

# Digital Vulnerabilities: A Statistical Analysis

Filipe Alexandre Araújo da Costa

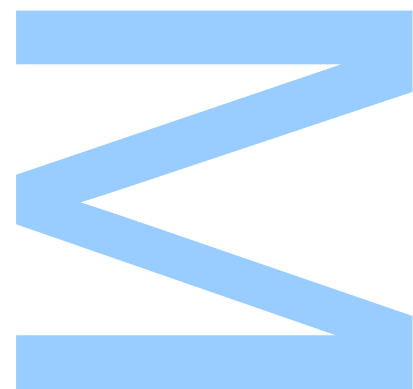
Mestrado de Estatística Computacional e Análise de Dados

Departamento de Matemática

2022

**Orientadora**

Prof. [Margarida Brito](#), Faculdade de Ciências





**U.** PORTO

**FC** FACULDADE DE CIÊNCIAS  
UNIVERSIDADE DO PORTO

Todas as correções determinadas  
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, \_\_\_\_ / \_\_\_\_ / \_\_\_\_

**W**

**S**

**Q**



UNIVERSIDADE DO PORTO

MASTERS THESIS

---

# Digital Vulnerabilities: A Statistical Analysis

---

*Author:*

Filipe COSTA

*Supervisor:*

Margarida BRITO

*A thesis submitted in fulfilment of the requirements  
for the degree of Mestrado de Estatística Computacional e Análise de Dados*

*at the*

Faculdade de Ciências da Universidade do Porto  
Departamento de Matemática

January 26, 2023



## *Declaração de Honra*

Eu, Filipe Alexandre Araújo da Costa, inscrito no Mestrado em Estatística Computacional e Análise de Dados da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação reflete as perspectivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega. Ao entregar esta dissertação, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Filipe Alexandre Araújo da Costa

30 de Novembro 2022





*“Your work is going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. If you haven't found it yet, keep looking. Don't settle. As with all matters of the heart, you'll know when you find it.”*

Steve Jobs



## *Acknowledgements*

I want to thank all my colleagues and teachers who accompanied me during this trip, especially my teacher and advisor Margarida Brito for the availability she offered with all kindness. Furthermore, I would like to thank my parents, friends and family who supported me during this period. I especially want to thank Rodrigo Silva for the endless and unquestionable support that allowed me to have the strength to carry out this dissertation.



UNIVERSIDADE DO PORTO

## *Abstract*

Faculdade de Ciências da Universidade do Porto

Departamento de Matemática

Mestrado de Estatística Computacional e Análise de Dados

### **Digital Vulnerabilities: A Statistical Analysis**

by [Filipe COSTA](#)

With the increasing use of the internet and the digital world, cybercrime records have accompanied this growth, with several categories of cybercrimes being reported such as malware, ransomware or spyware. For the implementation of these cybercrimes it is necessary that there are vulnerabilities in the systems for these tools to exploit them. In this dissertation, the discovery of vulnerabilities was modeled using point process models and extreme value theory. In addition, an alarm system was designed to detect days with more vulnerabilities registered. In the point models, the non-linear marked Poisson model obtained the best performance, confirming the behavior observed in the descriptive analysis. The model supported by extreme value theory, the POT model, was successful in modeling the time of exceedances in the proposed threshold. Finally, the alarm system performs better with the data grouped weekly. Several metrics were proposed for the evaluation of the alarm system, where they suggest that the model has better predictive capacity within 30 days.



UNIVERSIDADE DO PORTO

## *Resumo*

Faculdade de Ciências da Universidade do Porto

Departamento de Matemática

Mestrado em Estatística Computacional e Análise de Dados

### **Vulnerabilidades Digitais: uma Análise Estatística**

por [Filipe COSTA](#)

Com a crescente utilização da internet e do mundo digital, os registos de cibercrime têm acompanhado esse crescimento, sendo reportadas várias categorias de cibercrimes como malwares, ransomwares ou spywares. Para a implementação de estes cibercrimes é necessário que existam vulnerabilidades nos sistemas para estas ferramentas os explorarem. Nesta dissertação foram modeladas a descoberta das vulnerabilidades utilizando modelos de processos pontuais e de teoria de valores extremos. Além disso foi desenhado um sistema de alarme para a deteção de dias com mais vulnerabilidades registadas. No modelos pontuais, o modelo de Poisson marcado não linear obteve a melhor performance, confirmando os comportamentos observados na análise descritiva. O modelo apoiado em teoria de valores extremo, o modelo POT, foi bem sucedido na modelação do tempo de excedência do limiar proposto. Por fim, o sistema de alarme tem um melhor desempenho com os dados agrupados semanalmente. Foram propostas várias métricas para a avaliação do sistema de alarme, segundo as quais o modelo tem melhor capacidade preditiva no espaço de 30 dias.





# Contents

<b>Acknowledgements</b>	<b>vii</b>
<b>Abstract</b>	<b>ix</b>
<b>Resumo</b>	<b>xi</b>
<b>Contents</b>	<b>xiii</b>
<b>List of Figures</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 State of Art</b>	<b>3</b>
2.1 Some basic definitions . . . . .	3
2.2 Vulnerabilities categorization . . . . .	4
2.3 Common modulation techniques in vulnerability modeling . . . . .	7
2.4 NVD Database . . . . .	8
2.5 Individual data bases . . . . .	9
2.5.1 Android data base . . . . .	9
2.6 Note on modeling limitations . . . . .	12
<b>3 Modelling through Point Processes</b>	<b>13</b>
3.1 Point Processes . . . . .	13
3.2 Poisson point processes . . . . .	14
3.3 Marked point process . . . . .	15
3.4 Intensity estimation . . . . .	16
3.5 Model diagnostics . . . . .	16
3.5.1 Probability plots . . . . .	17
3.5.2 Chi-square test . . . . .	17
3.5.3 Anderson-Darling test . . . . .	18
3.5.4 Wald test . . . . .	18
3.5.5 Implementation . . . . .	19
3.6 Statistical Analysis . . . . .	20
3.6.1 Investigating Android Data . . . . .	20
3.6.2 Results . . . . .	23

<b>4</b>	<b>Modelling through Classic Extreme Value Theory</b>	<b>25</b>
4.1	Extreme value theory (EVT) for independent variables . . . . .	25
4.1.1	Convergence analysis . . . . .	27
4.2	EVT for dependent variables . . . . .	28
4.2.1	Maximum stationary series . . . . .	28
4.2.2	Convergence analysis . . . . .	29
4.3	Modelling using stationary series . . . . .	30
4.3.1	Block Maximum model . . . . .	30
4.3.2	Peak over threshold - POT model . . . . .	30
4.3.2.1	Iid case . . . . .	30
4.3.2.2	Dependent case . . . . .	31
4.4	POT model - Point process perspective . . . . .	31
4.5	Parameter estimation . . . . .	33
4.6	Modeling the Android data . . . . .	34
4.7	Results . . . . .	36
<b>5</b>	<b>Modelling through EVT for discrete data</b>	<b>39</b>
5.1	Introduction . . . . .	39
5.1.1	Estimation of the upper limit of return level . . . . .	40
5.1.2	Implementation . . . . .	42
5.2	Modeling Android data . . . . .	42
5.3	Metrics . . . . .	45
5.3.1	Binary metric . . . . .	46
5.3.2	General metric . . . . .	50
5.4	Results . . . . .	50
<b>6</b>	<b>Conclusion and future work</b>	<b>53</b>
<b>A</b>	<b>Equations for CVSS metric</b>	<b>55</b>
<b>B</b>	<b>Diagnose plots</b>	<b>59</b>
<b>C</b>	<b>NB curves</b>	<b>63</b>
	<b>Bibliography</b>	<b>65</b>

# List of Figures

2.1	Top 25 of most dangerous Software Weaknesses of 2022 . . . . .	5
2.2	CVSS equation metrics [15] . . . . .	6
2.3	CVSS metric groups for version 2.0 [15] . . . . .	7
2.4	CVSS metric groups for version 3.0 [15] . . . . .	7
2.5	Number of vulnerabilities registered at NVD . . . . .	8
2.6	Scores boxplot over the years . . . . .	9
2.7	Number of vulnerabilities registered at Android . . . . .	10
2.8	Score type in Android . . . . .	10
2.9	Cumulative vulnerabilities for Android data . . . . .	11
2.10	Market Share of Android . . . . .	12
3.1	Correlation plots . . . . .	20
3.2	Output of the simple Poisson model . . . . .	21
3.3	Output of the marked Poisson model . . . . .	21
3.4	Output of the non-linear marked Poisson model . . . . .	22
4.1	Clusters graphic representation. The clusters are defined by each grey rectangle . . . . .	35
4.2	Correlation plots . . . . .	35
4.3	Comparison between the model and observed data . . . . .	36
4.4	QQ-plot between the model and observed data . . . . .	36
5.1	Grid search plots . . . . .	43
5.2	Return levels for different aggregated data . . . . .	44
5.3	Alarme system for different aggregated data . . . . .	45
5.4	PR curves for various type of aggregated data . . . . .	47
5.5	NB curves for various type of aggregated data . . . . .	48
5.6	NB curves for various window sizes . . . . .	49
A.1	Equation for base calculation for CVSS in version 2.0 . . . . .	55
A.2	Equation for temporal calculation for CVSS in version 2.0 . . . . .	56
A.3	Equation for environmental calculation for CVSS in version 2.0 . . . . .	56
A.4	Equation for base calculation for CVSS in version 3.0 . . . . .	57
A.5	Equation for Temporal calculation for CVSS in version 3.0 . . . . .	57
A.6	Equation for Modified impact Sub-Score (MISS) calculation for CVSS in version 3.0 . . . . .	57
A.7	Equation for Environmental calculation for CVSS in version 3.0 . . . . .	57
A.8	Numerical values for version 3.0 . . . . .	58

B.1	Diagnose plots for Critical simple Poisson model . . . . .	59
B.2	Diagnose plots for High simple Poisson model . . . . .	59
B.3	Diagnose plots for Medium simple Poisson model . . . . .	60
B.4	Diagnose plots for Low simple Poisson model . . . . .	60
B.5	Diagnose plots for Critical marked Poisson model . . . . .	60
B.6	Diagnose plots for High marked Poisson model . . . . .	60
B.7	Diagnose plots for Medium marked Poisson model . . . . .	61
B.8	Diagnose plots for Low marked Poisson model . . . . .	61
B.9	Diagnose plots for Critical non linear marked Poisson model . . . . .	61
B.10	Diagnose plots for High non linear marked Poisson model . . . . .	61
B.11	Diagnose plots for Medium non linear marked Poisson model . . . . .	62
B.12	Diagnose plots for Low non linear marked Poisson model . . . . .	62
C.1	NB curve with accepting window size of 2 weeks . . . . .	63
C.2	NB curve with accepting window size of 6 weeks . . . . .	63

# Chapter 1

## Introduction

The use of the Internet and its services has been increasing over the last decade, expanding its influence and restructuring society and its behaviors. If we add the various emerging needs due to the Covid-19 pandemic, companies and services accelerated their digital transition even more. With these, the need to protect critical information such as credit card numbers and personal data has grown over the years, as has the attempt to obtain them. In Portugal and around the world, there has been a significant increase in cyber-crime cases [11], such as phishing, malware, ransomware and fraud. According to the World Economic Forum's 2022 Risk Report [46], 85% of the world's leaders in the WEF Cybersecurity Leadership Community show concern about the growth of cyber-crime cases, specifically ransomware, ranking them as a major public safety concern. Although the risk of cyber-crime happening has never been so high, the number of professionals and investigation has not kept up with the demand, with an estimated lack of 3 million professionals in the area to investigate, correct and mitigate the various vulnerabilities that affect the IT areas.

One type of cybercrime involves the creation of exploits to take advantage of vulnerabilities in software and systems infrastructure in order to gain unauthorized access, remove service operability or obtain valuable information. This type of attacks are very important to avoid as they can bring a lot of material and immaterial damage to the entities that suffer them. Recent cases such as Vodafone, Sonae or TAP are examples of this type of attack. Entities are interested in mitigating these events, either by strengthening their security systems or reducing the number of attack vectors, which in this case include vulnerabilities. In this dissertation we will focus on the study of the latter, that is, on the

vulnerability discovery process. There is already a lot of literature on this subject, mostly based on empirical models [5][29] or differential equations [36].

This work has the objective of exploring and applying statistical frameworks to model vulnerabilities, applying different approaches that offer different perspectives in each situation. The text is structured based on six chapters:

In Chapter I, a brief introduction to the topic is given, in the context of the cyber-crime landscape in Portugal and in the world.

In Chapter II, necessary definitions for the rest of the dissertation are introduced, as well as characteristics in their categorizations. In addition, it is briefly discussed the different common approaches to vulnerability modeling. Finally, the considered databases in this dissertation are described, offering some relevant information for further modeling.

In Chapter III, the topic of point processes is addressed. The concepts of simple and marked Poisson point process are explored, in addition to the required conditions for application. The model diagnoses that are applied in the rest of the dissertation are also presented. Finally, the parameters for the model are estimated using the available data and the results are discussed.

In Chapter IV, the classical extreme value theory is investigated, culminating in the definition of a general model called POT, peak over thresholds. The modelling of the data following this approach is presented, and its results are commented.

In Chapter V, the extreme value theory is applied again, but this time applied to discrete structures. Thus, an alarm system is introduced in order to predict the occurrence of an anomalous number of vulnerabilities. Various metrics are applied to this alarm system and corresponding results are discussed.

Finally, in Chapter VI, all the results are clustered, and a general comment is made comparing the performance of each approach. In addition, limitations in vulnerability modeling are discussed, as well as future work that can be developed to overcome these difficulties.

# Chapter 2

## State of Art

### 2.1 Some basic definitions

Throughout the evolution of computer security, the concept of vulnerability has also been improved. However, there is no consensus among the community on the exact definition, only suggestions from various organizations with influence in the area. For example, CVE (Common Vulnerability Enumeration) [13] defines vulnerability as “A weakness in computational logic found in software and hardware components that, when exploited, results in a negative impact on product confidentiality, integrity or availability”, while NIST (National Institute of Standards and Technology) [34] defines it as “ A weakness in an information system, security protocol, internal controls or implementation that could be exploited by a threat source”. As the data gathered on this dissertation are all represented by CVE, we followed the definition of this organization.

In addition to the concept of vulnerabilities, there are other concepts that are equally important for a clear understanding of this area. One of these is the vulnerability lifecycle. The vulnerability lifecycle is defined as a series of events where a state and an associated risk are reflected. In [41] the cycle is divided into four essential moments: vulnerability discovery, vulnerability registration, patch release and exploit availability. Although an overlap of events is often observed, it is important to have a clear distinction of definitions between them. In this way, the four main events are defined as follows:

1. **Vulnerability Discovery:** In this event, the vulnerability is discovered for the first time. Depending on who discovered it, this knowledge could be used for malicious purposes.

2. **Vulnerability Registration:** First time the vulnerability information has been reported by a reputable source. At this moment, the vulnerability has already been studied by experts of risk analysis.
3. **Launch of patch:** From this moment on, the system is protected and the vulnerability is removed.
4. **Availability of an exploit:** The occurrence of this event allows a person with access to the exploit to be able to exploit the vulnerability.

It should be noted that the existence of a vulnerability does not mean the existence of an exploit, and thus, compromise of the system. Therefore, the existence of a vulnerability does not represent a risk in on itself, only if there is an exploit for the specific vulnerability. Other important points to mention are the actors in the life cycle of vulnerabilities, which vary according to their position and influence. Some of these actors are:

1. **Vendor :** Entity that produces the hardware/software and is responsible for keeping it safe
2. **Hacker \* :** Entity that discovers and/or releases exploits for vulnerabilities
3. **Independent Organization:** Entity that independently discovers and reports vulnerabilities
4. **User:** Product user

The various actors have different roles within the lifecycle and can strongly affect the security of a system. An example is, no matter how fast the vendor is to provide a patch for the vulnerability in question, if the user does not update the software, the user's system will remain vulnerable to threats. It is possible to simulate the actions of each actor for a given vulnerability through Markov models [1], but this will not be the focus of this work.

## 2.2 Vulnerabilities categorization

Due to the flexibility and different applications of the term vulnerability, it is necessary to measure intrinsic properties in order to be able to make comparisons in different domains. For this, there are several suggestions of universal categorizations and identifications that

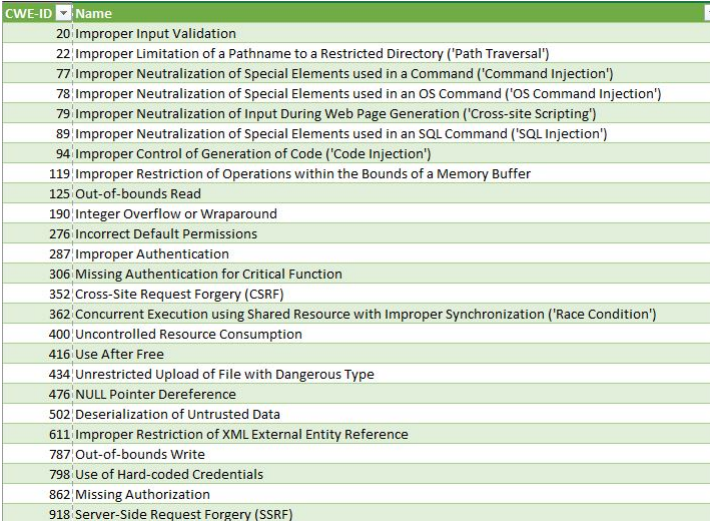
---

\*There are differences in the possible entities that can create an exploit or that gain access to them



try to group the different vulnerabilities. NIST (National Institute of Standards and Technology) is an American organization created in 1999 focused on identifying, defining and cataloging publicly disclosed vulnerabilities. In this way, organizations can publish discovered vulnerabilities so that cybersecurity professionals can easily identify and prioritize common issues.

To distinguish between the various types of vulnerabilities, a categorization often used in the industry is the CWE [16] (Common Weakness Enumeration) . This community-developed list of vulnerabilities was developed in 2006 and allows experts to categorize vulnerabilities by type, whether they are software or hardware. An example is the top 25 most dangerous vulnerabilities of 2022 according to their category (figure 2.1)



CWE-ID	Name
20	Improper Input Validation
22	Improper Limitation of a Pathname to a Restricted Directory ('Path Traversal')
77	Improper Neutralization of Special Elements used in a Command ('Command Injection')
78	Improper Neutralization of Special Elements used in an OS Command ('OS Command Injection')
79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scripting')
89	Improper Neutralization of Special Elements used in an SQL Command ('SQL Injection')
94	Improper Control of Generation of Code ('Code Injection')
119	Improper Restriction of Operations within the Bounds of a Memory Buffer
125	Out-of-bounds Read
190	Integer Overflow or Wraparound
276	Incorrect Default Permissions
287	Improper Authentication
306	Missing Authentication for Critical Function
352	Cross-Site Request Forgery (CSRF)
362	Concurrent Execution using Shared Resource with Improper Synchronization ('Race Condition')
400	Uncontrolled Resource Consumption
416	Use After Free
434	Unrestricted Upload of File with Dangerous Type
476	NULL Pointer Dereference
502	Deserialization of Untrusted Data
611	Improper Restriction of XML External Entity Reference
787	Out-of-bounds Write
798	Use of Hard-coded Credentials
862	Missing Authorization
918	Server-Side Request Forgery (SSRF)

FIGURE 2.1: Top 25 of most dangerous Software Weaknesses of 2022

Despite the different contexts and origins of vulnerabilities, one of the characteristics relevant for processing and mitigation is the severity they would have if they were exploited. Although severity often depends a lot on the context in which it appears, it is possible to highlight characteristics that are independent of those contexts to obtain a more objective view of them. For this, the FIRST [21] (Forum of Independent Report and Security Teams) created in 2005 a vulnerability severity classification system: CVSS (Common Vulnerability Scoring System). This classification makes it possible to «capture the main characteristics of the vulnerabilities and produce a numerical score that reflects the severity». Over time, the criteria used in the classification were improved to respond to changes in vulnerability landscapes. In the dataset analyzed here, two different versions of this classification were used, version 2.0 and 3.1. Both versions divide the metrics into

3 groups: base, temporal and environmental. The base metrics aim to represent the fundamental and intrinsic characteristics that are independent of context and time, while the other two groups seek to explain the variable metrics. This rating uses a scale of 0 to 10.

