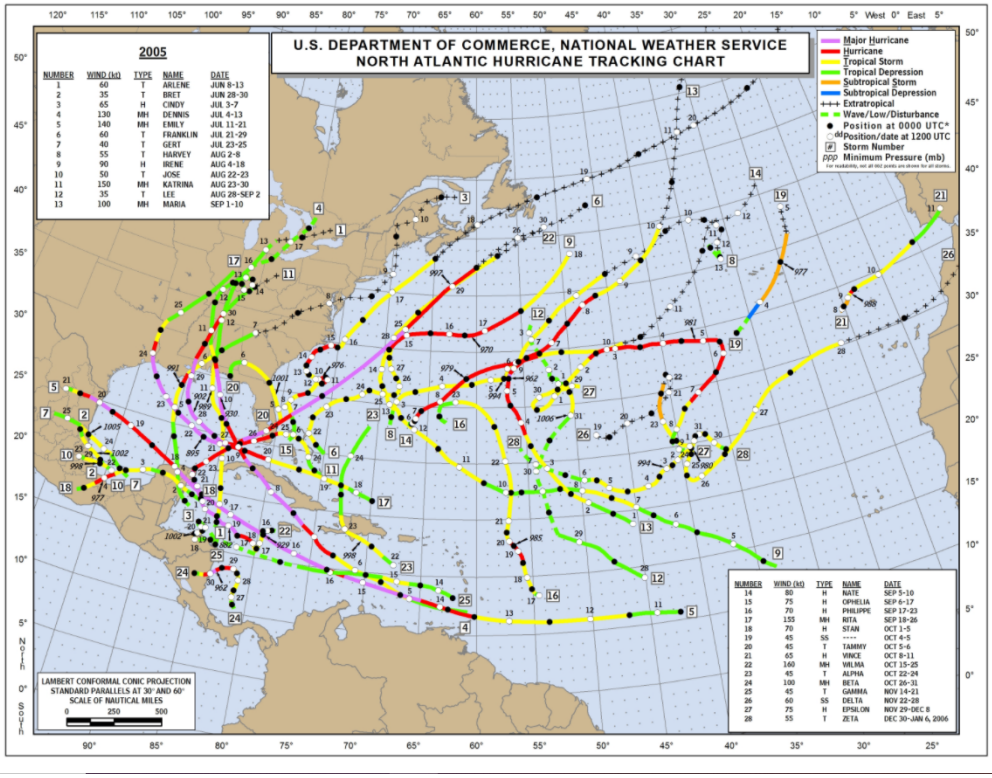

Figure 19: 2005 North Atlantic Hurricane Season Track Map

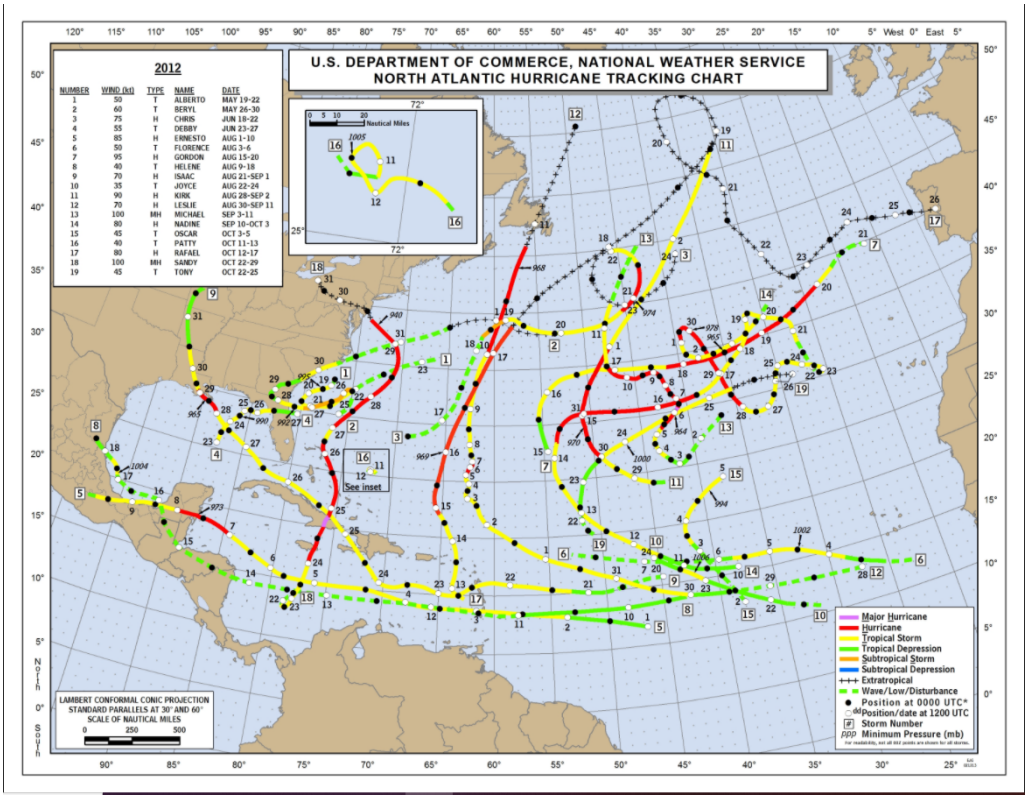

Figure 20: 2012 North Atlantic Hurricane Season Track Map

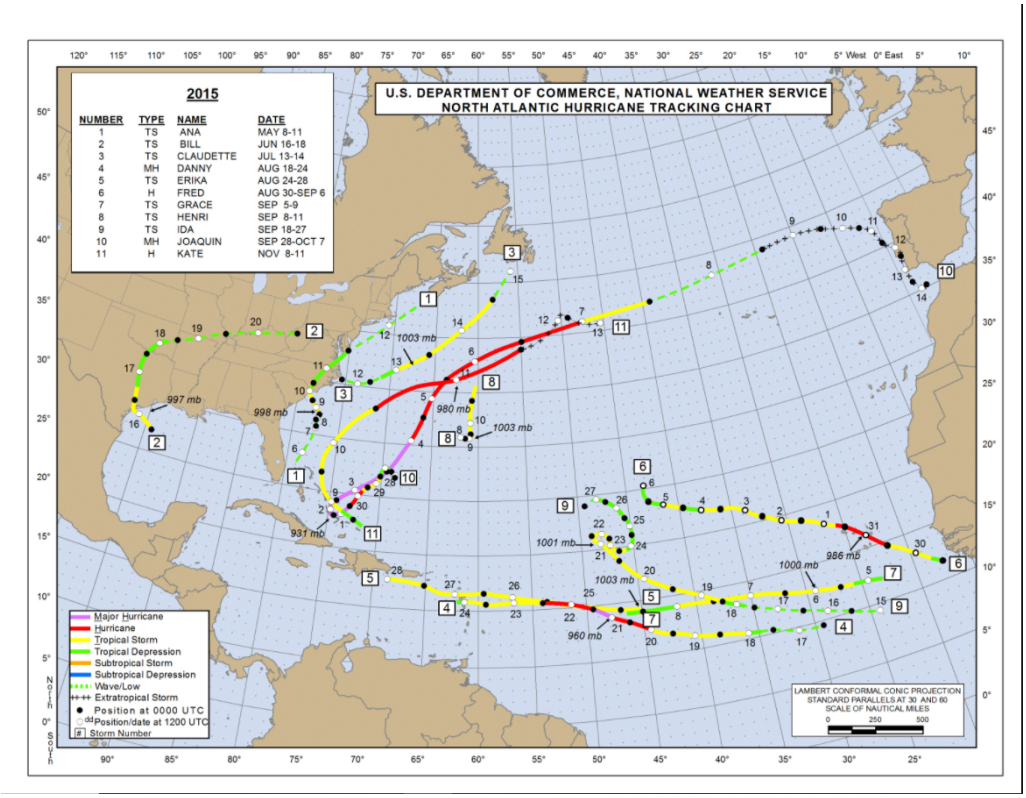

Figure 21: 2015 North Atlantic Hurricane Season Track Map

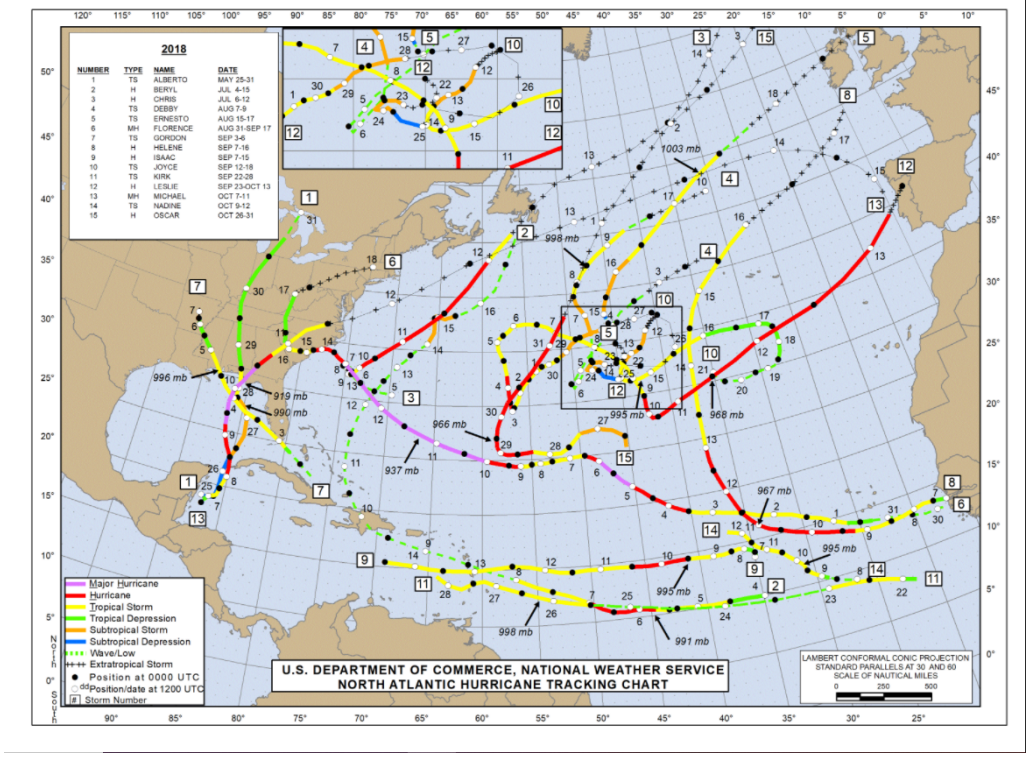

Figure 22: 2018 North Atlantic Hurricane Season Track Map

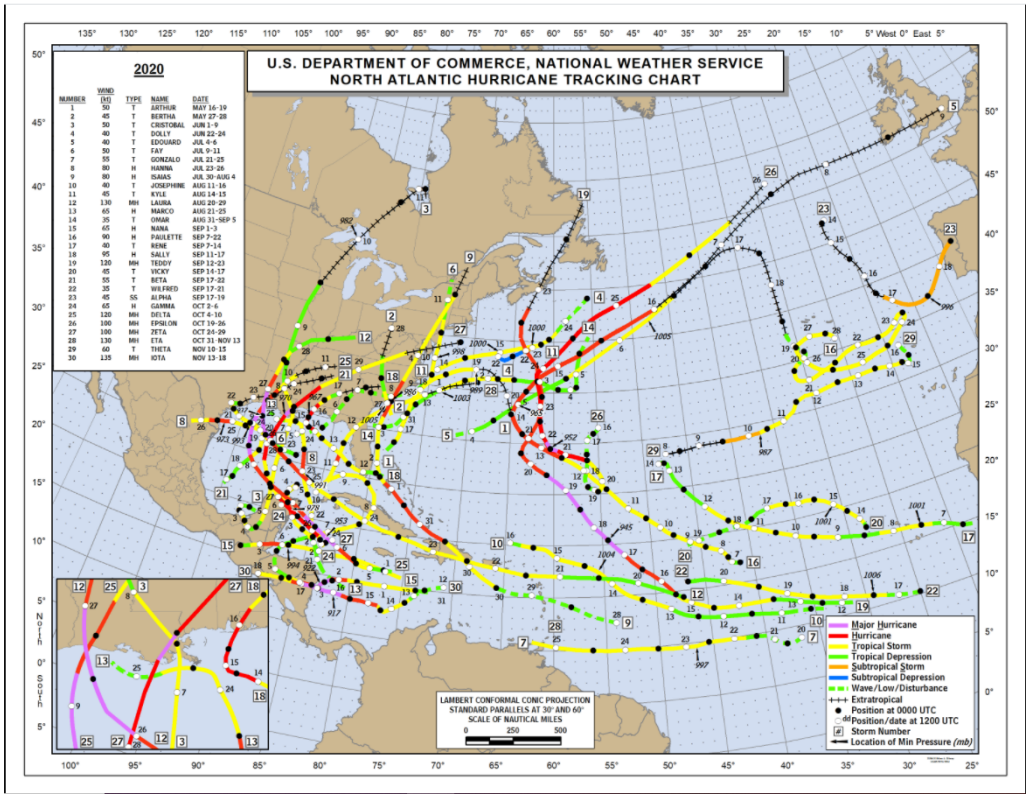

Figure 23: 2020 North Atlantic Hurricane Season Track Map

B Variables

The variables provided by the company are:

- *Type of Property* (T.o.P) with levels:
  - Content
  - Building
- *Year of Construction* (Y.o.C) with levels:
  - LEVEL 1 if the year of construction $\leq$ 1982
  - LEVEL 2 if 1982 $<$ year of construction $\leq$ 1992
  - LEVEL 3 if 1992 $<$ year of construction $\leq$ 2002
  - LEVEL 4 if the year of construction $>$ 2002
- *Framing of the Housing* (Framing) with levels:
  - Residential Cluster
  - Semi-detached house
  - Other
- *Type of Housing* (Type) with levels:
  - Apartment
  - Single-family house
  - Other
- *Type of Floor* (T.o.F) with levels:
  - LEVEL 1 = sub cave or ground floor or first floor or intermediate floor
  - LEVEL 2 = last floor
  - Not defined
- *Capital Insured* (Cap. Ins) with levels:
  - LEVEL 1 if capital insured $\leq$ 80000 €
  - LEVEL 2 if 80000 € $<$ capital insured $\leq$ 120000 €
  - LEVEL 3 if 120000 € $<$ capital insured $\leq$ 165000 €
  - LEVEL 4 if capital insured $>$ 165000 €

The variables obtained through public available information are:

- *Intensity* with levels
  - LEVEL 1 if D1 $<$ 54 Km
  - LEVEL 2 if 54 km $\leq$ D1 $<$ 78 Km

  - LEVEL 3 if 78 Km $\leq$ D1 $<$ 107 Km
  - LEVEL 4 if D1 $\geq$ 107 Km and D2 $<$ 44 Km
  - LEVEL 5 if D1 $\geq$ 107 Km and D2 $\geq$ 44 Km

- *Intensity 2* with levels
  - LEVEL 1 if D1 $<$ 54 Km
  - LEVEL 2 if 54 km $\leq$ D1 $<$ 78 Km
