

**TITLE**

Subtitle

Ana Clara do Carmo St. Aubyn

Dissertation presented as partial requirement for obtaining the Master’s degree in Advanced Analytics

MAA

Title:

Subtitle:

Ana Clara do Carmo St. Aubyn

2021





**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa

TITLE

by

Ana Clara do Carmo St. Aubyn

Dissertation presented as partial requirement for obtaining the Master’s degree in Advanced Analytics

**Advisor / Co Advisor:**Mauro Castelli and Maria Jordão

Month 2021

Dedication (optional)

Acknowledgements (optional)

Abstract

Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text Abstract text.

Keywords

Keyword 1; Keyword 2; Keyword 3; Keyword 4

INDEX

[1. Introduction 1](#_Toc69496393)

[1.1. Thesis Objective and Research Questions 1](#_Toc69496394)

[1.2. The Evolution of Domestic Violence in Portugal 2](#_Toc69496395)

[2. Literature Review 4](#_Toc69496396)

[3. Theoretical Background 6](#_Toc69496397)

[3.1. Panel Data 6](#_Toc69496398)

[3.2. Causal Relationships and *Ceteris Paribus* 6](#_Toc69496399)

[4. Data Exploration 8](#_Toc69496400)

[4.1. Dependent Variable 8](#_Toc69496401)

[4.2. Explanatory Variables 10](#_Toc69496402)

[5. Methodology 11](#_Toc69496403)

[6. Results and Discussion 12](#_Toc69496404)

[7. Conclusions 13](#_Toc69496405)

[8. Limitations and Recommendations for Future Works 14](#_Toc69496406)

[9. Bibliography 15](#_Toc69496407)

[10. Appendix (optional) 16](#_Toc69496408)

[11. Annexes (optional) 17](#_Toc69496409)

List of Figures

[Figure 1.1. Domestic Violence Occurrences (Portugal) 2](#_Toc69496384)

[Figure 1.2. Domestic Violence Occurrences by Category (Portugal) 3](#_Toc69496385)

[Figure 1.3. Domestic Violence Against Spouse or Analogous Occurrences (Portugal) 3](#_Toc69496386)

[Figure 2.1 - Age of Domestic Violence Victims (2013-2017) 4](#_Toc69496387)

[Figure 4.1. Absolute Change in Total and DVASA Occurrences (Portugal) 9](#_Toc69496388)

LIST OF TABLES

Table 4.1. Descriptive Statistics for the Dependent Variable (Portugal) 7

Table 4.2. Missing Values by Category (Municipalities) 8

Table 4.3. Missing Values for DVASA in Municipalities and Difference Between National Total and Municipality Total 9

List of Abbreviations and Acronyms

**APAV** Associação Portuguesa de Apoio à Vítima

**DGPJ** Direção-Geral da Política de Justiça

**DVASA** Domestic Violence Against Spouse or Analogous

**DVAM** Domestic Violence Against Minors

**OMA** Observatório de Mulheres Assassinadas da UMAR

# 

# Introduction

Domestic violence is a widely discussed issue. According to Portuguese news agencies, the number of victims seems to be rising each year. Given this, it certainly is of the utmost importance to identify and address the causes of this problem. It is also true that the public in general is increasingly aware of the reality about domestic violence and this topic is becoming more relevant in the official media channels.

According to the yearly report[[1]](#footnote-1) published by OMA – *Observatório de Mulheres Assassinadas da UMAR* – in 2017, during the same year there were 20 murders related to domestic violence in Portugal. Besides that, there were 28 cases of domestic violence that were considered attempted murders. The report[[2]](#footnote-2) from the following year states that the number of domestic violence related murders increased by 8, turning the reported number of murders related to domestic violence in 2018 into 28. The number of deaths related to domestic violence in 2019[[3]](#footnote-3) was even higher than in the previous years, with 31 registered deaths. It is important to acknowledge that throughout the year of 2019 there were a total of 89 willful murders registered, as stated by the official statistics provided by *Direção-Geral da Política de Justiça* (DGPJ). This number makes the previous number regarding murders related to domestic violence much more relevant and contextualized. The official report[[4]](#footnote-4) about murder victims in 2019, released by APAV – *Associação Portuguesa de Apoio à Vítima* in June 2020, also stated that, regarding the 44 willful murders that they followed, 48% of those were caused by domestic violence. **These numbers lift the veil on the sad reality Portugal is facing and clarify the need for addressing this problem**.