The interconnection between the different types of metrics, regardless of version, follows the logic embedded in the figure 2.2: In the version 2.0, base metrics are defined by

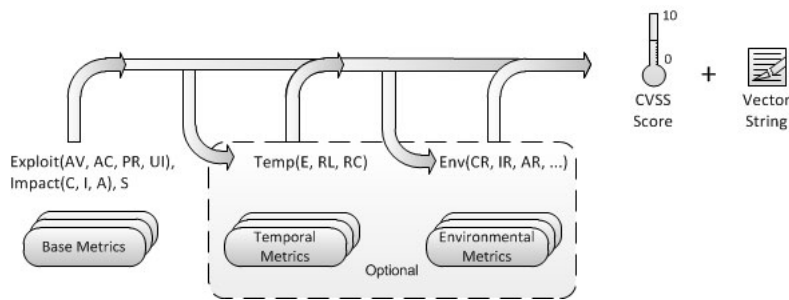


FIGURE 2.2: CVSS equation metrics [15]

attack vector, access complexity, authentication and impact on confidentiality, integrity, and availability. The attack vector defines the context where the exploit may be possible, that is, the more remote the exploitation is possible, the higher the score will be. The attack complexity reflects the conditions that the attack needs to have to be successful, the fewer conditions are needed, the higher the score, while authentication describes the need of privileges to be able to carry out the exploit. The remaining three impact metrics assess the impact the vulnerability could have on three relevant areas: confidentiality, integrity, and product availability.

With regard to temporal metrics, this version provides three: exploitability, remediation level and confidence in disclosure. The first metric describes the current state of the techniques that exploit this vulnerability, the remediation level determines if there is a fix available and confidence in disclosure measures whether the vulnerability has been confirmed by the seller itself. Environmental metrics measure severity in the context where they occur. There are five environmental metrics: potential collateral damage, target distribution and security requirements in the three areas considered (confidentiality, integrity and availability). Potential collateral damage, as the name suggests, shows whether exploiting the vulnerability could lead to possible collateral damage. The scope distribution measures the proportion of vulnerability in the system itself, i.e. whether the vulnerability is system-wide and security requirements measure the importance of the affected areas at the vendor. A graph with the different metrics by group is represented in the figure 2.3. The equation that calculates the scores known for the possible values for

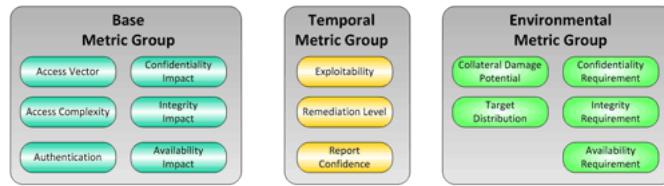


FIGURE 2.3: CVSS metric groups for version 2.0 [15]

each metric is available in [A](#).

In version 3.0, most metrics remain intact, while others are expanded. Most of the changes are located in the base metrics, with the addition of user interaction and scope metrics. The user interaction metric describes the need for some user, in addition to the attacker, to be necessary for the exploit to be successful. Scope measures the ability of exploiting the vulnerability to reach other systems that are beyond the vulnerability itself. Another relevant change is in the environmental metrics, where the relative importance for the supplier is maintained, but modified base metrics are added where the objective is to adapt them in the context of each entity. A graph with the different metrics added is represented in the figure [2.4](#). In addition, the formula that joins all metrics is also signifi-

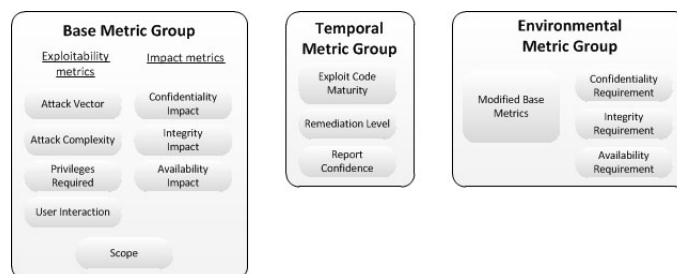


FIGURE 2.4: CVSS metric groups for version 3.0 [15]

cantly changed in [A](#).

## 2.3 Common modulation techniques in vulnerability modeling

In the literature there are several approaches to vulnerability modeling. Models that focus on modeling vulnerability discovery are called VDMs (Vulnerability Discovery Models). Within these models, two large groups can be considered: models based on the temporal component and models based on effort [3]. The first type of models models vulnerabilities taking into account when they happened, while the second is based on external factors such as the use of the software, its popularity, etc. Temporal modeling can be done through S models [4] [5], which assume that the behavior of vulnerabilities follows the

form of an S, differential equations [36] or through proposals for specific models [29]. In addition, models also try to incorporate metrics adjacent to vulnerabilities such as their severity [42]. There are also models that have the goal to simulate the process of discovering and exploiting vulnerabilities by simulating information systems [38] or hidden markov chains [1] [2]. Finally, there are also methods based on Machine learning, more specifically on neural networks [32], which avoid the parametric restrictions of the previously mentioned models.

## 2.4 NVD Database

To better understand the context and evolution of public vulnerabilities recorded over the years, we start by doing a brief exploratory analysis of the NVD database. The NVD [35] (National Vulnerability Database) is an American database that serves as a comprehensive repository/dictionary of recorded vulnerabilities. Created in 1999, this institution helped redefine the vulnerability cataloging system. Nowadays, almost all the world's software creators with a lot of users report their vulnerabilities in this public database. Here are stored the vulnerabilities identified with the CEV system, indexed by CVSS metric. There are some complementary databases such as "CVE details" [14], which provide information such as the CWE and the number of exploits created, if any. The period analyzed in this database includes data from 1988 to 2021, as this is the last year with complete data.

Through the figure 2.5 we can see how the number of registered vulnerabilities has evolved over the years. It is important to note that there are approximately 3 periods of

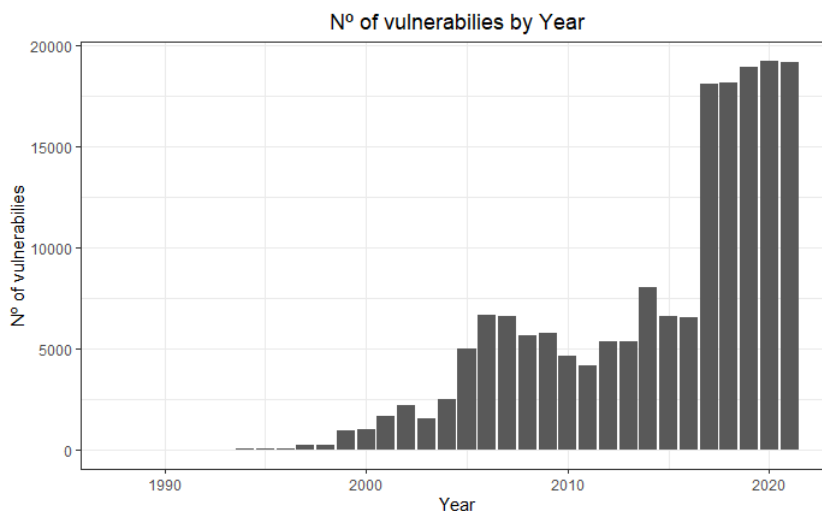


FIGURE 2.5: Number of vulnerabilities registered at NVD

distinct behavior present in the figure: 1988-2005, 2005-2016 and 2016-2021. Between each of the periods there is a significant increase between them, revealing different strategies regarding the detection, recording and prevention of vulnerabilities. In this way, the section of the different periods can be relevant for the further analysis of the data. In addition to the publication date, it is possible to visualize other factors such as integrity or access complexity over time to check its evolution overtime. In summary, it is possible to see that, despite observing a greater number of vulnerabilities, these tend to be less critical. A possible way for visualizing this behavior is through the vulnerability score, a metric that incorporates, through the formula [A.1](#), important information about a vulnerability and helps in measuring its severity. In the figure [2.6](#) we can see an increase in records of vulnerabilities with lesser severity in recent years, which may be evidence of greater attention to this area.

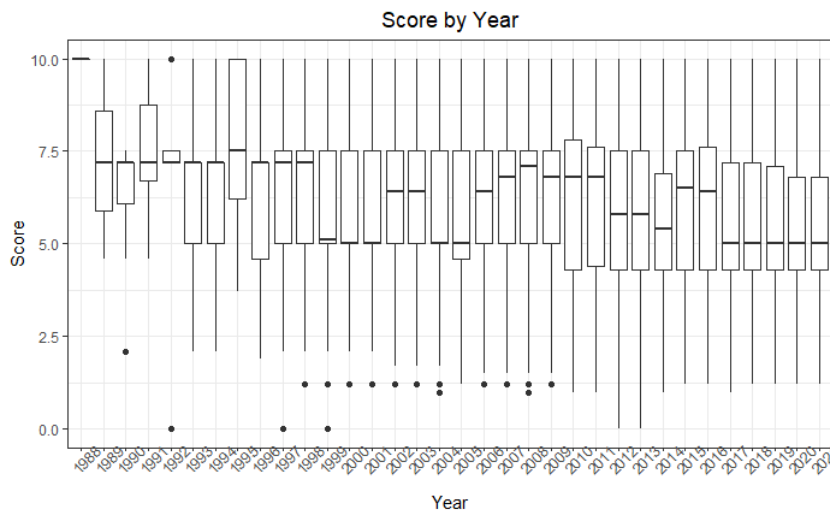


FIGURE 2.6: Scores boxplot over the years

While a global view of all vulnerabilities is excellent for a broader view of the problem, each product will have its own characteristics that may or may not match the global trend of the system. In this way, databases of singular products will also be studied to understand specific behaviors that they may have.

## 2.5 Individual data bases

### 2.5.1 Android data base

The Android product was first introduced in 2008 and, since then, it has reshaped the smartphone concept, being the most popular mobile phone operating system since 2012,

competing with Apple's iOS. With the popularity of the product, it is necessary that the security of this operating system to be as recent as possible, to avoid attacks on the sensitive information of their consumers. According to NVD, registered Android system vulnerabilities can be seen in the figure 2.7. It should be noted that, again, there are dif-

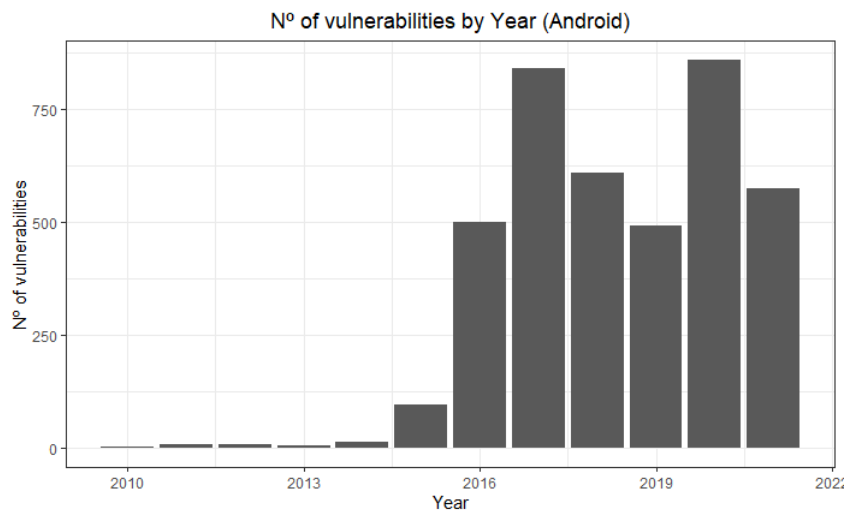


FIGURE 2.7: Number of vulnerabilities registered at Android

ferent behaviors in the number of vulnerabilities recorded over time, more specifically in the period 2010-2014 and in the period 2015-2020. During the prior period, there are very few vulnerabilities registered (in the order of the tens) while in the later period they are registered in the order of the hundreds, which justifies a separation of these periods. As the analysis with more recent data is more interesting, we will only deal with Android data in the period 2015-2020. Regarding the evolution of the score over time, it is pos-

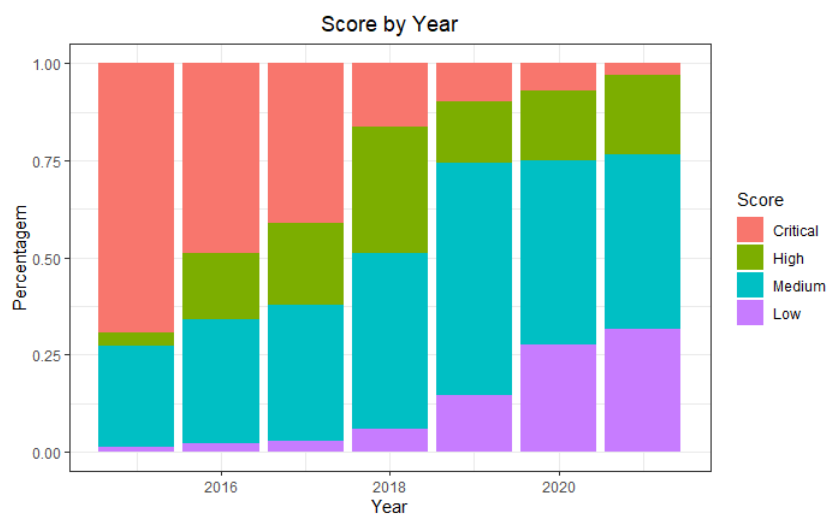


FIGURE 2.8: Score type in Android

sible to observe in the figure 2.8 that, in percentage terms, there is a notable decrease in the most critical vulnerabilities and an increase in the smallest. Such behavior suggests an improvement in the detection and prevention of vulnerabilities over the years considered. If we compare with the global trend of the data, the Android system seems to behave similarly to the rest of the products. The Android system, as a product in continuous development, needs several changes to add new content and correct previously developed errors. For this, the responsible for the product create patches, that is, periodic updates where they obtain the necessary feedback to continue to develop in the direction that the consumer prefers. Each update brings new content, and thus, new possibilities of vulnerabilities. In this way, it may be interesting to visualize the number of vulnerabilities with the information of these updates. This is represented in the figure 2.9, where we can see some increases in the release of new updates, as represented in the vertical lines. This information may be useful in modeling the data in later chapters. Another external

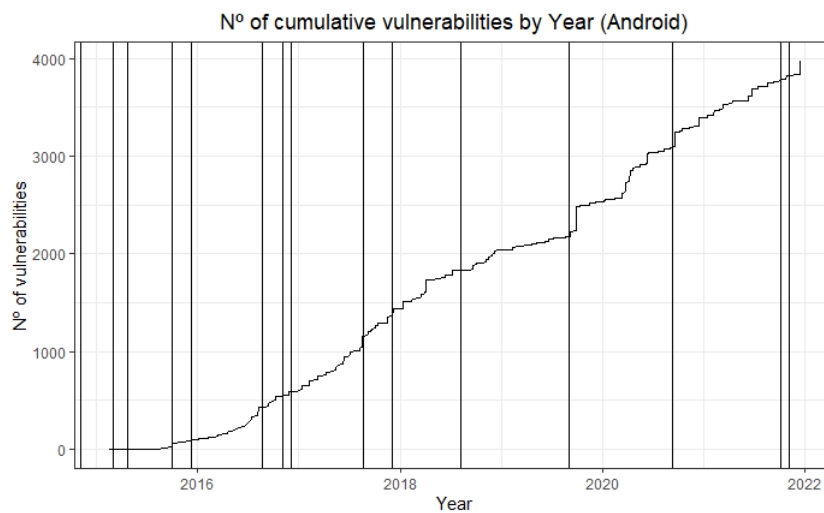


FIGURE 2.9: Cumulative vulnerabilities for Android data

factor that may influence the number of vulnerabilities is the influence of the product on the market [3]. In this sense, according to [43], Android has a variation in the influence on the mobile phone operating system market according to the figure 2.10. It should be noted that there has been an approximately linear increase from 2010 to 2017, where it has stabilized around 40 %.

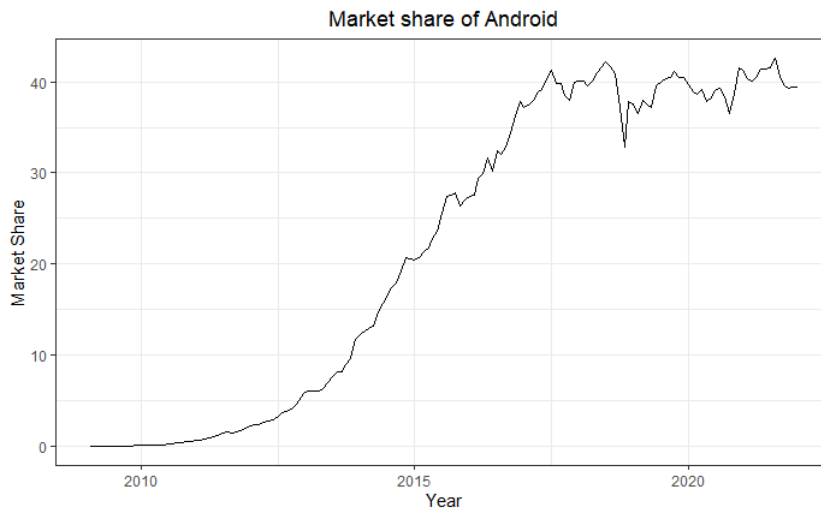


FIGURE 2.10: Market Share of Android

## 2.6 Note on modeling limitations

As mentioned in the section 2.1, there is a time difference between vulnerability discovery and registration. This topic is well known in the literature [23], however, over the years, information on this type of differences has become private, accessible only within the entity itself and vulnerability and risk analysis entities. In such a way, it will be assumed, for lack of more data, that the discovery time is equal to the registration time. This will be one of the major limitations of the analysis carried out from now on.

With the aforementioned limitations and with the help of exploratory analysis, three approaches were chosen: point processes due to the temporal counting nature of the data, classical extreme value theory for an analysis of the most critical vulnerabilities and, finally, extreme value theory but in an aspect of avoiding estimating distribution functions and using relevant statistics for the creation of models.



## Chapter 3

# Modelling through Point Processes

Among the approaches chosen, one starts with point processes. These processes are naturally applied in these situations because the vulnerability record can be represented in a timeline. The theory of point processes will be introduced, followed by some more popular models that best fit the described context. For this, [17] and [30] will be used as a basis, also using references to [9].

### 3.1 Point Processes

There are several ways to define point processes, whether informal, historical or detailed using measure theory. To avoid going too deeply into concepts of measure theory, all proofs and technical definitions will be referred to chapter 9 of [17] and to [7].

To define point processes with some rigor, let's start by defining a space  $\Omega$  with a  $\sigma$ -Borel algebra  $\mathcal{B}$  and  $\mathcal{B}_0$  the class of bounded Borel sets. Being  $\mathcal{N}$  the  $\sigma$ -algebra defined by:

$$\mathcal{N} = \sigma(\{x \in N : n(x_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0) \quad (3.1)$$

then we can define a point process by 3.1.