  - LEVEL 3 if 78 Km $\leq$ D1 $<$ 107 Km
  - LEVEL 4 if D1 $\geq$ 107 Km and D2 $<$ 25 Km
  - LEVEL 5 if D1 $\geq$ 107 Km and D2 $\geq$ 25 Km

- *Altitude* is the height over the sea level, has levels:
  - LEVEL 1 if altitude $\leq$ 90 m
  - LEVEL 2 if 90 m $<$ altitude $\leq$ 200 m
  - LEVEL 3 if altitude $>$ 200 m

  The quotients of location [8] (Q.L) for each council, of the variables *Forest Area*, *Bush Area* and *Urban Area*, have been obtained from the report [9] of 2018 from the INE (Instituto Nacional de Estatistica).

- *Forest Area* represents the Q.L of the forest area in the council. Has levels:
  - LEVEL 1 if Q.L $\leq$ 1,45
  - LEVEL 2 if Q.L $>$ 1,45

- *Bush Area* represents the Q.L of the bush area in the council. Has levels:
  - LEVEL 1 if Q.L $\leq$ 0,19
  - LEVEL 2 if Q.L $>$ 0,19

- *Urban Area* represents the Q.L of the urban area in the council. Has levels:
  - LEVEL 1 if Q.L $\leq$ 2,12
  - LEVEL 2 if 2,12 $<$ Q.L $\leq$ 3,35
  - LEVEL 3 if Q.L $>$ 3,35

[8] the ratio between the proportion of each class of use and occupation of the territory in that specific council and the respective proportion in the continent

[9] The full report can be consulted at `https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=435668469&DESTAQUESmodo=2`

- *Region* [10] with levels:
  - North
  - Center
  - Metropolitan Area of Lisbon (MAL)
  - Alentejo
  - Algarve

Significance codes:

Significance code	p-value
***	[0, 0.001]
**	(0.001, 0.01]
*	(0.01, 0.05]
.	(0.05, 0.1]
	(0.1, 1]

Table 22: Significance codes Table

[10] continental Portugal has been divided in 5 regions as done in the NUTS II https://www.pordata.pt/en/What+are+NUTS

References

[1] AON. Global catastrophe recap first half of 2021 , year = 2021, url = http://thoughtleadership.aon.com/Documents/20212107$_a nalytics - if - 1H - global - recap.pdf, urldate = 2021 - 07 - 21$.