## Thesis Objective and Research Questions

The present project aims to develop a model that allows an understanding of the causes of domestic violence in Portugal and explains while quantifying the effect of each explanatory variable in the number of domestic violence occurrences. Taking into account the available data and the characteristics of it, the application of a panel data regression was selected as a viable solution in order to achieve the main goal.

During this research, the importance of several possible causes for domestic violence was tested. Since the objective is to explain the number of occurrences as well as possible modifications to some variables were considered during the process as well as alternative models.

Keeping this in mind, the present dissertation proposes to answer the following questions:

* How did the number of domestic violence occurrences in Portugal evolve between 2008 and 2019?
* How well can panel data regression explain this evolution?
* What are the main causes of domestic violence?
* How does each explanatory variable affect the number of domestic violence occurrences?

## The Evolution of Domestic Violence in Portugal

In order to have a better understanding of the numbers, one can take a look at Figure 1.1. This figure represents the total number of domestic violence occurrences registered by police authorities in Portugal by year in the time interval between 2008 and 2019. From the figure we can see a rise in the number of occurrences between 2008 and 2010, followed by a decrease from 2010 to 2012. In the period between 2012 and 2018 the number of occurrences remained relatively stable. However, one can witness a new rise from 2018 to 2019. **The key point of this study is to find the causes for this evolution and explain their influence in the number of domestic violence occurrences.**



Figure 1.1. Domestic Violence Occurrences (Portugal)

In Portugal, under Article 152 of the Criminal Code, an aggression is categorized as domestic violence if the aggressor, repeatedly or not, inflicts physical or psychological ill-treatment, including physical punishment, deprivations of freedom and sexual offenses to:

* A spouse or ex-spouse.
* Someone from either the same or any other gender with whom the aggressor keeps or has kept a relationship analogous to that of spouses, even if without cohabitation.
* A parent of common offspring in first degree.
* Someone who is particularly helpless, possibly because of age, disability, illness, pregnancy or economic dependency with whom the aggressor cohabits.

Having this in mind, one can identify three categories of domestic violence as shown in the official statistics provided by DGPJ. The first one, which is called **domestic violence against spouse or analogous** includes all the topics mentioned above except for the last one. The second one, which is called **domestic violence against minors**, includes all aggressions to minors in which the aggressor cohabits with the victim. Finally, the last category, which is called **others**, includes the last of the topics mentioned above except for cases of domestic violence against minors. Figure 1.2 is a breakdown of Figure 1, splitting the total occurrences into these three categories. It becomes clear that most of the domestic violence occurrences in Portugal fit into the first category – domestic violence against spouse or analogous – represented in the figure by the green line. The minimum value for this category was 20394 in 2008, whilst the maximum value was 25129 in 2010. The second most prominent category is others, with a maximum value of 4651 in 2011 and a minimum value of 3083 in 2008. Finally, the least represented category is domestic violence against minors, with a maximum value of 680 in 2008 and a minimum value of 430 in 2017.



Figure 1.2. Domestic Violence Occurrences by Category (Portugal)

Figure 1.3. allows one to take a closer look at the evolution of the number of domestic violence against spouse or analogous occurrences and notice that not only it represents the majority of domestic violence occurrences in Portugal, as it follows the pattern detected for the total occurrences described in Figure 1.1. This is the main category for domestic violence occurrences in Portugal and will also be the one used as a dependent variable during the course of this study.



Figure 1.3. Domestic Violence Against Spouse or Analogous Occurrences (Portugal)

# Literature Review

An article[[5]](#footnote-5) by Ellsberg et al., written in 2001, compares three studies on domestic violence in Nicaragua. Two of them are focused on urban areas of the country (León and Managua) and the remaining one is a national-wide Demographic and Health Survey that included other themes besides domestic violence. All of them are interview-based studies. When comparing the results of the studies, the authors of the article come to the conclusion that **domestic violence occurrences tend to be underestimated when the source relies on self-reporting**. This underestimation is not random, as it depends on numerous factors such as the number of individuals present in the room or the way the questions are asked. The article also mentions that some groups seem to be more at risk for domestic violence than others. It seems like **women with more children are more prone to suffer assaults**. Also, **younger women were found to be at higher risk of violence**.