**Definition 3.1.** A point process  $X$  defined on  $\Omega$  is a measurable mapping defined on some probability space  $(\Omega, \mathcal{B}, \mathcal{P})$  and taking values in  $(N, \mathcal{N})$ . The distribution  $P_X$  of  $X$  is given by  $P_X(B) = P(\{\omega \in \Omega : X(\omega) \in B\})$  for  $B \in \mathcal{N}$ . We shall sometimes identify  $X$  and  $P_X$  both as a point process.

A definition that will be useful later is the measure  $\mu$ . A measure  $\mu(A)$  of a set  $A \subseteq \Omega$  can be expressed in the equation  $\mu(A) = \int_B \rho(\xi) d\xi$ , where  $\rho$  is the intensity function of the point process.

This mathematical basis allows the elaboration of point processes with several properties necessary for the construction of effective models. One of the simplest but fundamental processes as it serves as a base for the generalization to other processes is the Poisson point process.

### 3.2 Poisson point processes

The definition of the Poisson point process can be done using different tools. For this work it is only necessary an intuitive but rigorous vision to understand the fundamental properties of this process. We will closely follow the approach taken by [9]. Let's consider a counting process where we wait  $X_n$  for the  $n$ th event to occur, where  $X_n$  is a random variable in a probability space where there are no overlapping events. So we can define a strictly increasing quantity  $S_n = X_1 + .. + X_n$  that indicates the time of occurrence of an  $n$ th event. If a finite number of events occur in finite time intervals, we can say that:

$$0 = S_0(\omega) < S_1(\omega) < \dots, \sup_n S_n(\omega) = \infty \quad (3.2)$$

or equivalently ,

$$X_1(\omega) < X_2(\omega) < \dots, \sum_n X_n(\omega) = \infty \quad (3.3)$$

If conditions are met for each  $\omega$ , these conditions will be called **condition 0**. Note that we no longer impose restrictions on the variables  $X_n$ , whether they are identically distributed or independent.

Another quantity that is also important is the number of  $N_t$  events that occur up to time  $t$  defined as:

$$N_t = \max\{n : S_n \leq t\} \quad (3.4)$$

It is possible to interconnect the quantities through the expression:

$$[N_t \geq n] = [S_n \leq t] \quad (3.5)$$

From which it is deduce that

$$[N_t = n] = [S_n \leq t < S_{n+1}] \quad (3.6)$$

This particular equation is important for other developments of theories such as in the area of dynamical systems . If we assume the zero condition, it is possible to show that it is equivalent to assume the independence of exponentially distributed variables  $X_n$  with parameter  $\alpha$  and that the increments are independent and follow a Poisson distribution:

$$P[N_t - N_s = n] = e^{-\alpha(t-s)} \frac{(\alpha(t-s))^n}{n!} \quad (3.7)$$

where  $s < t$ . In fact, we can enumerate a relevant theorem to this study:

**Theorem 3.2.** *If condition zero holds and  $[N_t : t \geq 0]$  has independent increments and no fixed discontinuities, then each increment has a Poisson distribution.*

This condition is important because the zero condition and the independence of increments are sufficient for the increments to follow a Poisson distribution. It should be noted that the increments following a Poisson distribution is not enough to formalize the process, that is, the finite-dimensional distribution only defines part of a process. It is necessary to define paths (equivalent to setting  $\omega$  and varying  $t$  of the function  $N_t(\omega)$ ) to completely define the process. These paths were defined by the zero condition, hence their importance for the process construction.

### 3.3 Marked point process

A possible generalization for a Poisson process [30] defined in space  $S$  is to associate to each point of the Poisson process a random variable, called mark  $m$ , defined in space  $M$ . Thus, the new process is defined in  $S^* = \{(X, m); X \subseteq S\}$  in the space  $S \times M$ . In the specific case where the  $m$  marks belong to a space  $M = \{1, 2, \dots, N\}$ , we call it a multi-type Poisson process, a specific type of marked Poisson process. Although this generalization is not complex, it allows introducing more variety to the process as marks can be correlated with their locations or with each other. In addition, marks are able to carry information about the locations  $x_i$  that would not be possible otherwise.

The correlation between marks and location is an important factor in creating a process that correctly reflects reality. In this sense, it is possible to present two types of marks:

1. **Independent Marks:**  $N$  has independent marks if the marks  $\{m_i\}$  are mutually independent random variables such that their distribution only depends on their locations  $x_i$ .

2. **Unpredictable Marks** :  $N$  has unpredictable marks if the distribution of marks  $m_i$  is independent of location and other marks.

It is important to note that, in the construction of the marked Poisson process, the importance of respecting the restrictions of the base Poisson process, specifically independence. This generalization would not be valid if these conditions do not hold.

### 3.4 Intensity estimation

A common method for estimating the intensity of a process through its points is through likelihood maximization (MLE), ie, it seeks to maximize the likelihood function (or its logarithm) of a process. For a non-homogeneous Poisson process with intensity  $\lambda_\theta(x)$ , where  $\theta$  is the parameter to be estimated, the logarithm function of the likelihood, up to a constant, represented in 3.8:

$$\log(L_\theta) = \sum_{i=1}^n \log(\lambda_\theta(x_i)) - \int_W \lambda_\theta(u) du \quad (3.8)$$

where  $W$  is the window where the data is. Despite the generality of this function, not knowing the explicit relationship between intensity  $\lambda_\theta(x)$  and the parameter  $\theta$ , numerical maximization may not behave correctly as well as there may exist several maximums. To combat this problem it is usual to assume a loglinear relationship [7] between the intensity and the  $\theta$  parameter, especially in spatial or health problems.

In the Marked Poisson case we can divide the marked point process  $Y$  into sub processes  $X^{(1)}, \dots, X^{(M)}$  where each sub process  $X^{(m)}$  corresponds to the behavior of the  $m = 1, \dots, M$  mark. For parameter estimation it is possible to adapt the expression 3.8 to the marked process  $y = \{(u_1, m_1), \dots, (u_n, m_n)\}$ , getting the expression 3.9, up to a constant:

$$\log(L_\theta) = \sum_{i=1}^n \log(\lambda_\theta(u_i, m_i)) - \sum_{m \in M} \int_W \lambda_\theta(u, m) du \quad (3.9)$$

### 3.5 Model diagnostics

In order to understand the adequacy of the models for the behavior exhibited by the data, several techniques, tests and metrics are used to evaluate the quantitative and graphic evaluation of the models' goodness of fit. In this dissertation, in addition to its applications, each technique will be described and the conditions of use will be identified.

### 3.5.1 Probability plots

While quantitative diagnostic tests are more specific, graphical tests allow to visualize how and where data is being best modeled. The graphical methods used in this dissertation are the quantile plots (QQ-plot). If we order the observations of a sample of the uniform distribution in the form  $U_{(1)} < \dots < U_{(n)}$ , then we have that [40]:

$$\mathbb{E}(X_j) = \frac{j}{n+1} \quad (3.10)$$

With this information, considering the sample  $X_{(n)}$  from a distribution  $F$ , we can represent the ordered statistic  $X_{(k)}$  against  $F^{-1}(\frac{k}{n+1})$ , where we should get a direct proportional relationship, independent of  $F$ . In fact,  $F^{-1}(\frac{k}{n+1})$  is actually the quantile  $\frac{k}{n+1}$  of the  $F$  function, which can then be represented in a QQ-plot quantile graph.

### 3.5.2 Chi-square test

The Chi-square test is a test widely used in its application as a measure of the success of model adjustment [40]. Its formalism in the application of point Poisson processes consists in testing the null hypothesis that the intensity is of the form  $\lambda_\theta(t)$  for certain values of  $\theta$ . For this, the space is divided into equal  $B_j$  quadrants with  $n_j$  points in each of these spaces. These points will be realizations of the Poisson process with average  $\mu_j$ . This average can be estimated using the estimated intensity of the model:

$$\hat{\mu}_j = \int_{B_j} \hat{\lambda}_\theta(t) dt \quad (3.11)$$

Then we can calculate the test statistic:

$$X^2 = \sum_j \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_j \frac{(n_j - \hat{\mu}_j)^2}{\hat{\mu}_j} \quad (3.12)$$

Associating the statistic  $X^2$  with the distribution  $\chi^2$  with  $m - p$  degrees of freedom, where  $m$  is the number of quadrants and  $p$  the number of parameters in the model. The null hypothesis  $H_0$  proposes that the model adequately fits the data, implying that the number of points in each quadrant should be close to the expected number. Although the division of the quadrants may be arbitrary, it is empirically recommended that there is no data cell containing less than 5 observations, as this may impact the value of the statistic. Equivalently, it is possible to calculate a p-value for the statistic, which is then compared with the previously imposed level of confidence.

### 3.5.3 Anderson-Darling test

The Anderson-Darling test is a non parametric test that was originally developed to measure the deviation of a sample from normality. This test is usually considered as an alternative to the Kolmogorov-Smirnoff test, however different studies [19] show that Anderson-Darling is more sensitive and stable for samples with different tail behaviors. Darling and Pettit [6] generalized the test to compare two samples  $\{X_1, \dots, X_n\}$  and  $\{Y_1, \dots, Y_m\}$  that come from distributions  $F$  and  $G$ , with empirical distribution function  $F_n$  and  $G_m$ , respectively. The null hypothesis of the test becomes that the distributions  $F$  and  $G$  are identical, thus obtaining the equation 3.13 :

$$AD = \frac{nm}{N} \int_{-\infty}^{\infty} \frac{\{F_n(x) - G_m(x)\}^2}{H_N(x) \{1 - H_N(x)\}} dH_N(x) \quad (3.13)$$

where  $N = n + m$  and  $H_N(x)$  is the distribution function of the combined samples  $H_N(x) = \{nF_n(x) + mG_m(x)\} / N$ . It is possible to simplify the expression 3.13 to the form 3.14:

$$AD = \frac{1}{mn} \sum_{i=1}^{n+m} (N_i Z_{(n+m-ni)})^2 \frac{1}{i Z_{(n+m-i)}} \quad (3.14)$$

where  $Z_{(n+m)}$  represents the ordered statistic of combined sample of the initial samples  $\{X_1, \dots, X_n\}$  and  $\{Y_1, \dots, Y_m\}$  of size  $n$  and  $m$ , respectively. Furthermore  $N_i$  represents the number of observations of  $X_n$  that are less than or equal to the  $i$ th observation of  $Z_{(n+m)}$ . If the calculated statistic is greater than the critical value, defined in [6] and dependent on the significance level  $\alpha$ , then the null hypothesis that the two distributions are equal is rejected. This is going to be one of the main statistics to be used to compare samples from the model and the data.

### 3.5.4 Wald test

Even if an intensity is estimated, it is necessary to understand if the parameters that compose it are relevant to the construction of the model. In this way we can submit the parameters to Wald's tests. This test has the null hypothesis that each individual parameter  $\phi$  is zero, that is, we are testing whether this parameter is statistically different from zero. The statistic of this model can be represented in the form:

$$\sqrt{W} = z = \frac{\hat{\phi}}{se(\hat{\phi})} \quad (3.15)$$

where  $se(\hat{\phi})$  is the standard deviation calculated using MLE. This test is also called the Z test because the statistic usually follows a normal distribution asymptotically. When the statistic is greater than the critical value established by the confidence level  $1 - \alpha$ , we can then reject the null hypothesis, thus stating that the parameter  $\phi$  is non-zero statistically.

### 3.5.5 Implementation

Regarding data analysis, the data will be modeled according to a simple (non-homogeneous) Poisson process and a marked Poisson process, where the marks will be the score associated with each vulnerability. Previous studies [42] point to different behaviors depending on the severity recorded, suggesting the best performance of the marked Poisson model.

Before the modeling itself, it is necessary to verify to what extent the data follows the necessary requirements of the applied models. For this, it will be checked if there are records of simultaneous vulnerabilities as well as the independence of the inter arrival times. If there are simultaneous records of vulnerabilities, a possible approach would be to consider these overlaps as clusters and consider a cluster or cox process, but this will not be explored in this work (although a solution to the clustering problem is found using another perspective proposed in chapter 4). If this problem exists, the model is reformulated to model the days when vulnerabilities occurred not the individuals instances of vulnerabilities, solving the overlapping problem. As there may be days where vulnerabilities of different severity are recorded, the points are shifted from their original position by an insignificant amount (about 0.001 of a day). On the other hand, the independence of the inter arrival times is not possible to solve without changing the nature of the data, therefore the only thing that can be done is to simply model the dependence and discard the Poisson models.

After verifying that the data follow the requirements, the intensities are estimated using the MLE method and all parameters are submitted to the Wald test at 5% significance level. The parameters accepted in the models are those that are statistically different from zero. As observed in the 3.4 intensities estimation chapter, the choice of the observation window is a crucial factor for a correct estimation, because in addition to using information from where there are points, it also uses where they are not. For that, the chosen window must have as little non-relevant space as possible. One of the ways to guarantee this condition is for the window to have the same dimension as the data plus one day in the time component ( $x$ ), the choice of dimension  $y$  is not relevant in this particular study.

A particularity of stochastic models like the ones presented is the variability that exists in the creation of a model instance, that is, simulating a model with the same intensity parameters results in different realizations. Therefore, all diagnostic tests applied to a model will be slightly different due to different realizations being performed.

The software used for data modeling was the R [37], using the packages mentioned in [44], [48], [24], [47], [39], [18] and [8]. In particular, the most used package in this section was *spatstat* [8].

## 3.6 Statistical Analysis

### 3.6.1 Investigating Android Data

The first step to be considered is studying the independence of the inter arrival times. This is equivalent to analyzing the dependence for the differentiated data, ie the differences between data points, verifying with the graphs of ACF (autocorrelation function) and PACF (partial autocorrelation function) represented in the figures 3.1a and 3.1b, respectively. As

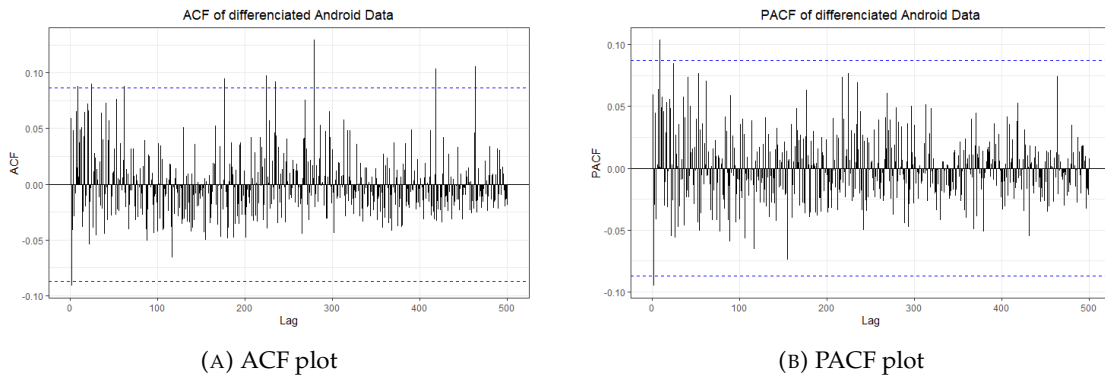


FIGURE 3.1: Correlation plots

seen in the 3.1 plot, there are some dependency peaks for certain lags in both graphs, but they are not in significant numbers to talk about a dependency relationship between arrival times. Thus, we can assume that the data respect the conditions required by the models.

Under these conditions, the intensities were estimated for the simple Poisson model, marked Poisson model and non-linear marked Poisson model. Graphical comparisons such as QQ-plot comparisons are represented in B. Statistical tests are represented in table 3.1.



Table of test statistics (p-value)					
Model	Chi-Squared	A.D - Critical	A.D - High	A.D - Medium	A.D - Low
Simple	2.2e-16	0.06073	0.2383	0.4209	0.04967
Marked	2.2e-16	0.1487	0.3456	0.2257	0.2702
Non-linear Marked	2.2e-16	0.5272	0.4572	0.413	0.5093

TABLE 3.1: Table of test statistics

The output of the simple Poisson model is reproduced in the figure 3.2, and the equation 3.16 can be constructed based on these parameters:

	Estimate <dbl>	S.E. <dbl>	CI95.lo <dbl>	CI95.hi <dbl>	Ztest <fctr>	Zval <dbl>
(Intercept)	-1.176004e+01	1.060344e-01	-1.196786e+01	-1.155222e+01	***	-110.907792
x	4.137156e-04	7.359279e-05	2.694764e-04	5.579548e-04	***	5.621687

FIGURE 3.2: Output of the simple Poisson model

$$\hat{\lambda}_s = \exp(-11.76 + 4.1372 \cdot 10^{-4}t) \quad (3.16)$$

where  $\hat{\lambda}_s$  is the intensity of the simple Poisson model and  $t$  is the number of days since the first vulnerability recorded. Although the model is non-homogeneous, the value of the time-dependent component is an order of magnitude smaller than the independent component, leading to an almost homogeneous intensity. As a simple Poisson model, it is assumed that the behavior is the same for all severities, with no distinction between them. For model diagnoses, the p-values of the Anderson-Darling test are close to the threshold established by the significance level. It should be noted that there are orders of magnitude differences between the different severities.

Adding the formalism of marks to the simple Poisson model, we obtain the output in the figure 3.3, being able to represent the model in the equation 3.17:

	Estimate <dbl>	S.E. <dbl>	CI95.lo <dbl>	CI95.hi <dbl>	Ztest <fctr>	Zval <dbl>
(Intercept)	-1.141059e+01	0.2027508078	-1.180797e+01	-1.101320e+01	***	-56.2788656
marksHigh	-1.554839e-01	0.2810619601	-7.063552e-01	3.953875e-01		-0.5532014
marksLow	-1.477856e+00	0.3682727819	-2.199658e+00	-7.560549e-01	***	-4.0129393
marksMedium	-1.422593e-01	0.2771924896	-6.855466e-01	4.010280e-01		-0.5132150
x	-2.487895e-05	0.0001575368	-3.336453e-04	2.838874e-04		-0.1579247
marksHigh:x	4.025448e-04	0.0002082735	-5.663873e-06	8.107534e-04		1.9327696
marksLow:x	1.003792e-03	0.0002474347	5.188294e-04	1.488756e-03	***	4.0567975
marksMedium:x	4.749220e-04	0.0002043140	7.447392e-05	8.753702e-04	*	2.3244711

FIGURE 3.3: Output of the marked Poisson model

$$\hat{\lambda}_s = \exp(-11.41 - 1.4778M_l - 0.14226M_m - 2.488 \cdot 10^{-5}t + 1.003 \cdot 10^{-3}M_l t + 4.749 \cdot 10^{-4}M_m t) \quad (3.17)$$

where  $M_l$  and  $M_m$  is an indicator function that is one when the observed severity is low and medium, respectively. It should be noted that the terms  $t$  and  $M_m$  are not significant, but the interaction between them is, so they will be necessary for the construction of a coherent model. This model has critical severity as a benchmark, that is, the model is based on observations with critical severity and estimates the difference in behavior for the other severities. This is relevant because, for example, the term that corresponds to the differences between critical and high severities is not significant, that is, the Wald test does not reject the hypothesis that this term is different from zero. This means that there is no significant evidence for different behaviors between critical and high severities. The marks in this model are independent as they depend on the temporal quantity.

When we talk about statistical tests, there is an improvement over the previous model. In addition the marked model obtained more uniform statistics for the various categories, on average, closer to 1 than the simple model.