[2] Lixion A. Avila. Tropical Cyclone Report Hurricane Rafael 12-17 October 2012 . Technical report, National Hurricane Center, NOAA, 01 2013.

[3] Robbie Berg. Tropical Cyclone Report Hurricane Joaquin 28 September-7 October 2015. Technical report, National Hurricane Center, NOAA, 01 2016.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Daniel P. Brown. Tropical Cyclone Report SUBTROPICAL STORM ALPHA 17-19 September 2020. Technical report, National Hurricane Center, NOAA, 01 2021.

[6] Markus G Donat, Tobias Pardowitz, GC Leckebusch, Uwe Ulbrich, and Olaf Burghoff. High-resolution refinement of a storm loss model and estimation of return periods of loss-intensive storms over germany. *Natural Hazards and Earth System Sciences*, 11(10):2821–2833, 2011.

[7] James L. Franklin. Tropical Cyclone Report Hurricane Vince 8-11 October 2005. Technical report, National Hurricane Center, NOAA, 02 2006.

[8] Christian LE Franzke. Impacts of a changing climate on economic damages and insurance. *Economics of Disasters and Climate Change*, 1(1):95–110, 2017.

[9] Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[10] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[11] Reindert J Haarsma, Wilco Hazeleger, Camiel Severijns, Hylke De Vries, Andreas Sterl, Richard Bintanja, Geert Jan Van Oldenborgh, and Henk W van den Brink. More hurricanes to hit western europe due to global warming. *Geophysical Research Letters*, 40(9):1783–1788, 2013.

[12] Robert E Hart and Jenni L Evans. A climatology of the extratropical transition of atlantic tropical cyclones. *Journal of Climate*, 14(4):546–564, 2001.

[13] P. Heneka, T. Hofherr, B. Ruck, and C. Kottmeier. Winter storm risk of residential structures ndash; model development and application to the german state of baden-württemberg. *Natural Hazards and Earth System Sciences*, 6(5):721–733, 2006.

[14] Peter Hoeppe. Trends in weather related disasters – consequences for insurers and society. *Weather and Climate Extremes*, 11:70–79, 2016. Observed and Projected (Longer-term) Changes in Weather and Climate Extremes.

[15] Barry D Keim, Robert A Muller, and Gregory W Stone. Spatial and temporal variability of coastal storms in the north atlantic basin. *Marine Geology*, 210(1-4):7–15, 2004.

[16] Munich.Re. Hurricanes, typhoons and cyclones, url = https://www.munichre.com/en/risks/natural-disasters-losses-are-trending-upwards/hurricanes-typhoons-cyclones.html, urldate = .

[17] The National Oceanic and Atmospheric Administration (NOAA). Total and average number of tropical cylones by month (1851-2017), year = 2016, url = http://www.aoml.noaa.gov/hrd/tcfaq/tcfaqE.html, urldate = 2016-06-01.

[18] Richard J. Pasch. Preliminary Report Hurricane Jeanne 21 September - 1 October, 1998. Technical report, National Hurricane Center, NOAA, 02 1999.

[19] Richard J. Pasch and David P. Roberts. Tropical Cyclone Report Hurricane Leslie 23 SEPTEMBER–13 OCTOBER 2018 . Technical report, National Hurricane Center, NOAA, 03 2019.

[20] B. F. Prahl, D. Rybski, M. Boettle, and J. P. Kropp. Damage functions for climate-related hazards: unification and uncertainty analysis. *Natural Hazards and Earth System Sciences*, 16(5):1189–1203, 2016.

[21] Boris F Prahl, Diego Rybski, Olaf Burghoff, and Jürgen Peter Kropp. Comparison of storm damage functions and their performance. *Natural Hazards and Earth System Sciences*, 15(4):769–788, 2015.