The evolution of data does not simply follow a linear trend, more particularly when there are global advances in understanding and detecting vulnerabilities. This can be incorporated, for example, by adding a non-linear component  $t^2$ . The output of the non-linear marked Poisson model is represented in the figure 3.4, building from this the equation 3.18:

	Estimate <dbl>	S.E. <dbl>	CI95.lo <dbl>	CI95.hi <dbl>	Ztest <fctr>	Zval <dbl>
(Intercept)	-1.196479e+01	2.552130e-01	-1.246500e+01	-1.146458e+01	***	-46.8815893
marksHigh	-2.727354e-01	3.077782e-01	-8.759696e-01	3.304988e-01		-0.8861426
marksLow	-1.799574e+00	4.116946e-01	-2.606481e+00	-9.926677e-01	***	-4.3711388
marksMedium	-2.823333e-01	3.042936e-01	-8.787379e-01	3.140712e-01		-0.9278319
x	1.396095e-03	3.613457e-04	6.878704e-04	2.104320e-03	***	3.8635989
I(x^2)	-6.356697e-07	1.413341e-07	-9.126793e-07	-3.586600e-07	***	-4.4976395
marksHigh:x	5.005715e-04	2.335449e-04	4.283202e-05	9.583110e-04	*	2.1433635
marksLow:x	1.248392e-03	2.825037e-04	6.946946e-04	1.802089e-03	***	4.4190277
marksMedium:x	5.905893e-04	2.296684e-04	1.404475e-04	1.040731e-03	*	2.5714868

FIGURE 3.4: Output of the non-linear marked Poisson model

$$\hat{\lambda}_s = \exp(-11.96 - 1.800M_l - 0.2823M_m - 0.2727M_h + 1.3960 \cdot 10^{-3}t - 6.3567 \cdot 10^{-7}t^2 + 1.248 \cdot 10^{-3}M_l t + 5.905 \cdot 10^{-4}M_m t + 5.0057 \cdot 10^{-4}M_h t) \quad (3.18)$$

In this model, the indicator functions  $M_l$  and  $M_h$  are not significant, but they were added because their interactions with time are significant. There is an order of magnitude disparity between parameters where the time component is used because  $t$  is measured in days. The marks in this model, like the previous one, are independent. Regarding the statistical tests, this model is the one that shows the best results, obtaining values close to 0.5 in the Anderson-Darling test. This indicates that there is more difficulty in distinguishing the non-linear marked Poisson model from data than the rest.

Comparing the different graphical diagnoses presented in [B](#) for the three proposed models, they corroborate the quantitative statistics. More specifically, the QQ-plots of the marked non-linear model are much closer to the bisector of the odd quadrants than the other models, with the simple model being the one that departs the most, especially in the tails of the distributions.

### 3.6.2 Results

Poisson models are simple statistical models that allow to represent counting events quite accurately. In this context, we apply three types of progressively more complex Poisson models: a simple model, a marked model and yet another marked non-linear model. After verifying that the data can be assumed to satisfy the model constraints, the intensities for the different models were estimated. Both graphical and quantitative diagnoses suggest that the marked non-linear model is better suited to model these data.



## Chapter 4

# Modelling through Classic Extreme Value Theory

Although it is possible to model the data as a whole, there are cases in which it will be more interesting to analyze subsections of it. More specifically, in our particular case, there is an intrinsic interest in the study and modeling of days when an abnormal number of vulnerabilities are observed or when a vulnerability with critical impacts for the entity is recorded. Generally, this type of observations may not follow the distribution of the other data, thus requiring special treatment. In this way, we will apply extreme value theory in this section. First, the extreme value theory in the context of independent variables will be introduced. Next, the adaptation of the theory to stationary series and correlated data will be presented as well as the models built on this basis. The theory section closely follows [12] and [33], while the modeling part follows [28].

### 4.1 Extreme value theory (EVT) for independent variables

The extreme value theory is based on the study of asymptotic distributions of maxima and/or minima of a set of observations. This statistic, formally defined as  $M_n$ , can be written as:

$$M_n = \max(X_1, X_2, \dots, X_n) \quad (4.1)$$

where  $X_n$  are random variables iid following an  $F$  distribution. The distribution of this statistic can be derived for all  $n$ :

$$P \{M_n \leq z\} = Pr \{X_1 \leq z, X_2 \leq z, \dots, X_n \leq z\} = F^n(z) \quad (4.2)$$

where we intend to study this distribution when  $n \mapsto \infty$ . Normally, to avoid degeneracies, the distribution of the normalized statistic  $M_n^* = \frac{M_n - b_n}{a_n}$  is studied, where the appropriate choice of  $a_n$  and  $b_n$  will stabilize this statistic at the desired threshold.

One of the most important results in this theory was introduced by [22] and generally proved by [25] where it is shown that when there is convergence of the distribution of the statistic  $M_n^*$ , it will be for one of the families of generalized extreme values (GEV). More formally, the theorem can be written as:

**Theorem 4.1** (Theorem of Generalized Extreme value distribution). *If there is a set of constants  $\{a_n > 0\}$  and  $\{b_n\}$  such that*

$$P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \rightarrow G(z) \quad (4.3)$$

*when  $n \rightarrow \infty$  for a non-degenerative distribution function  $G$ , then  $G$  is a member of the GEV family,*

$$G(z) = \exp \left\{ \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \quad (4.4)$$

*defined on  $\left\{ z : 1 + \xi \left( \frac{z - \mu}{\sigma} \right) > 0 \right\}$  where  $-\infty < \mu < \infty, \sigma > 0$  and  $-\infty < \xi < \infty$*

This model has three parameters: a location parameter  $\mu$ , a scale parameter  $\sigma$  and a shape parameter  $\xi$ . The three families, known as those of Gumbel, Fréchet and Weibull, are specific cases of values for these parameters, more specifically they correspond to the cases  $\xi = 0$ ,  $\xi > 0$  and  $\xi < 0$ , respectively. The estimation of parameters, through maximum likelihood for example, allows to know which family and tail behavior best adapts to the data. It should be noted that the estimation of the sets of constants  $a_n$  and  $b_n$  is not relevant because, for large  $n$ ,  $G\left(\frac{z - b_n}{a_n}\right) = G^*(z)$ , that is, if the convergence is verified for the statistic  $M_n^*$ , then it also exists for  $M_n$  only with different parameters of location and scale, making the problem practically the same when it comes to parameter estimation.

An intuitive way to create a model would be to simply group the data into sequences with  $n$  observations and choose only their maximum  $M_n$ . To estimate the extreme quantiles, it would be enough to invert the equation 4.4:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{-\log(1 - p)\}^{-\xi}], & \text{for } \xi \neq 0, \\ \mu - \sigma \log \{-\log(1 - p)\}, & \text{for } \xi = 0, \end{cases} \quad (4.5)$$

where  $G(z_p) = 1 - p$ . Usually  $z_p$  is designated as the return level associated with the return period  $1/p$ , that is, the level  $z_p$  is expected to be exceeded, on average, once for every  $1/p$  time levels. Return times are of particular importance in this dissertation and will be discussed in more detail in the chapter 5.

#### 4.1.1 Convergence analysis

The study of extreme values is only possible if the probability distribution convergence  $P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} = P \{M_n \leq u_n\}$ , where  $u_n = z \cdot a_n + b_n$  is verified. Thus, it will be interesting to study under what conditions this occurs. An essential theorem for the study of convergence, especially in the dependence situation, can be summarized in the theorem 4.2:

**Theorem 4.2.** *Let  $\{X_1, X_2, \dots, X_n\}$  be random iid variavels with a commum distribution function  $F$ . Chossing  $-\infty < \tau < \infty$  and let  $\{u_n\}$  be a sequence of number such that :*

$$n(1 - F(u_n)) \rightarrow \tau \text{ as } n \rightarrow \infty \quad (4.6)$$

then

$$P \{M_n \leq u_n\} \rightarrow e^{-\tau} \text{ as } n \rightarrow \infty \quad (4.7)$$

Using the notation  $x_F = \sup \{x; F(x) < 1\}$ , we can deduce a corollary of the theorem 4.2

**Corollary 4.3.** •  $M_n \rightarrow x_F (\leq \infty)$  with probability one as  $n \rightarrow \infty$

- If  $x_F < \infty$  and  $F(x_{F-}) < 1$  and if for a sequence  $\{u_n\}$ ,  $P \{M_n \leq u_n\} \rightarrow \rho$  as  $n \rightarrow \infty$ , then  $\rho = 1$  or  $0$

With this theorem and corollary we can see that bounded and discrete distributions will not satisfy the above conditions, converging to a degenerate function. This is why GEVs cannot be applied directly to our discrete data and other methods will be explored in the section 5. However, other more complex formulations of the extreme value theory may be useful.

## 4.2 EVT for dependent variables

Although the classical extreme value theory encompasses many real and interesting cases, the fact that the variables are iid is a very strong assumption for temporal data. In this way, the incorporation of some kind of data dependency will be necessary for the creation of more realistic models. It is possible to build this theory in more detail, but for reasons of brevity, more technical details can be consulted in [33].

### 4.2.1 Maximum stationary series

Although there are numerous ways to incorporate dependency into the data, one of the most popular and simple ways will be to create a stationary process, that is, a process in which the marginal distribution of observations does not vary with time and for which the observations become almost independent the farther apart they are. This can be represented by the condition  $D(u_n)$ , formalized in the definition 4.4 :

**Definition 4.4.** A stationary series  $X_1, X_2, \dots, X_n$  is said to satisfy the  $D(u_n)$  condition if, for all  $i_1 < \dots < i_p < j_1$  with  $j_i - j_p > l$

$$\left| \Pr \left\{ X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n \right\} - \Pr \left\{ X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n \right\} \Pr \left\{ X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n \right\} \right| \leq \alpha(n, l) \quad (4.8)$$

where  $\alpha(n, l) \rightarrow 0$  for some sequence  $l_n$  such that  $l_n/n \rightarrow 0$  as  $n \rightarrow \infty$ .

With this condition, we have a theorem similar to the theorem of the generalized distribution of extreme values for stationary series:

**Theorem 4.5.** Let  $\{X_n\}$  be a stationary sequence and  $\{a_n\}$  and  $\{b_n\}$  given constants such that  $P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\}$  converges to a non-degenerate distribution function  $G(z)$ . Suppose that  $D(u_n)$  is satisfied for  $u_n = z/a_n + b_n$  for which real  $z$ . Then  $G(z)$  is part of the GEV family presented in theorem 4.4

As the only difference between the theorem 4.5 and 4.1 is the condition  $D(u_n)$ , it will be interesting to investigate the relationship between a sequence of iid variables with maximums  $M_n$  and a stationary process with the same marginal distribution with maxima  $M_n^*$ . Based on the theorem 4.2, it is possible to prove the theorem 4.6 :

**Theorem 4.6.** Let  $\{X_n^*\}$  be a stationary process and  $\{X_n\}$  random iid variables with the same marginal distribution. Defining  $M_n = \max \{X_n\}$  and  $M_n^* = \max \{X_n^*\}$ . Under conditions of



convergence, i.e. satisfying condition  $D(u_n)$

$$P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \rightarrow G_2(z) \quad (4.9)$$

as  $n \rightarrow \infty$  for normalizing sequences  $\{a_n > 0\}$  and  $\{b_n\}$  where  $G_2$  is a non-generate function if and only if

$$P \left\{ \frac{M_n^* - b_n}{a_n} \leq z \right\} \rightarrow G_1(z) \quad (4.10)$$

where  $G_2(z) = G_1^\theta(z)$  for a constant  $\theta$  such that  $0 < \theta \leq 1$ .

This theorem implies that, if the dependent series converges, it is related to the limit distribution of the independent series through the parameter  $\theta$ , normally called extreme index. Substituting the explicit expression 4.4 in the theorem 4.6, we verify that the function  $G_1(\xi, \mu, \sigma)$  differs from the function  $G_2(\xi, \mu^*, \sigma^*)$  only in the location and scale parameters, where  $\mu^* = \mu - \frac{\sigma}{\xi}(1 - \theta^{-\xi})$  and  $\sigma^* = \sigma\theta^\xi$ . In fact, the introduction of dependency between the data does not change the behavior of the tail of the distribution, it simply introduces the formation of clusters. Another useful way of thinking about the extremal index is to define it as  $\theta = \{\text{limiting mean cluster size}\}^{-1}$ , that is, the extreme index is the inverse of the average size of observed clusters. Consequently, the estimation of this parameter may indicate whether the stationary series tends to form clusters at high thresholds, which is relevant for the correct modeling of the data.

#### 4.2.2 Convergence analysis

As in the iid case, the convergence analysis of the distributions is essential to verify under which conditions we can apply the previously announced theorems. To maintain the condition given in the 4.2 theorem, it is necessary to add a new constraint to maintain the upper bound of the limit:

**Definition 4.7.** The condition  $D'(u_n)$  will be said to hold for the stationary sequence  $\{X_n\}$  and sequence  $\{u_n\}$  of constants if

$$\lim_{n \rightarrow \infty} \sup n \sum_{j=2}^{\lfloor n/k \rfloor} P \left\{ X_1 > u_n, X_j > u_n \right\} \rightarrow 0 \text{ as } k \rightarrow \infty \quad (4.11)$$

where  $\lfloor \cdot \rfloor$  denotes the integer part

This new constraint ensures that there will not be multiple observations in a point process of exceedances, a necessary condition for the construction of models such as the Poisson process.

### 4.3 Modelling using stationary series

Based on the theory and theorems of extreme value theory, it is possible to build several models that offer different perspectives on how to treat observations. In this way, we will explore three models: the block maximum model, the peak over threshold model and the peak over threshold models through the lens of point processes.

#### 4.3.1 Block Maximum model

The most intentional model based on the GEV theory will be the block maxima model. This model consists of dividing the data into blocks of equal length and modeling the maximums of these blocks with the GEV distributions. Choosing  $m$  blocks of size  $n$  is crucial for a good model. In general, choosing a large  $n$  will lead to a better approximation of the GEV distribution and little bias in the estimation of parameters, while a high  $m$  brings more data to the estimation, leading to a smaller variance in the estimation of the parameters. As previously seen in the theory, there is no difference in the GEV family between the iid and dependent case, leading only to different estimations of the location and scale parameters. However, the convergence rate for the GEV distribution is smaller in the dependent case,  $n\theta$ , than in the iid case,  $n$ , leading to a less accurate approximation.

#### 4.3.2 Peak over threshold - POT model

Despite its simple conditions, the block maxima model condenses the observations into their maxima, rendering the rest of the observations unusable, if they exist. To combat this problem and use all available observations, another approach requiring the use of thresholds can be used. Let's start by analyzing the iid model and then generalize to stationary series.

##### 4.3.2.1 Iid case

Similar to GEV theory, we define a set  $\{X_n\}$  of iid variables with a common marginal distribution  $F$ . Choosing an arbitrary high threshold  $u$  and its exceedance  $y$ , we can describe the stochastic behavior of extreme events in the conditional form:

$$P\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, y > 0 \quad (4.12)$$

Taking advantage of the result given by the theorem 4.4, we can apply the conditional to obtain the theorem 4.8

**Theorem 4.8.** *According to the theorem 4.4, where for some large  $n$   $P\{M_n < z\} \approx G(z)$  where  $G(z)$  is a member of the family GEV given by 4.4. Then, for large enough  $u$ , the distribution function of  $(X - u)$ , conditional on  $X > u$  is:*

$$P\{X \leq u + y | X > u\} \approx H(y) = 1 - \left(1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right)^{-1/\xi} \quad (4.13)$$

defined on  $\left\{y : y > 0 \text{ and } 1 + \frac{\xi y}{\sigma + \xi(u - \mu)} > 0\right\}$

The family of functions given by the equation 4.13 is called **Generalized Pareto Distributions** (GPD). By the theorem 4.8 we can assume that if the maximums per blocks are approximated by the GEV, then the threshold exceedances followed a GPD. Furthermore, the GPD parameters are uniquely determined by the GEV, especially the tail parameter  $\xi$ , which is the same for both distributions. The difference lies in the variability of the parameters: a choice of block sizes  $n$  will change the parameters of the GEV, but not those of the GPD, since  $\xi$  is invariant to the block size and the scale and location parameters will offset each other .

#### 4.3.2.2 Dependent case

As seen in the theory in 4.2, the difference between the iid and dependent case is in the tendency of excess clusters when we introduce an extremal index  $\theta < 1$ . This suggests changes in the modeling of the GPD, as it needs the excesses to be independent. One of the most used methods to get around the problem is *Declustering*. This method consists of defining clusters of exceedances and identifying their maximums. So we can assume that these will be independent, being able to model with the GPD. This method has its limitations, namely being quite dependent on the criterion of cluster formation and the loss of information in only considering the maximums of the clusters.

## 4.4 POT model - Point process perspective

Another different way of looking at modeling is to consider threshold exceedances as an event in time and use a point process for modeling the occurrence of these events. The basic theory of point processes has already been described in the chapter 3, so it will be

introduced promptly. The point peak over threshold (POT) model can be summarized by the theorem

**Theorem 4.9.** *Let  $\{X_n\}$  be a set of iid random variavels for which there are sequences of constants  $\{a_n > 0\}$  and  $\{b_n\}$  such that*

$$P \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \rightarrow G(z) \quad (4.14)$$

where

$$G(z) = \exp \left\{ \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \quad (4.15)$$

and let  $z_-$  and  $z_+$  be the lower and upper endpoints of  $G$ , respectively. Then, the sequence of point processes defined in  $\mathbb{R}^2$

$$N_n = \left\{ (i/(n+1), (X_i - b_n)/a_n) : i = 1, \dots, n \right\} \quad (4.16)$$

converges on regions  $A$  of the form  $(0, 1) \times [u, \infty)$  for any  $u > z_-$ , to a Poisson process with intensity meausre on  $(t_1, t_2) \times [z, z_+)$  given by

$$\Lambda(A) = (t_2 - t_1) \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \quad (4.17)$$

The proof of this theorem can be seen in detail in [33] or more briefly in [12]. It is important to note that the great advantage of this model lies in the independence of the process parameters with the chosen threshold. This offers stability and flexibility to the model that was not offered in previous models. In fact, it is possible to prove that the block maxima and threshold exceedance model can be enclosed in the POT model, showing its generality. The block maxima model is the particular case where  $N_n(A_z) = 0$  and the threshold exceedance model is the conditional construction of the POT model. Although the model is general, there are still assumptions in its construction that have to be respected in order to ensure that it is a good choice for modeling real data. These assumptions can be summarized in the following items:

1. The exceedances time follows a homogeneous Poisson process
2. Exceedances are idd and are independent of the time of exceedance
3. Exceedances follow a generalized Pareto distribution

Although this model is successful in several situations, the correlations that exist between points break the prepositions of the model. In this way, the POT model cannot be

applied directly in our case. A possible solution is to decluster the series, as suggested in the threshold excess model. The creation of clusters is a ad-hoc, depending on the conditions of the series itself. In this work we will follow the "runs" method. In this method, the cluster is started with the first observation that exceeds the pre-established threshold. The cluster is terminated only at the next occurrence of the threshold exceedance, except when the distance between exceedances is smaller than a pre-set distance  $r$ .

## 4.5 Parameter estimation

The estimation of model parameters is an essential step in the modeling process itself. Even if the model is adequate, the incorrect estimation of the parameters can mean the failure of a particular model. In this way, the choice of the estimation method is as or more important than the choice of the model itself. In the classical extreme value theory there are several methods for estimation, each with its advantages and disadvantages. In this work, the maximum likelihood method will be used. Although this method is quite popular, there are specific limitations for GEV families since the cut-off points of the distributions depend on its parameters. Thus, in [12] concludes that:

1. When the shape parameter  $\xi > -0.5$ , the maximum likelihood estimator is regular
2. When the shape parameter  $-1 < \xi < -0.5$ , the maximum likelihood estimator is obtainable, but may not have regular properties
3. When the shape parameter  $\xi < -1$ , the maximum likelihood estimator is probably not obtainable

With these restrictions in mind, the maximum likelihood logarithm function, for GEV families with  $\xi \neq 0$  with a sample  $z_1, \dots, z_m$ , is expressed in the equation 4.18

$$l(\mu, \sigma, \xi) = -m \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^m \log\left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^m \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi} \quad (4.18)$$

and for  $\xi = 0$ ,

$$l(\mu, \sigma) = -m \log(\sigma) - \sum_{i=1}^m \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^m \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\} \quad (4.19)$$

For models that require GP, such as the POT model, the sampled logarithmic function  $\{y_1, \dots, y_k\}$  of  $k$  exceeds a threshold  $u$  will be expressed in the form 4.20:

$$l(\sigma, \xi) = -k \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^k \log\left(1 + \frac{y_i \xi}{\sigma}\right) \quad (4.20)$$

or,

$$l(\sigma, \xi) = -k \log(\sigma) - (\sigma^{-1}) \sum_{i=1}^k \log(y_i) \quad (4.21)$$

where  $\xi \neq 0$  and  $\xi = 0$ , respectively

## 4.6 Modeling the Android data

In this work, the POT model will be used since it encapsulates all the other models previously explored. The first step in building the POT model is choosing a suitable threshold. In the case study, due to the classification presented above, the limits are already predefined. As the modulation of the most impacting cases are the ones of greatest interest, the initial threshold chosen is 9 on the 0-10 scale of the vulnerability score, that is, the most critical vulnerabilities. An extremal index estimated was 0.2764, indicating that there is a cluster process in this data. After applying the "runs" method of declustering to the series of vulnerabilities, we chose only the maximum of these clusters of observations, obtaining 88 extreme points. The clusters, represented by black bars, are represented in the figure. 4.1

The objective of this declustering would be to break the dependencies that exist in the occurrence of vulnerability clusters to obtain a homogeneous Poisson process with the maximum of each cluster. This can be confirmed by looking at the figure 4.2, where both ACF and PACF show no correlation between interarrival periods. This makes it possible to apply the Poisson model to the time of exceedances. Similar to the chapter 3, we can estimate the intensity of the model and verify its adequacy through graphic and quantitative diagnoses.

Both diagnostic graphs 4.3 and 4.4 present a satisfactory model, with the Anderson-Darling test obtaining a p-value of 0.34, thus corroborating the visual results.

To complete the POT modelling, it is still necessary to model the exceedances themselves. There is, however, the problem that the information about the exceedances, that is, the severity, is represented on a limited scale bounded by the value 10. This will naturally lead to unexpected behavior for values of severity at the upper end, precisely where

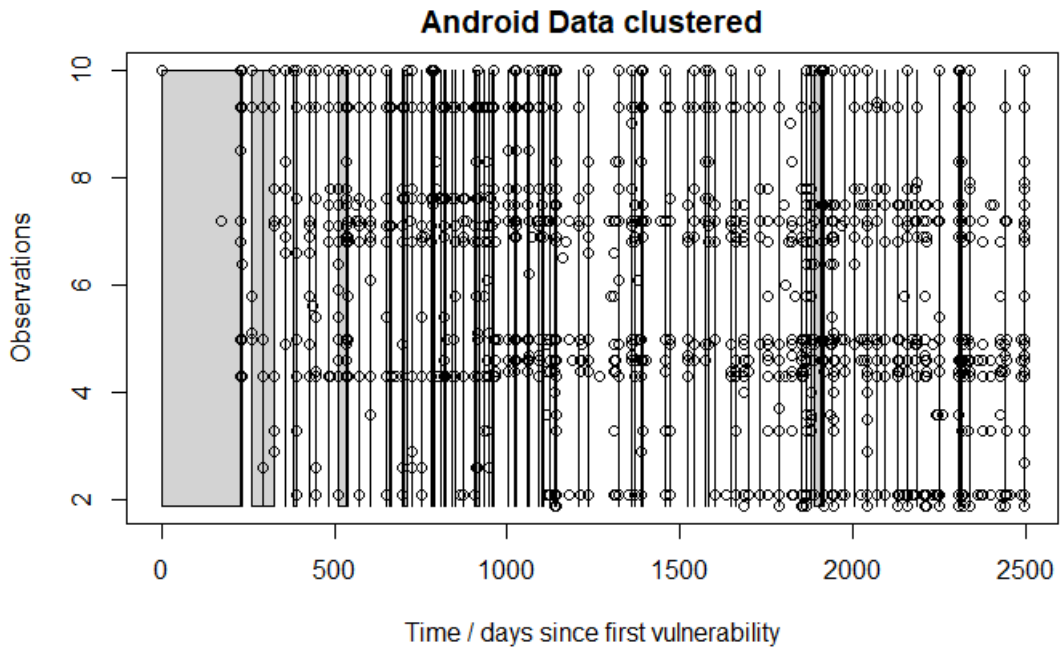


FIGURE 4.1: Clusters graphic representation. The clusters are defined by each grey rectangle

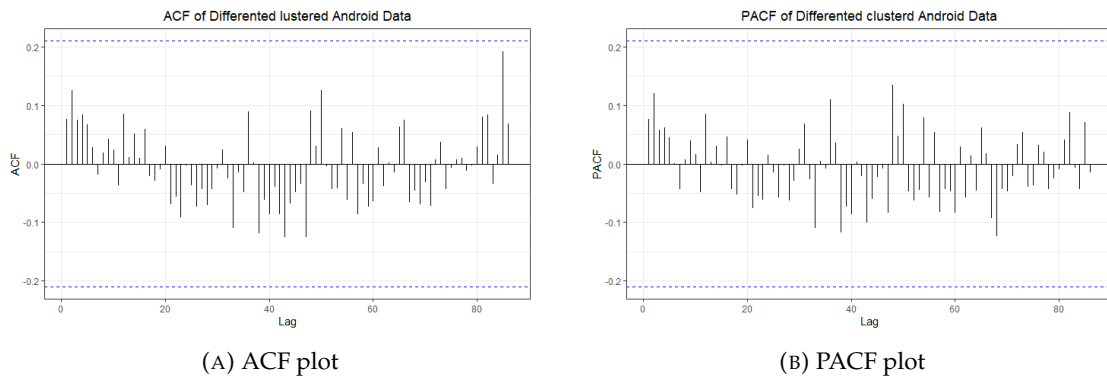


FIGURE 4.2: Correlation plots

the observations of interest are located. If we try to model a GPD on these excesses, we get meaningless results. The most plausible alternative would be to replace the severity with a quantity that is directly correlated but not bounded. One of these quantities could be the potential losses to the entity if the vulnerability were exploited. This information was public information until the mid-2010s, where, with the growing importance of systems security, it became confidential information. Without these data it is impossible to complete the POT model satisfactorily.

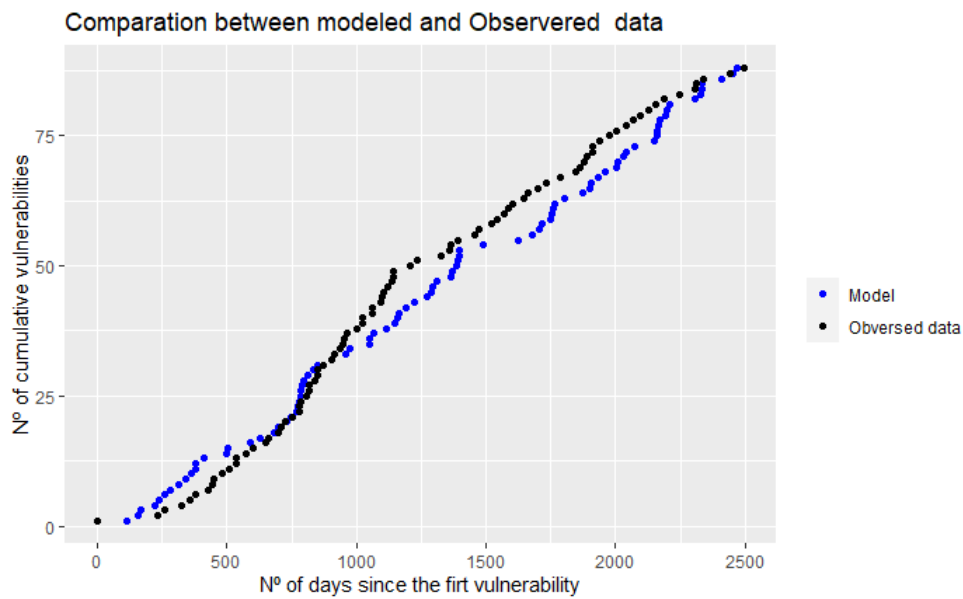


FIGURE 4.3: Comparison between the model and observed data

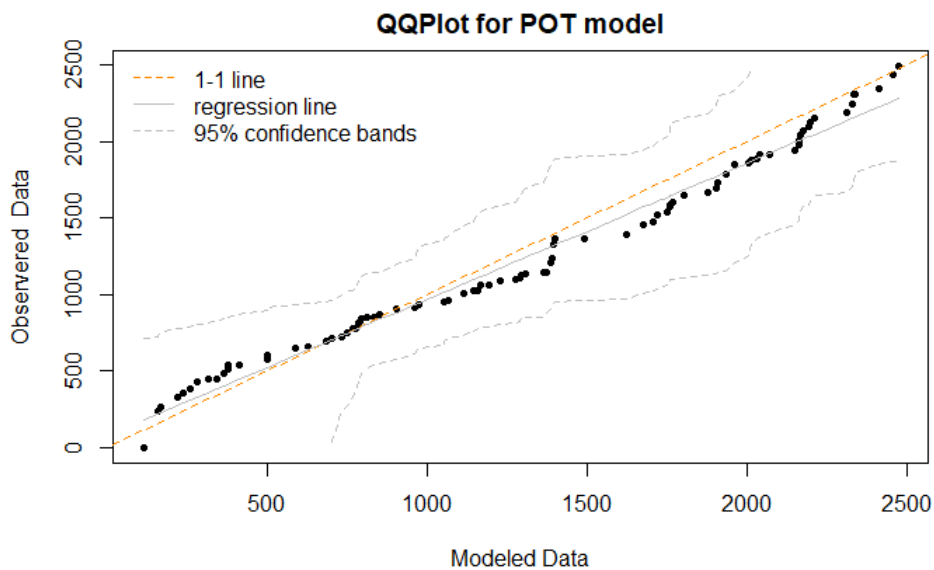


FIGURE 4.4: QQ-plot between the model and observed data

## 4.7 Results

In this chapter, the classical extreme value theory was introduced, as well as adaptations for the introduction of dependence. The POT model is introduced as a possible generalization of other models as well as its restrictions. To satisfy the model conditions, a declustering process is performed for the exceedances. The time of exceedances is successfully modeled using a Poisson model, supported by graphical and quantitative diagnostic tests. Exceedances are not possible to be modeled by a GPD as they have an associated



upper bound. This creates unexpected behavior in modeling the extreme value function. One way of circumventing the problem itself is to not estimate the distribution function directly but other important statistics that are associated with it. This will be done in the next section.



## Chapter 5

# Modelling through EVT for discrete data

### 5.1 Introduction

Another possible perspective on the problem is to avoid estimating the distribution of the  $F$  data and to use other adequate statistics in order to obtain the desired information. One of these statistics, commonly used in the extreme values approach, is the return level. Return level, which was introduced in the section 4 is a relevant statistics when we want to know tail behaviors such as rare or extreme events, important for various sectors such as risk analysis, flood studies or epidemiological studies. More specifically, the return level seeks to answer the question of what is the threshold  $z_t$ , associated with a set of variables  $(X_1, \dots, X_t)$  with marginal distribution  $F$ , where we expect to see a value greater than  $z_t$  occurred in  $t$  times, corresponding to solving the equation 5.1:

$$\mathbb{E}\left(\sum_{i=1}^t \mathbb{1}(X_i > z_t)\right) = 1 \quad (5.1)$$

Since the variables iid have a marginal distribution  $F$ , we can rewrite the equation 5.1 in the 5.2 :

$$F(z_t) = 1 - \frac{1}{t} \quad (5.2)$$

Thus, we can see that the threshold  $z_t$  is just a quantile  $p$  with  $p = 1 - \frac{1}{t}$

In the context of extreme value theory, the return level is a well documented and perceived measure, especially in the continuous case where large sample sizes are available. There the asymptotic behavior of the tail can be modeled through a particular distribution

on some parameters. In a GEV model, through the block maximum method, it is possible to obtain an exact expression for return levels, thus obtaining a natural estimation through the estimated parameters of the model:

$$\hat{z}_t = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} [1 - (-\log(1 - 1/t))^{\hat{\xi}}] \quad (5.3)$$

In the same way, we can use a threshold exceeding model (POT) to get a exact expression of the return levels, estimating it naturally by using that equation 5.4:

$$\hat{z}_t = \hat{z}_* + \frac{\hat{\sigma} + \hat{\xi}(\hat{z}_* - \mu)}{\hat{\xi}} [(t\hat{F}(\hat{z}_*))^{-\hat{\xi}} - 1] \quad (5.4)$$

using a GPD that fits the conditioned tail  $\mathbb{P}(X > Z | X > z_*)$  choosing a fixed threshold  $z_*$ . However, when dealing with discrete variables, as in our case, or with small sample sizes, the assumptions for building models in EVT do not hold [33], forcing us to look for other solutions. Fortunately it is possible to work around this problem not by estimating exactly the return level  $z_t$ , but an upper limit  $b_t$ . Although this upper limit is not guaranteed to be very close to the return level  $z_t$ , in the context in which we present ourselves, this will not be a problem because, even if the knowledge of the approximate number of vulnerabilities at a given time is valuable information, a pessimistic view of this value is enough to make informed decisions in order to mitigate the possible effects of vulnerabilities.

### 5.1.1 Estimation of the upper limit of return level

The following derivations closely follow [27]. The inequality that allows estimating the upper bound is the Markov inequality written as follows:

$$\bar{F}(z) \leq \frac{\mathbb{E}(h(X))}{h(z)} \quad (5.5)$$

where  $\bar{F}(z) = 1 - F(z)$ ,  $z > 0$  and  $h$  a positive and increasing function in a latter sense. The study of different candidates for the  $h$  function has been extensively applied, leading to different bounds such as the Chernoff or moment bounds [27]. However, it is possible to construct a function  $h$  that is simple and provides a better alternative than the aforementioned bounds. Let  $h(z) = u(z)v(F(z))$ , where  $u$  and  $v$  represent non-negative and non-decreasing functions defined on  $[0, \infty)$  and  $[0, 1]$ , respectively. In the particular case where  $u(x) = x^\alpha$  and  $v(x) = x^\beta$ , with integer  $\alpha$  and  $\beta$  positive, we obtain the PWM (probability weight moments), introduced in [31]. Thus, we can rewrite the 5.6 equation in the

form:

$$z_t \leq b_t(u, v), \quad b_t(u, v) = u^{\leftarrow} \left[ \frac{t\theta(u, v)}{v(1 - 1/t)} \right] \quad (5.6)$$

where  $\theta(u, v) = \mathbb{E}(u(X)v(F(X)))$  and  $u^{\leftarrow}$  is the generalized inverse of  $u$ . The moment  $\theta(u, v)$  is the main parameter in this estimation, as it is the only one dependent on  $X$ , that is, on the data. Some properties of this moment are important to be studied, especially when it comes to convergence properties. This process is well documented in [27]. To estimate the return level, it will be necessary to estimate  $\theta(u, v)$  and choose the respective functions  $u$  and  $v$  so that the estimate reaches the real value as close as possible. A simple way to estimate  $\theta(u, v)$  is through the equation 5.7:

$$\frac{1}{n} \sum_{i=1}^n u(X_i)v(F(X_i)) \quad (5.7)$$

But as we do not know the  $F$  distribution, we resort to an L statistic, that is, a linear combination of order statistics, thus obtaining:

$$\hat{\theta}_n(u, v) = \frac{1}{n} \sum_{i=1}^n u(X_i)v\left(\frac{i}{n}\right) \quad (5.8)$$

In this way, choosing appropriate  $u$  and  $v$ , we can then estimate the return levels. The family of these functions is only restricted to non-decreasing and positive functions, so the possible choices are vast. As mentioned above, a simple but effective choice is to choose  $u(x) = x^\alpha$  and  $v(x) = x^\beta$ , with  $\alpha$  and  $\beta$  positive integers, making the momentum estimation of the form::

$$\hat{\theta}_n(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n (X_i)^\alpha \left(\frac{i}{n}\right)^\beta \quad (5.9)$$

and the upper bound estimation:

$$\hat{b}_t(\alpha, \beta) = \left[ \frac{t\hat{\theta}(\alpha, \beta)}{(1 - 1/t)^\beta} \right]^{\frac{1}{\alpha}} \quad (5.10)$$

With the information of the upper limits of the return levels, it is possible to build a system where we associate a return time to an observation and verify, in the future, if there is a point that exceeds the initial observation within the return time. This is possible if we assume independence (or low correlation) because we can say that  $\mathbb{E}(\sum_{i=1}^T \mathbb{1}(X_i \geq X_{t_0})) = 1$ , that is, we only expect an observation greater than  $X_t$  during  $[t, t + b_t]$ .

### 5.1.2 Implementation

As described in the previous section and following [26], the alarm system can be summarized in the following steps:

1. Estimate return times  $\hat{b}_t$  taking into account vulnerability records
2. Assign to each observation  $x_t$  a return time  $\hat{b}_t$
3. For each new observation  $x_{t_0}$ , check if there are previous observations  $x_t$  where  $t_0 \leq t + \hat{b}_t$  and sound the alarm if  $x_t < x_{t_0}$

It should be noted that, for the estimation of  $\hat{\theta}_n$  and consequent  $\hat{b}_t$ , it is necessary to choose the parameters  $\alpha$  and  $\beta$  for each of the times considered. Naturally we want our upper bound to be as small as possible, so we must numerically find:

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}[\hat{b}_t(\alpha, \beta) : \alpha \in [\varepsilon, \alpha_{max}], \beta \in [\varepsilon, \beta_{max}]] \quad (5.11)$$

where  $\varepsilon$  is a value close to zero. The choice of  $\alpha_{max}$  and  $\beta_{max}$  hugely influences the results and is heavily dependent on data [10]. In this way, it will be important to focus first on an adequate choice of these parameters to succeed in the correct estimation of the return times and, therefore, of the alarm system as a whole.

These variables will be optimal when their increase does not change the behavior of the return levels, that is, when there is stabilization of the process. This search process can be done through a grid search, that is, an intensive search where we analyze all combinations of parameters within a pre-established limit. When increasing the parameters does not bring changes to the return levels, then  $\alpha_{max}$  and  $\beta_{max}$  will be the minimum of the parameters where this phenomenon was observed.

The software used for data modeling was R core team [37], adapting the code provided by [10].

## 5.2 Modeling Android data

For Android data, a grid search of values between one and ten was performed for each of the parameters. At the beginning of the search, that is, in the 5.1a graph, the return levels have a linear behavior, which is uncharacteristic behavior for this quantity. Increasing the parameter  $\beta_{max}$  to 5 (graph 5.1b) did not significantly change the behavior. Alternatively, when parameter  $\alpha_{max}$  was increased to 5 (graph 5.1c), it revealed a logarithmic behavior,

which is expected. There are no different results by increasing further the parameter  $\beta_{max}$  or  $\alpha_{max}$  from these values, which suggests that the process has stabilized. Thus, the values of  $\alpha_{max}$  and  $\beta_{max}$  5 and 1, respectively, were chosen. Another factor that can be decisive

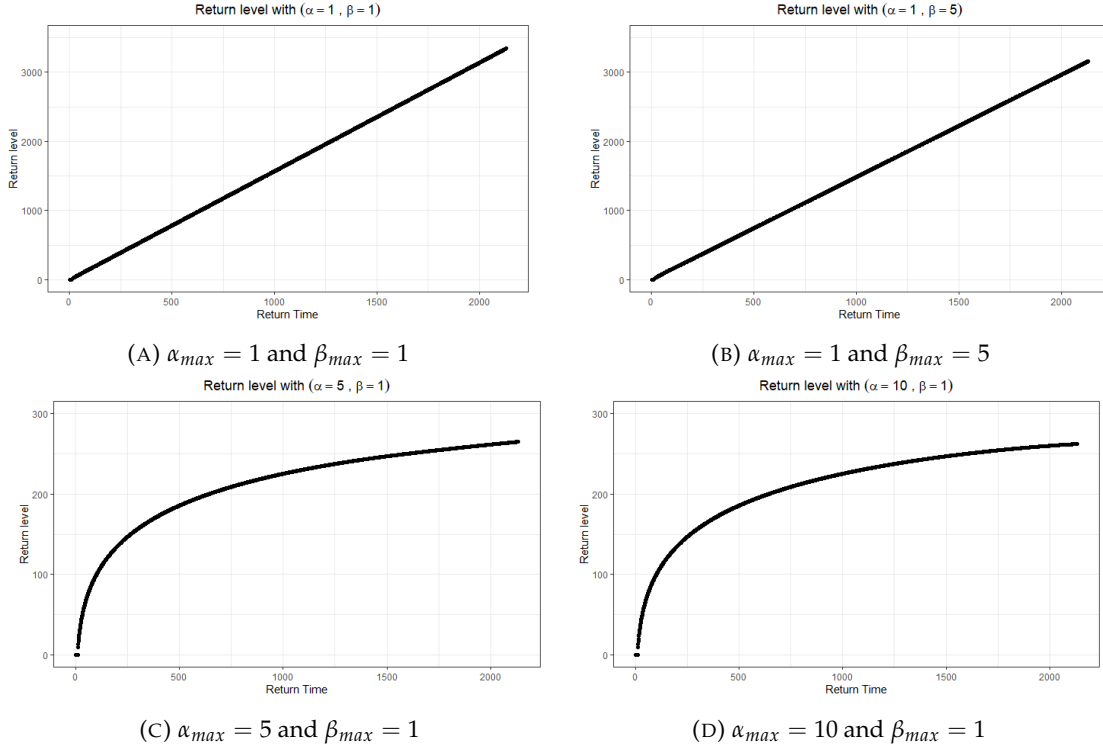


FIGURE 5.1: Grid search plots

for the good performance of the model is the grouping of the data. If no data aggregation is done, we will have many consecutive days where few vulnerabilities are observed or not observed at all, which can influence the alarm system in unwanted ways. On the other hand, vulnerabilities with different impacts have different behaviors, which may justify implementation for critical vulnerabilities only. In this way, the alarm system was considered in four cases: all daily and weekly vulnerabilities, as well as only daily and weekly critical vulnerabilities. It is possible to observe the estimated return levels for the different data groupings (figure 5.2). All of these exhibit similar behaviors, just on different scales. A possible interpretation of the graphs 5.2 would be, for example, is expected to wait 500 days to observe approximated 180 vulnerabilities for daily, or wait 200 weeks to observe approximated 80 critical vulnerabilities for weekly data. Applying the alarm system to the data described above we obtain the following results represented in the figure 5.3:

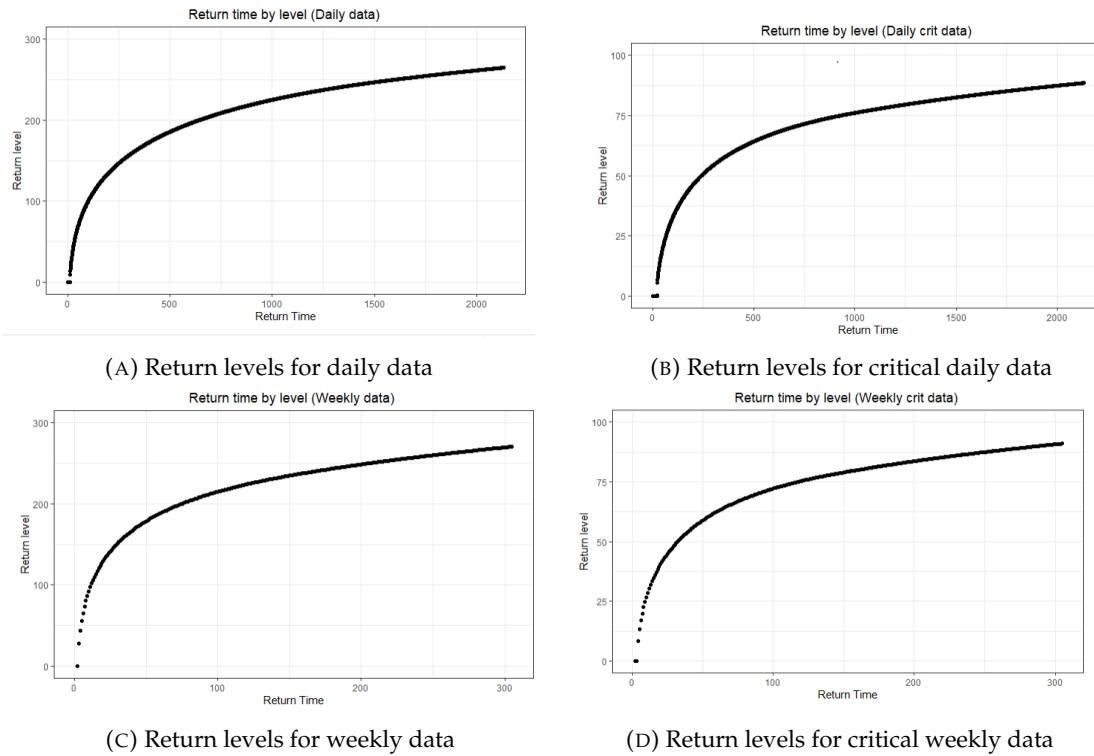
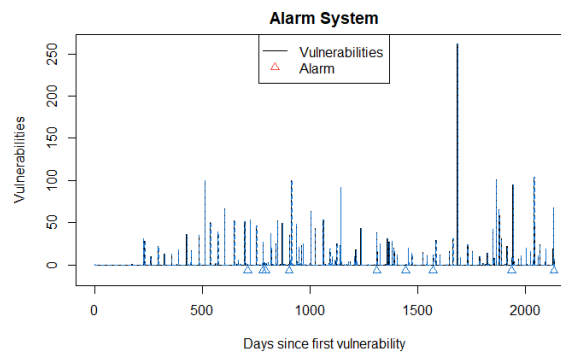


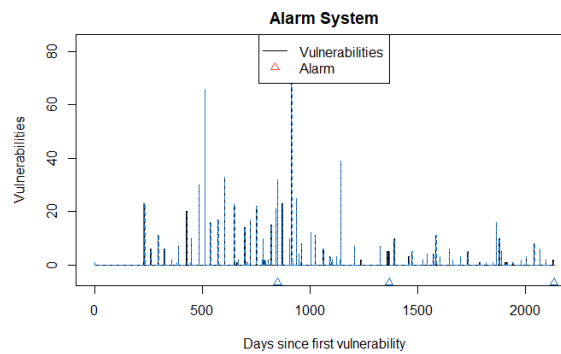
FIGURE 5.2: Return levels for different aggregated data

The daily alarm system (figure 5.3a) in addition to creating few alarms for the vulnerabilities present, these are generally not concentrated close to those of the days when more vulnerability records occurred. Worse happens for the daily critical alarm system (figure 5.3b), where there were only 3 alarms, where only one of them is close to the relevant days for the study. The opposite is observed for the weekly alarm system. In addition to more alarms being observed (figure 5.3c), these are more condensed near the days with the highest vulnerability peaks. The same can be said of the critical weekly alarm system (figure 5.3d) which, despite registering fewer alarms than when considering all the data, these are still close to the relevant days.

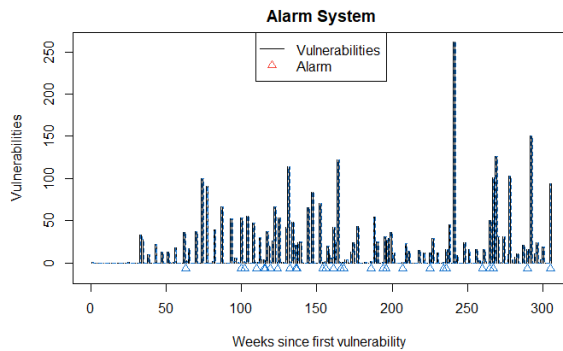




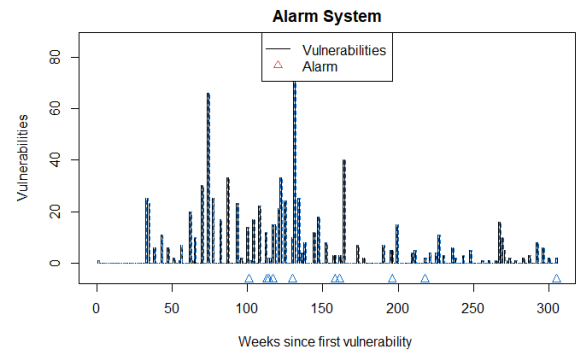
(A) Alarm system for daily data



(B) Alarm system for critical daily data



(C) Alarm system for weekly data



(D) Alarm system for critical weekly data

FIGURE 5.3: Alarme system for different aggregated data

### 5.3 Metrics

Although the simple direct comparison between the alarm and the vulnerabilities gives a good idea on how adequate the alarm system was to detect them, it does not correctly or accurately represent the circumstances in which the alarms will be useful for the entities in question nor does it tell us the advantages and disadvantages of grouping the data. More specifically, the alarm system must be able to predict (within a time window) where the greatest number of vulnerabilities occur, in order to warn the entity to be more careful in

that period of time. Therefore, a numerical comparison between the alarm systems must be considered. For this purpose, two metrics are proposed: a binary and a general metric.

### 5.3.1 Binary metric

The binary metric proposes a zone immediately close to the alarm where it would be more useful to detect amounts of abnormal vulnerabilities. More specifically, we classify a good alarm if it precedes, within the fixed window, an observation with vulnerabilities above a pre-established threshold. As the observations with more vulnerabilities are of greater interest, we studied the effect of the threshold chosen through four high quantiles closely (0.90, 0.95, 0.975 and 0.99) and represented lower threshold more spaced (from 0.20 to 0.90 by spaces of 0.1).

One way of representing this metric is to create pseudo confusion matrices for each threshold and for each window size considered. These confusion matrices are constituted by the number of alarms that precede vulnerabilities within the chosen range (true positives), number of alarms that do not precede vulnerabilities within the chosen range (false positives) and by the number of vulnerabilities above the threshold that were not detected (false negatives). Thus, we can evaluate classic metrics such as sensitivity. It is important to note that, for a fairer comparison between groups of data, the range of acceptance regions are equal for weekly and daily data.

Instead of comparing confusion matrices, another way of proceeding is to represent these statistics through a Precision-Recall (PR) curve, that is, a comparison of the quality of hits (precision) against the amount of hits (recall). Classically the precision is defined by the equation 5.12:

$$P = \frac{TP}{TP + FP} \quad (5.12)$$

where TP are true positives and FP false positives, while Recall is defined by the equation 5.13:

$$R = \frac{TP}{TP + FN} \quad (5.13)$$

where FN are the false negatives. This curve is popular for comparisons of unbalanced data as it gives primacy to positive classification by varying the cut-off point. As our model is not probabilistic, the classical definition of a cut-off point is not applicable. However, it is possible to redefine the cut-off point by associating it with a quantile of vulnerabilities, as described above. Thus, it is possible to draw the PR curve for the alarm system,

for each window size considered.

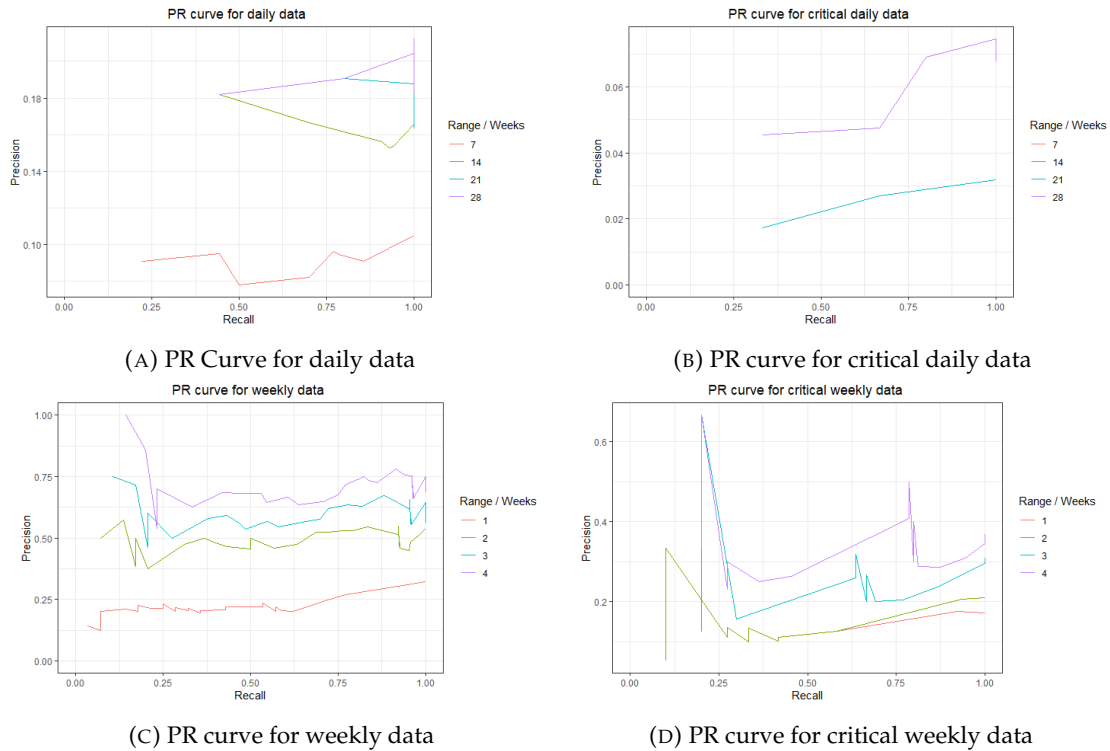


FIGURE 5.4: PR curves for various type of aggregated data

As far as data aggregation is concerned, the behavior of PR curves is different depending on the type of data being dealt with. It can be seen that for the weekly data (figure 5.4c and 5.4d), the curves present a higher precision than the daily ones (figure 5.4a and figure 5.4b). When comparing the plots where all the data are incorporated (figure 5.4a and figure 5.4c) with only the critical ones (figure 5.4b and figure 5.4d), the former have higher accuracies. Thus, following the results of this metric, the most useful aggregation will be the weekly one with all available data, represented in the figure 5.4c.

The difficulty in interpreting this metric is finding a compromise between the flexibility and predictability of the model, that is, choosing the acceptance window in order to have good results with relevant information. This choice is a recurring problem in different contexts, particularly in clinical settings. Within this area, the choice of when to apply diagnosis or treatment to patients who are likely to be sick, especially if it could harm the patient, is crucial. Previous studies [45] in this area propose a method of aid for this choice called net benefit. This metric has a simple composition, being only the difference between the ratio of true positives and false positives, influenced by the chosen cut-off

point:

$$NB = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{p_c}{1 - p_c} \quad (5.14)$$

where  $NB$  is the net benefit,  $TP$  are the true positives,  $FP$  the false negatives,  $N$  the number of vulnerabilities and  $p_c$  the chosen cut-off point. Comparing the model's net benefit curve with baseline curves (in which case there is an alarm on each day and in a case in which there is no alarm), we can observe the cut-off points where the alarm has better performance. In addition, if we compare different acceptance windows, we help the decision to choose this one. First, we will compare clusters to corroborate the results presented in the PR curves. For a fair comparison between clusters, acceptance windows of identical size were chosen.

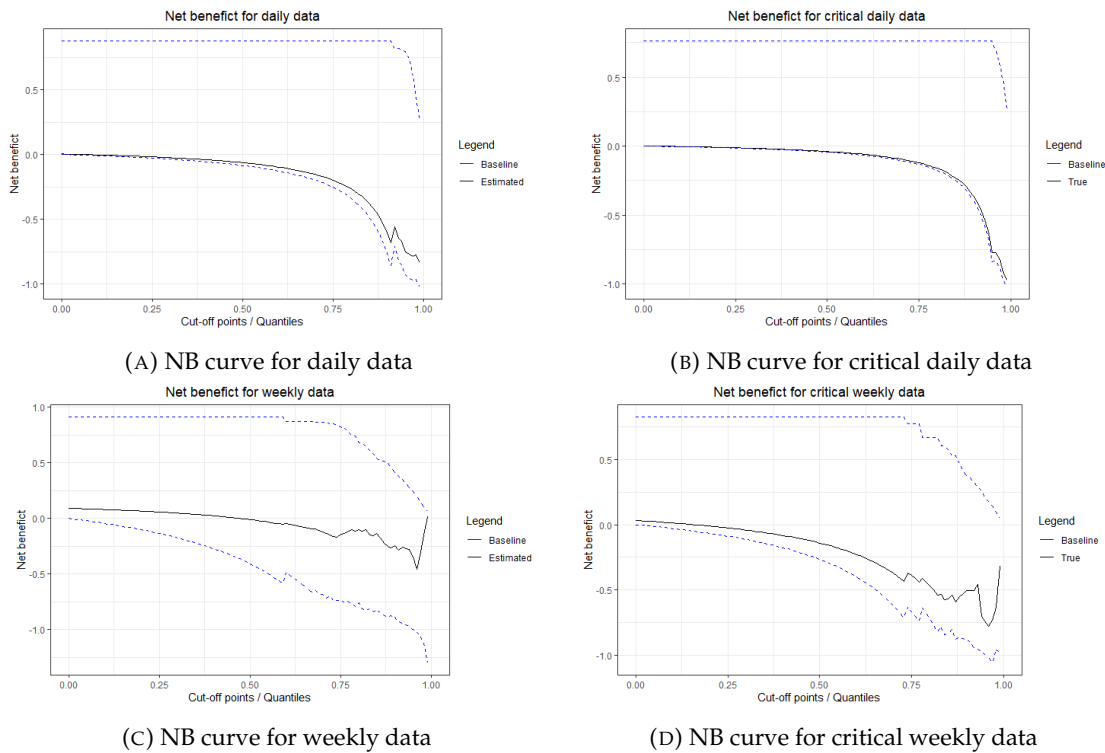


FIGURE 5.5: NB curves for various type of aggregated data

Observing the NB curves, we noticed that in none of the data clusters the curve of the estimated model exceeds the upper baseline model, which means that none of the models would be better, according to this metric, than the model in which an alarm was triggered every day. All comparisons that were observed in the PR curves can also be observed in the NB curves, that is, the clustering with all the data (figure 5.5a and 5.5c) fits better than than only critical data (figure 5.5b and 5.5d) and alarms with weekly data (figure 5.5c and 5.5d) have better performance than the daily (figure 5.5a and 5.5b), since

the curve have higher values. One of the particularities of this metric is the sensitivity to cut-off points. It is possible to observe this property in high cut-off points, more notable in the data grouped weekly where there is a growth in the metric near the cut-off of 0.9. Finally, different acceptance windows will be compared to help in the decision of choosing a better acceptance window. As an example, only data grouped weekly with activation windows between [3,5] will be analyzed (in C they are analyzed between [2,6]).

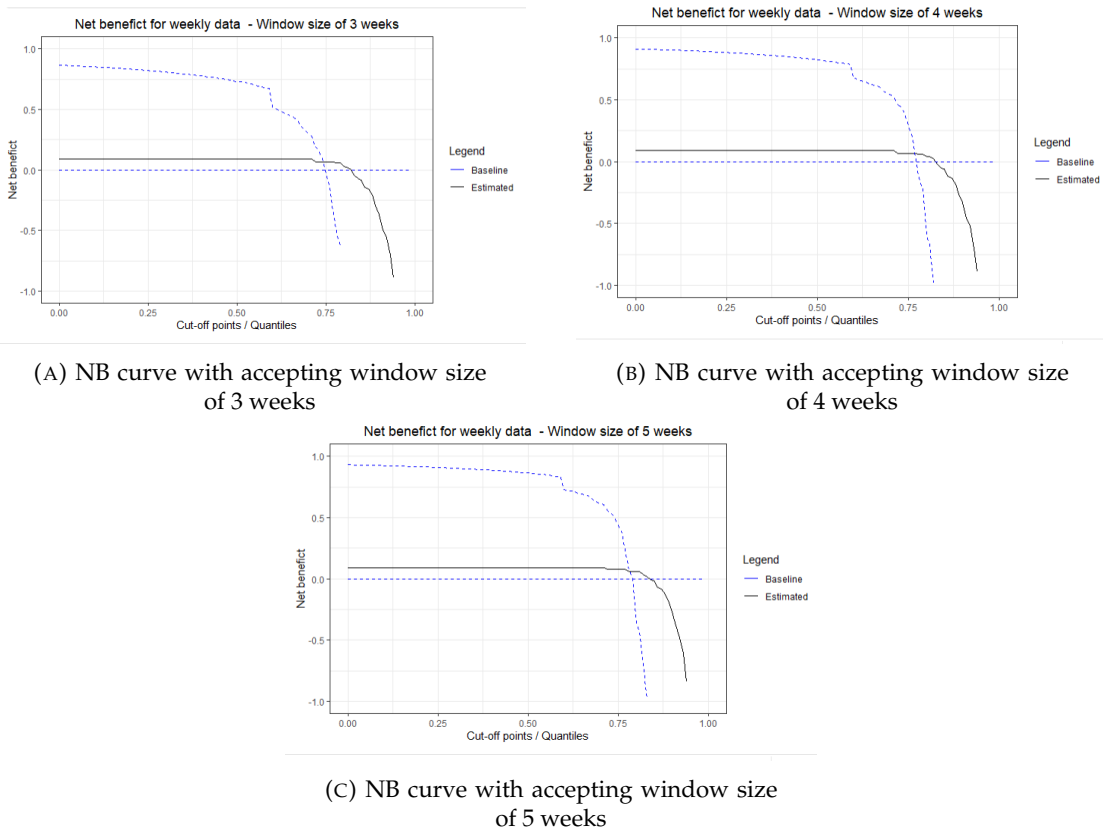


FIGURE 5.6: NB curves for various window sizes

Observing the graphs in the figure 5.6, it is possible to see that there is an improvement in the metric with the increase of the acceptance window, as expected. The new information gained from these charts is that there is a greater jump in performance from the third to the fourth week, not being registered such an impressive gain from the fourth to the fifth week. This suggests that the increase from the fourth week onwards does not justify the loss of model predictability that occurs with the increase in the acceptance window.

### 5.3.2 General metric

The continuous metric aims to classify the system as a whole, giving more importance to alarms that predate large amounts of recently recorded vulnerabilities. This can be obtained through the following formula 5.15:

$$m_c = \sum_a^n \frac{\pi_{l,i} \pi_{w,i}}{n} \quad (5.15)$$

, where an average of the sum of products is calculated between the weights associated with the location of the alarm in relation to each vulnerability  $\pi_{l,i}$  and the number of vulnerabilities registered at that point  $\pi_{w,i}$ , for each of the registered alarms  $a$ . The weights  $\pi_{l,i}$  and  $\pi_{w,i}$  are scaled to the unity, being close to 1 when vulnerabilities are close to alarm, decaying normally, or when the number of observed vulnerabilities is close to the observed maximum, respectively. These scalings cause the proposed metric to vary between 1 and 0.

Table for general metric				
Data Grouping Type	Accepting window of 1 week	Accepting window of 2 week	Accepting window of 3 week	Accepting window of 4 week
Daily	0.0391	0.0845	0.1185	0.1404
Critical daily	0	0	0	0.0376
Weekly	0.0275	0.09085	0.1272	0.1571
Critical Weekly	0.1209	0.1238	0.1620	0.2032

In this metric, we see that grouping the critical daily data is not suitable for this alarm system, while the grouping through the critical weekly data seems to behave better. For all data types, daily grouping is better than weekly only when the reach is 1 week, with weekly being better for subsequent larger windows. It should also be noted that, from a range of 3 weeks with the exception of critical data, all groupings are of the same order of magnitude.

## 5.4 Results

It can be seen that the two proposed metrics highlight different qualities that are intended for an ideal alarm system. Taking this into account, the two metrics suggest that weekly data is a better choice than daily ones, disagreeing only on whether to perform the model

only with critical data or not. In addition, the alarm system takes extreme values into account and incorporates all observations in its model, characteristics that were not observed in other models mentioned above. However, it is important to note that there are limitations associated with the proposed model, namely being heavily dependent on the observed data and being static in time, thus being vulnerable to changes, whether sudden or gradual, in the conditions and environment of the observed data.





## Chapter 6

# Conclusion and future work

In this dissertation, three different approaches were conducted that explored different characteristics of vulnerabilities. It was possible to model the vulnerabilities using point processes, where the most successful models were those where extra information about the vulnerabilities was incorporated, such as their severity. In addition, the inclusion of non-linear terms with time helped to obtain a better performance. These results coincide with the information taken from the descriptive analysis, where different behaviors were observed in vulnerabilities with different severities in addition to non linear behaviors over time.

With regard to modeling through classical extreme value theory, the POT model was applied, obtaining satisfactory results in the declustering process as well as in the modeling of inter arrival times. However, it was not possible to complete the model as the exceedances are limited, thus not being possible to apply a GPD. With additional information about the possible losses of entities given a vulnerability, this modeling could be completed, revealing critical information for interested entities.

Finally, the extreme value theory was applied again, this time without trying to estimate (or approximate) distribution functions, using only useful statistics from this theory. Using the return times, it was possible to build an alarm system to detect anomalous numbers of registered vulnerabilities. By grouping the data weekly, it was possible to build a satisfactory alarm system with a predictive capacity of 1 month. This allows the system to alert the entity of the possibility of an anomalous record of vulnerabilities in the next month. Several metrics were applied to the alarm system to measure its predictability and help in choosing its parameters.

Although it is possible to model vulnerabilities, there are several limitations and assumptions that have been made. One of these limitations is the fact that the vulnerability registration date is considered to be the same when it was first discovered. There is a lot of literature that confirms that this assumption is false, so this analysis will have to be done with these types of data. Another limitation associated with the previous one is the lack of public details of some relevant data, such as information on possible losses from exploiting the vulnerability. This type of information is sensitive to each entity and is only disclosed within the closed circle of risk analysts. In any case, the inclusion of these data in this analysis would allow finalizing the POT model and building other more complex models.

This dissertation was developed by using only the Android dataset. Other datasets like IOS, Windows7, Internet Explorer can and will be explored in future developments.

For future work, it would be natural to expand the Poisson point models to cluster methods such as the Cox model. The alarm system built in this dissertation can be improved if the estimation of an upper limit for the return time is more accurate. In addition, the incorporation of other metrics such as SEID [20], may provide other information about the performance of alarm systems.

# Appendix A

## Equations for CVSS metric

```
BaseScore = round_to_1_decimal(((0.6*Impact)+(0.4*Exploitability)-1.5)*f(Impact))
Impact = 10.41*(1-(1-ConfImpact)*(1-IntegImpact)*(1-AvailImpact))
Exploitability = 20* AccessVector*AccessComplexity*Authentication
f(impact)= 0 if Impact=0, 1.176 otherwise

AccessVector    = case AccessVector of
    requires local access: 0.395
    adjacent network accessible: 0.646
    network accessible: 1.0

AccessComplexity = case AccessComplexity of
    high: 0.35
    medium: 0.61
    low: 0.71

Authentication  = case Authentication of
    requires multiple instances of authentication: 0.45
    requires single instance of authentication: 0.56
    requires no authentication: 0.704

ConfImpact      = case ConfidentialityImpact of
    none: 0.0
    partial: 0.275
    complete: 0.660

IntegImpact     = case IntegrityImpact of
    none: 0.0
    partial: 0.275
    complete: 0.660

AvailImpact     = case AvailabilityImpact of
    none: 0.0
    partial: 0.275
    complete: 0.660
```

FIGURE A.1: Equation for base calculation for CVSS in version 2.0

```

TemporalScore = round_to_1_decimal(BaseScore*Exploitability
                                   *RemediationLevel*ReportConfidence)

Exploitability = case Exploitability of
    unproven:          0.85
    proof-of-concept: 0.9
    functional:        0.95
    high:              1.00
    not defined:       1.00

RemediationLevel = case RemediationLevel of
    official-fix:      0.87
    temporary-fix:     0.90
    workaround:        0.95
    unavailable:       1.00
    not defined:       1.00

ReportConfidence = case ReportConfidence of
    unconfirmed:       0.90
    uncorroborated:   0.95
    confirmed:         1.00
    not defined:       1.00

```

FIGURE A.2: Equation for temporal calculation for CVSS in version 2.0

```

EnvironmentalScore = round_to_1_decimal((AdjustedTemporal+
(10-AdjustedTemporal)*CollateralDamagePotential)*TargetDistribution)

AdjustedTemporal = TemporalScore recomputed with the BaseScore's Impact sub-
equation replaced with the AdjustedImpact equation

AdjustedImpact = min(10,10.41*(1-(1-ConfImpact*ConfReq)*(1-IntegImpact*IntegReq)
                    *(1-AvailImpact*AvailReq)))

CollateralDamagePotential = case CollateralDamagePotential of
    none:          0
    low:           0.1
    low-medium:    0.3
    medium-high:   0.4
    high:          0.5
    not defined:   0

TargetDistribution = case TargetDistribution of
    none:          0
    low:           0.25
    medium:        0.75
    high:          1.00
    not defined:   1.00

ConfReq = case ConfReq of
    low:           0.5
    medium:        1.0
    high:          1.51
    not defined:   1.0

IntegReq = case IntegReq of
    low:           0.5
    medium:        1.0
    high:          1.51
    not defined:   1.0

AvailReq = case AvailReq of
    low:           0.5
    medium:        1.0
    high:          1.51
    not defined:   1.0

```

FIGURE A.3: Equation for environmental calculation for CVSS in version 2.0

**ISS =  $1 - [(1 - Confidentiality) \times (1 - Integrity) \times (1 - Availability)]$**

Impact =	
If Scope is Unchanged	$6.42 \times ISS$
If Scope is Changed	$7.52 \times (ISS - 0.029) - 3.25 \times (ISS - 0.02)^{15}$
Exploitability =	$8.22 \times AttackVector \times AttackComplexity \times PrivilegesRequired \times UserInteraction$
BaseScore =	
If Impact $\leq 0$	0, else
If Scope is Unchanged	Roundup (Minimum [(Impact + Exploitability), 10])
If Scope is Changed	Roundup (Minimum [1.08 × (Impact + Exploitability), 10])

FIGURE A.4: Equation for base calculation for CVSS in version 3.0

**TemporalScore = Roundup (BaseScore × ExploitCodeMaturity × RemediationLevel × ReportConfidence)**

FIGURE A.5: Equation for Temporal calculation for CVSS in version 3.0

**MISS = Minimum ( 1 - [ (1 - ConfidentialityRequirement × ModifiedConfidentiality) × (1 - IntegrityRequirement × ModifiedIntegrity) × (1 - AvailabilityRequirement × ModifiedAvailability) ], 0.915)**

ModifiedImpact =	
If ModifiedScope is Unchanged	$6.42 \times MISS$
If ModifiedScope is Changed	$7.52 \times (MISS - 0.029) - 3.25 \times (MISS \times 0.9731 - 0.02)^{13}$
ModifiedExploitability =	$8.22 \times ModifiedAttackVector \times ModifiedAttackComplexity \times ModifiedPrivilegesRequired \times ModifiedUserInteraction$

FIGURE A.6: Equation for Modified impact Sub-Score (MISS) calculation for CVSS in version 3.0

**EnvironmentalScore =**

If ModifiedImpact $\leq 0$	0, else
If ModifiedScope is Unchanged	Roundup ( Roundup [Minimum [(ModifiedImpact + ModifiedExploitability), 10] × ExploitCodeMaturity × RemediationLevel × ReportConfidence)
If ModifiedScope is Changed	Roundup ( Roundup [Minimum [1.08 × (ModifiedImpact + ModifiedExploitability), 10] × ExploitCodeMaturity × RemediationLevel × ReportConfidence)

FIGURE A.7: Equation for Environmental calculation for CVSS in version 3.0

Metric	Metric Value	Numerical Value
Attack Vector / Modified Attack Vector	Network	0.85
	Adjacent	0.62
	Local	0.55
	Physical	0.2
Attack Complexity / Modified Attack Complexity	Low	0.77
	High	0.44
Privileges Required / Modified Privileges Required	None	0.85
	Low	0.62 (or 0.69 if Scope / Modified Scope is Changed)
User Interaction / Modified User Interaction	High	0.27 (or 0.5 if Scope / Modified Scope is Changed)
	None	0.85
Confidentiality / Integrity / Availability / Modified Confidentiality / Modified Integrity / Modified Availability	Required	0.62
	High	0.56
	Low	0.32
Exploit Code Maturity	None	0
	Not Defined	1
	High	1
	Functional	0.97
	Proof of Concept	0.94
Remediation Level	Unproven	0.91
	Not Defined	1
	Unavailable	1
	Workaround	0.97
	Temporary Fix	0.96
Report Confidence	Official Fix	0.95
	Not Defined	1
	Confirmed	1
	Reasonable	0.96
Confidentiality Requirement / Integrity Requirement / Availability Requirement	Unknown	0.92
	Not Defined	1
	High	1.5
	Medium	1
	Low	0.5

FIGURE A.8: Numerical values for version 3.0

# Appendix B

## Diagnose plots

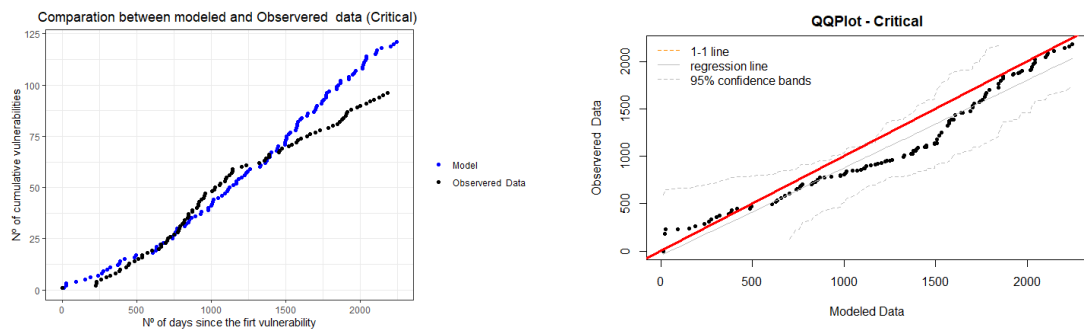


FIGURE B.1: Diagnose plots for Critical simple Poisson model

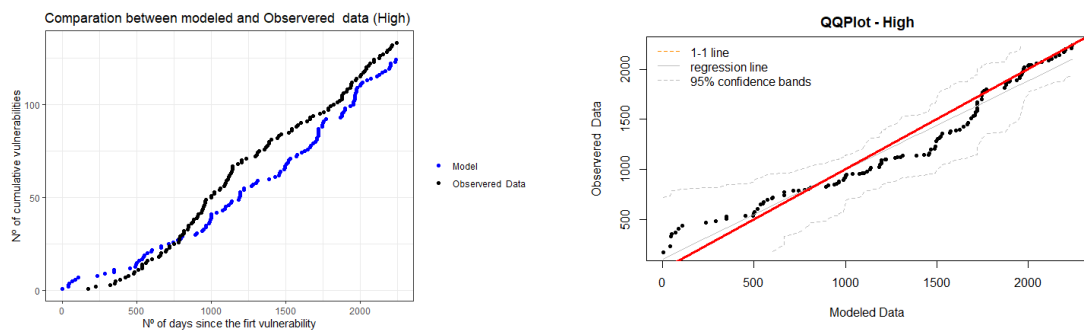


FIGURE B.2: Diagnose plots for High simple Poisson model

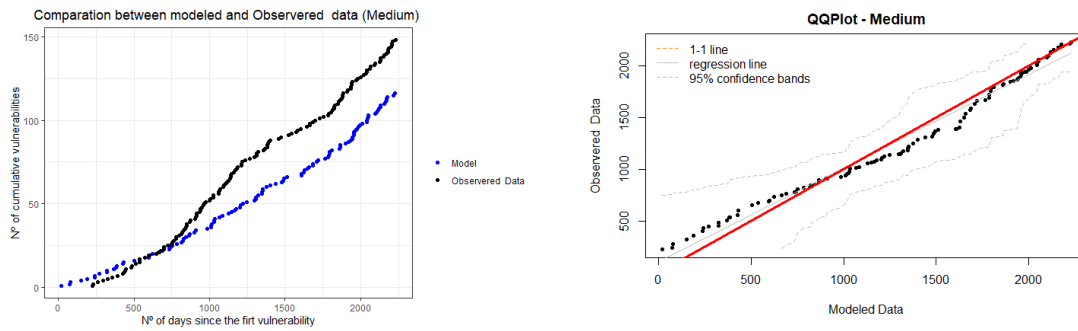


FIGURE B.3: Diagnose plots for Medium simple Poisson model

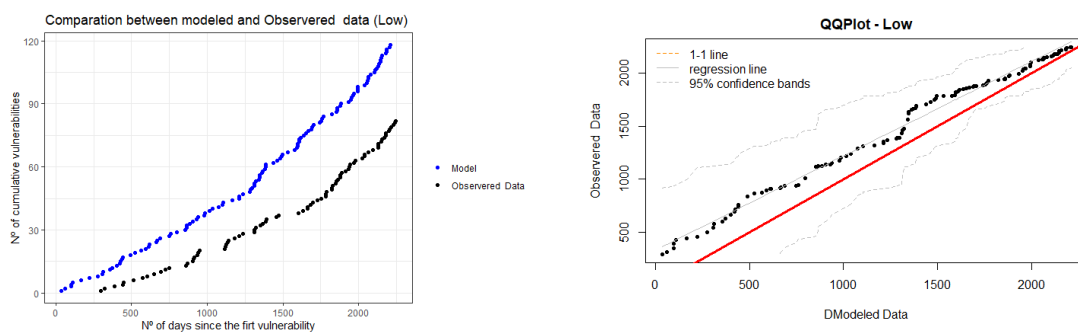


FIGURE B.4: Diagnose plots for Low simple Poisson model

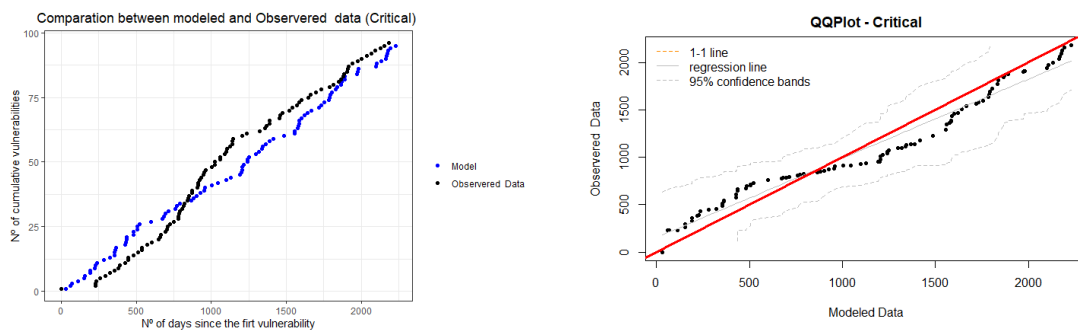


FIGURE B.5: Diagnose plots for Critical marked Poisson model

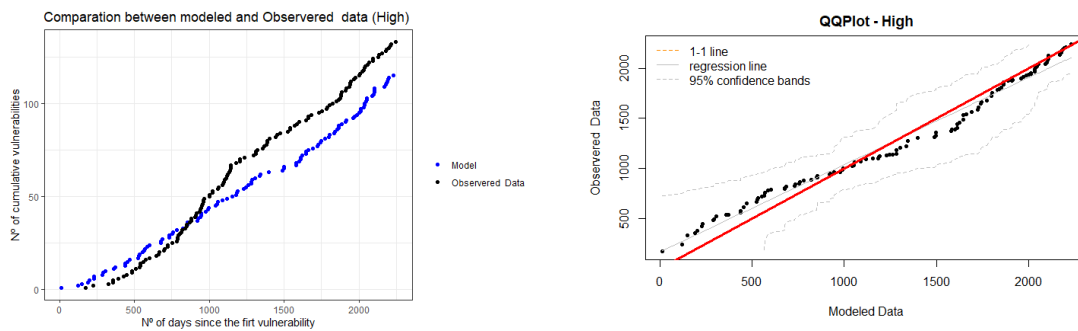


FIGURE B.6: Diagnose plots for High marked Poisson model



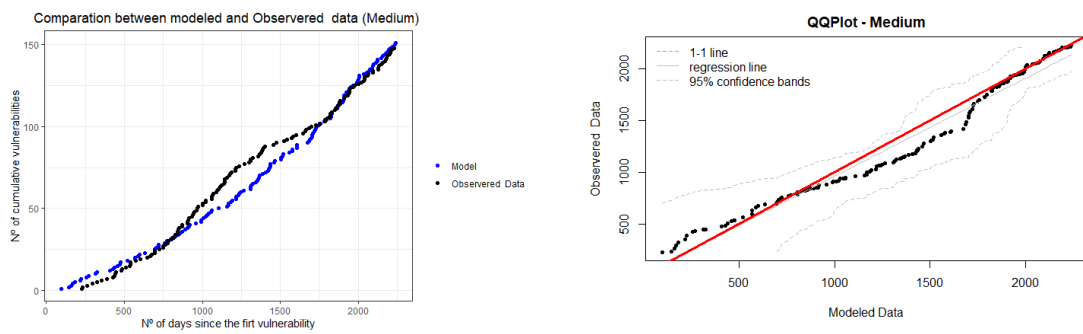


FIGURE B.7: Diagnose plots for Medium marked Poisson model

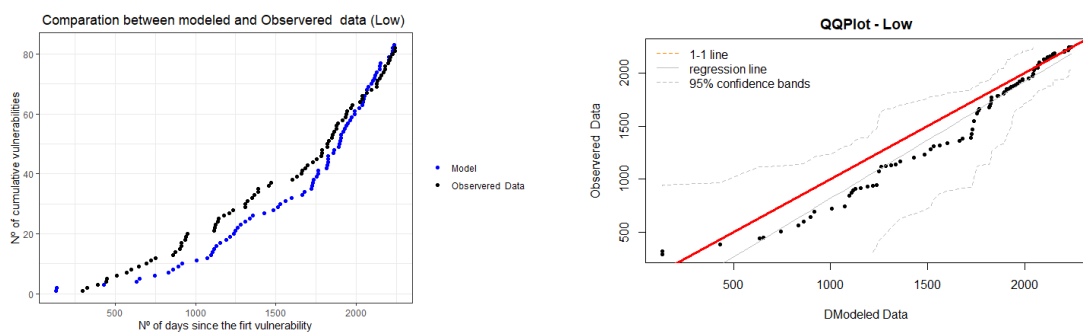


FIGURE B.8: Diagnose plots for Low marked Poisson model

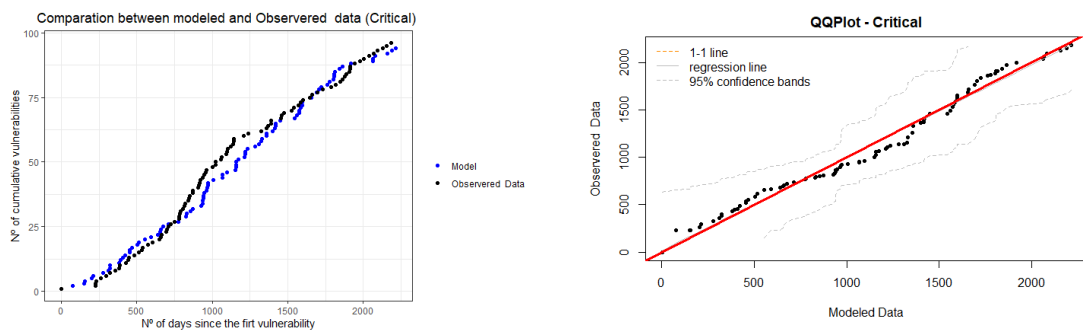


FIGURE B.9: Diagnose plots for Critical non linear marked Poisson model

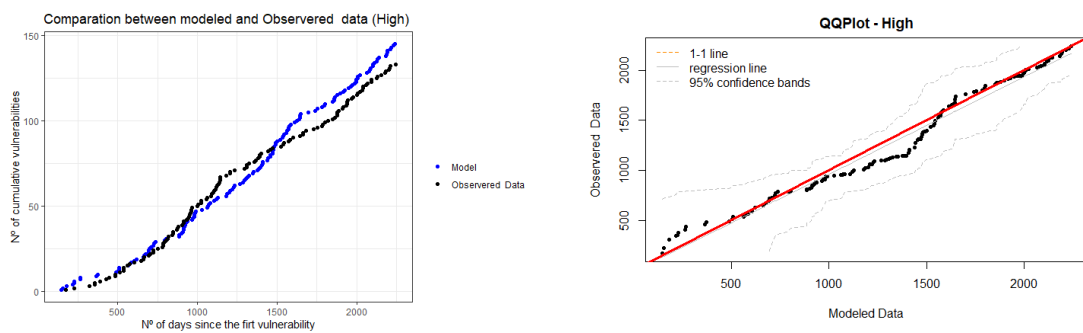


FIGURE B.10: Diagnose plots for High non linear marked Poisson model

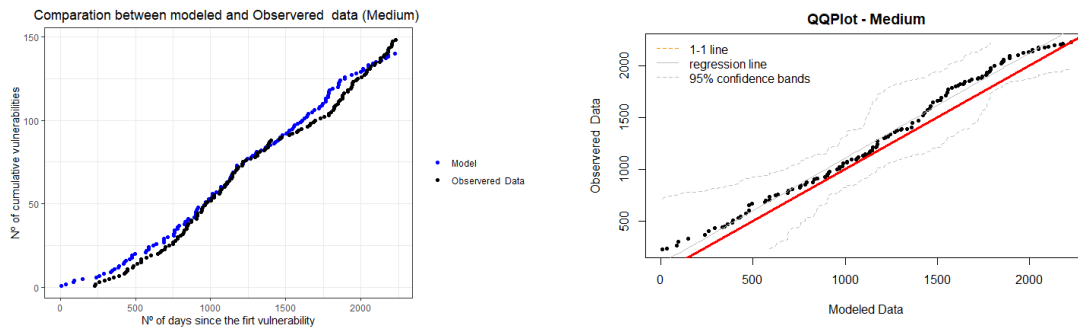


FIGURE B.11: Diagnose plots for Medium non linear marked Poisson model

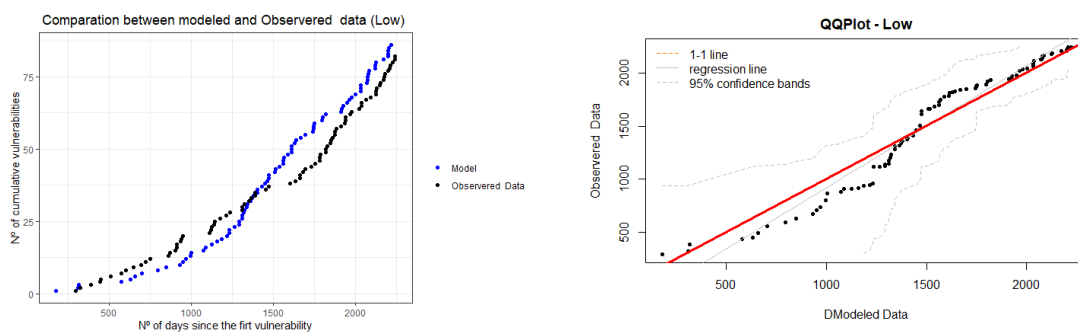


FIGURE B.12: Diagnose plots for Low non linear marked Poisson model

# Appendix C

## NB curves

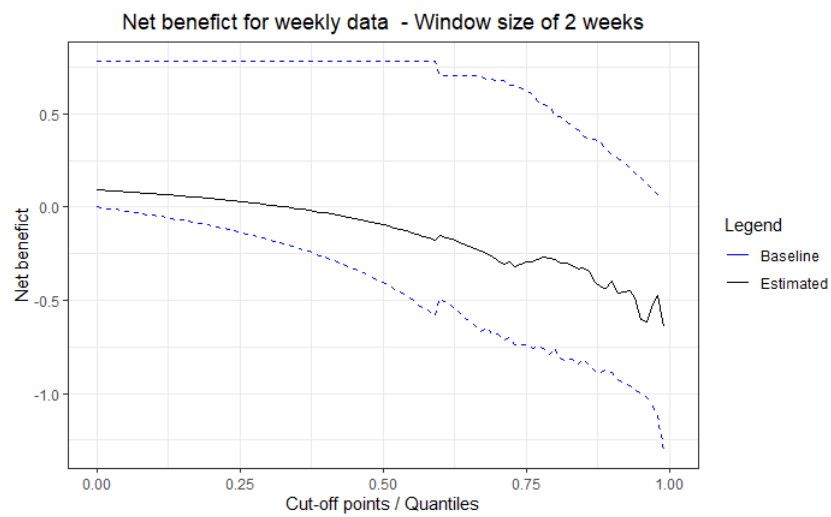


FIGURE C.1: NB curve with accepting window size of 2 weeks

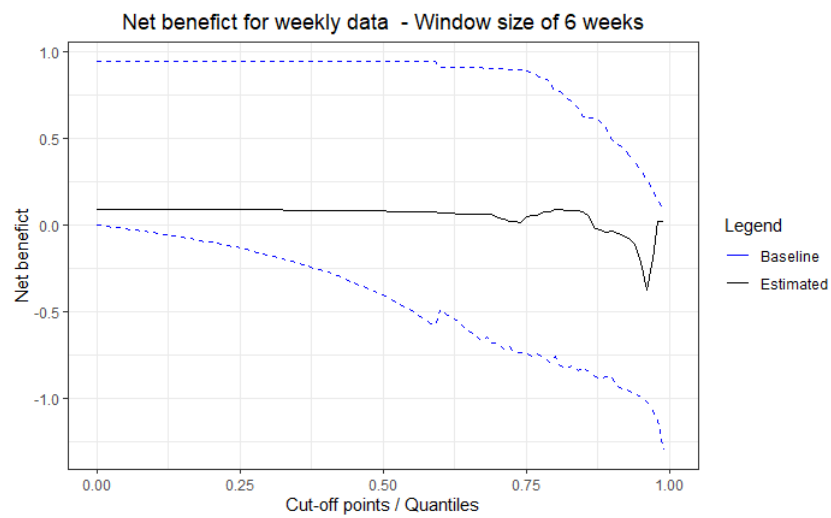


FIGURE C.2: NB curve with accepting window size of 6 weeks



# Bibliography

- [1] Abimbola, O., Akinyemi, B., Aladesanmi, T., Aderounmu, G., and Hamidja, K. (2019). An improved stochastic model for cybersecurity risk assessment. *Computer and Information Science*, 12:96 – 97. [Cited on pages [4](#) and [8](#).]
- [2] Abraham, S. and Nair, S. (2015). A novel architecture for predictive cybersecurity using non-homogenous markov models. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1. [Cited on page [8](#).]
- [3] Alhazmi, O. and Malaiya, Y. (2005). Quantitative vulnerability assessment of systems software. In *IEEE*. [Cited on pages [7](#) and [11](#).]
- [4] Alhazmi, O. and Malaiya, Y. (2008). Application of vulnerability discovery models to major operating systems. *IEEE Transactions on Reliability*, 57:14–22. [Cited on page [7](#).]
- [5] Anderson, R. (2002). Security in open versus closed systems - the dance of boltzmann, coase and moore. *Open Source Software : Economics, Law and Policy*. [Cited on pages [2](#) and [7](#).]
- [6] A.N.Pettit (1976). A two-sample anderson-darling rank statistic. *Biometrika*, pages 161–168. [Cited on page [18](#).]
- [7] Baddeley, A., Rubak, E., and Turner, R. (2016). *Spatial Point Patterns: Methodology and Applications with R*. Chapman Hall / CRC Press. [Cited on pages [13](#) and [16](#).]
- [8] Baddeley, A. and Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6). [Cited on page [20](#).]
- [9] Billingsley, P. (1995). *Probability and Measure*. John Wiley and Sons, third edition. [Cited on pages [13](#) and [14](#).]

- [10] Borchani, A. (2010). Statistiques des valeurs extrêmes dans le cas de lois discrètes. ESSEC Working paper. Document de Recherche ESSEC / Centre de recherche de l'ESSEC ISSN : 1291-9616 10009. [Cited on page 42.]
- [11] CNC (2021). Relatório riscos e conflitos 2021. Technical report, Centro Nacional de Cibersegurança. [Cited on page 1.]
- [12] Coles, S. (2004). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series. [Cited on pages 25, 32, and 33.]
- [13] CVE (2022). *Common Vulnerabilities and Exposures*. <https://cve.mitre.org/index.html>[Accessed : March 2022]. [Cited on page 3.]
- [14] CVE Details (2022). *CVE Details*. <https://www.cvedetails.com/>[Accessed : March 2022]. [Cited on page 8.]
- [15] CVSS (2022). *Common vulnerability scoring system SIG*. <https://www.first.org/cvss>[Accessed : March 2022]. [Cited on pages xv, 6, and 7.]
- [16] CWE (2022). *Common weakness enumeration*. <https://cwe.mitre.org/>[Accessed : March 2022]. [Cited on page 5.]
- [17] Daley, D. J. and Vere-Jones, D. (2005). *An introduction to the theory of Point Processes*. Springer. [Cited on page 13.]
- [18] Dowd, C. (2022). *twosamples: Fast Permutation Based Two Sample Tests*. R package version 2.0.0. [Cited on page 20.]
- [19] Engmann, S. and Cousineau, D. (2011). Comparing distributions: The two-sample anderson-darling test as an alternative to the kolmogorov-smirnov test. *Journal of applied quantitative methods*, pages 1–17. [Cited on page 18.]
- [20] Ferro, C. and Stephenson, D. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting - WEATHER FORECAST*, 26. [Cited on page 54.]
- [21] FIRST (2022). *Forum of Independent Report and Security Teams*. <https://www.first.org/>[Accessed : March 2022]. [Cited on page 5.]

- [22] Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190. [Cited on page 26.]
- [23] Frei, S., May, M., Fiedler, U., and Plattner, B. (2006). B.: Large-scale vulnerability analysis. *Association for Computing Machinery*, pages 1–9. [Cited on page 12.]
- [24] Gilleland, E. and Katz, R. W. (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, 72(8). [Cited on page 20.]
- [25] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d’une serie aleatoire. *Annals of Mathematics*, 44(3). [Cited on page 26.]
- [26] Guillou, A., Kratz, M., and Le Strat, Y. (2010). An extreme value theory approach for the early detection of time clusters with application to the surveillance of salmonella. *Statistics in medicine*, 33(28), pages 5015–5027. [Cited on page 42.]
- [27] Guillou, A., Naveau, P., Diebolt, J., and Ribereau, P. (2009). Return level bounds for discrete and continuous random variables. *TEST*, 18:584–604. [Cited on pages 40 and 41.]
- [28] J.McNeil, A., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press. [Cited on page 25.]
- [29] Joh, H. (2011). *Quantitative Analysis of Software Vulnerabilities*. PhD thesis, Colorado State University. [Cited on pages 2 and 8.]
- [30] Kingman, J. F. C. (1993). *Poisson Processes*, volume 3. The Clarendon Press Oxford University Press. [Cited on pages 13 and 15.]
- [31] Landwehr, J., Matalas, N., and Wallis, J. (1979). Probability weighted moments compared with some traditional techniques in estimating gumbel parameters and quantiles. *Water Resources Research*, 15. [Cited on page 40.]
- [32] Movahedi, Y., Cukier, M., and Gashi, I. (2020). Predicting the discovery pattern of publically known exploited vulnerabilities. *IEEE Transactions on Dependable and Secure Computing*, PP:1–10. [Cited on page 8.]
- [33] M.R.Leadbetter, Lindgren, G., and Rootzén, H. (2012). *Extremes and related properties of random sequences and processes*. Springer New York, NY. [Cited on pages 25, 28, 32, and 40.]

- [34] NIST (2022). *National Institute of Standards and Technology*. <https://www.nist.gov/>[Accessed : March 2022]. [Cited on page 3.]
- [35] NVD (2022). *Nation Vulnerability Database*. <https://nvd.nist.gov/>[Accessed : March 2022]. [Cited on page 8.]
- [36] Pokhrel, N. R., Khanal, N., Tsokos, C., and Pokhrel, K. (2020). Cybersecurity: a predictive analytical model for software vulnerability discovery process. *Journal of Cyber Security Technology*, 5. [Cited on pages 2 and 8.]
- [37] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [Cited on pages 20 and 42.]
- [38] Rajasooriya, S., Tsokos, C., and Kaluarachchi, P. (2016). Stochastic modelling of vulnerability life cycle and security risk evaluation. *Journal of Information Security*, 07:269–279. [Cited on page 8.]
- [39] Ribatet, M. and Dutang, C. (2022). *POT: Generalized Pareto Distribution and Peaks Over Threshold*. R package version 1.1-10. [Cited on page 20.]
- [40] Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury. [Cited on page 17.]
- [41] Shahzad, M., Shafiq, M., and Liu, A. (2012). A large scale exploratory analysis of software vulnerability life cycles. In *Proceedings - International Conference on Software Engineering*, pages 771–781. [Cited on page 3.]
- [42] Shukla, A., Katt, B., and Nweke, L. (2019). Vulnerability discovery modelling with vulnerability severity. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6. [Cited on pages 8 and 19.]
- [43] Stat Counter Global Stats (2022). *Operating system market share worldwide*. <https://gs.statcounter.com/os-market-share>[Accessed : March 2022]. [Cited on page 11.]
- [44] Stoffer, D. and Poison, N. (2022). *astsa: Applied Statistical Time Series Analysis*. R package version 1.15. [Cited on page 20.]
- [45] Vickers, A. J., Van Calster, B., and Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352. [Cited on page 47.]



- 
- [46] WFE (2022). *Global risks report 2022*. <https://www.weforum.org/reports/global-risks-report-2022>[Accessed : March 2022]. [Cited on page 1.]
- [47] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. [Cited on page 20.]
- [48] Wickham, H., François, R., Henry, L., and Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9. [Cited on page 20.]