According to an APAV report[[6]](#footnote-6) published in 2018 that studies the characteristics of victims of domestic violence in Portugal between 2013 and 2017, during this period this organization registered a total of 36.528 support processes in cases of domestic violence. In 31.317 of these processes, the victim was female (representing 85,73% of the total). Following the same line of thought, in 32.134 of the processes the author of the crime was male (representing 85,93% of the total). **This suggests that it might be useful to include some measure of the gender structure of the population as an explanatory variable for the present study**. It is also mentioned in the same report that 41% of the victims had ages comprehended between 26 and 55 years. However, in another report[[7]](#footnote-7) published by APAV also in 2018 that focuses on cases of domestic violence with male victims, it is mentioned that the victim’s age group with higher frequency is the 65 years or above one, representing around 28% of the processes contemplated by the report. Figure 2.1. below shows the percentage of processes for each age group (from 18 years and to 65+) both in the general report (23.193 processes) and in the male only report (2.745 processes). One can see that the patterns for each report are different, **showing the importance of including the age structure of the population by gender in the present study**.



Figure 2.1 - Age of Domestic Violence Victims (2013-2017)

Another key finding from the general APAV report is related to the family structure of the victims. Once again, it is shown that **children are an important factor when determining the risk for domestic violence**. From the 36.528 processes considered, 41,86% of the victims were in a nuclear family with children. Finally, it is also mentioned in the report that the marital status of the victim is also a relevant factor, as around 34% of the victims were married. This is a relevant percentage when compared to the 20,8% that were single, 16% whose marital status was unknown, 11,6 who were in a non-marital partnership, 8,7% who were divorced, 5,6% who were separated and 3,3% who were widowed. Therefore, **including a measure of the number of marriages may be relevant**.

As explained in the 2010 article[[8]](#footnote-8) entitled *The Gender Wage Gap and Domestic Violence*, written by Anna Aizer, **there are several theories regarding wage gender inequality and domestic violence against women**. The one supported b0y this article states that, as the wage gap between genders decreases, women get more bargaining power and, consequently, domestic violence decreases as well. However, there are other hypotheses possible. The first one is the “male backlash” hypothesis which states that, as the wage gap decreases, violence increases against women because aggressive men feel as if their traditional gender role may be threatened. The other hypothesis, the model of exposure reduction, states that as the wage gap decreases, the labor force participation of women increases and, consequently, domestic violence against them declines because women spend less time with violent partners. Either way, **it is unquestionable that this is a variable of interest when it comes to justifying changes in domestic violence occurrences**. One of the key points of this article is that the relative or potential salary is more important in justifying violence patterns than the actual one. Keeping this in mind, if possible, it is better to use a variable that reflects the potential wage of women vs. men instead of the actual wage gap. The results of this study, conducted in the state of California, in the United States, show that the decline in the wage gap between 1990 and 2003 explains nine percent of the decrease of domestic violence against women in the same period.

# Theoretical Background

## Panel Data

Data can be collected in multiple formats. The most widely discussed ones are cross section data, pooled cross section data, time series data and panel data.

**Cross section** data is data that is collected for multiple units across the same time period. Each observation represents a unit of the population being studied. It is the “common” dataset structure. When we combine cross section data from different time periods we create a **pooled cross section** dataset. In this case, each observation represents a unit of the population in a specific period in time. It is not necessarily true that the same units are studied for the different periods. If we are studying the same unit across different periods in time, we create a **time series**. A time series shows the evolution of that unit through a specified time interval. Finally, **panel data**, also called **longitudinal data**, is a combination of cross section and time series data. In this type of data, we have one time series for each included unit. The identifier of the unit and the period the data refers to are shown as variables in the dataset. The present study focuses on panel data, as there is one yearly discrete time series for each Portuguese municipality.

## Causal Relationships and *Ceteris Paribus*

The goal in regression analysis is to find **causal relationships** between variables, that is, to determine whether a change in *x* causes a change in *y*. To express our ideas regarding the relationships between variables we use functions. For example, to express a relationship between consumption and income one can write:

The previous function is a possible notation to say that the consumption of a certain individual is a function of his income. This is the same as saying the individuals’ consumption depends on his income. However, the consumption may not only depend on income, it may depend on many other factors as well. Keeping this in mind, to understand the relationship between consumption and income it is important to set aside the impacts of the remaining factors on consumption. The idea of ***ceteris paribus*** (c.p.) means to hold all other factors constant and is a key point in establishing causal relationships. Without holding the remaining variables constant one does not prove that the change observed in *y* is caused by the change in *x*. This is also the reason why a simple correlation study is not enough to analyze causal relationships.

When we are studying the causal effect of *x* on *y*, the remaining variables that influence *y* are called the **control variables**. The reason to control these variables is simple: we believe that *x* is correlated with other factors influencing *y*, which means that not holding these variables constant will make their effects reflect on the coefficient for *x*. Since we feed data to the regression model, it is important to correctly determine the control variables that need to be held fixed. This is a critical part of regression analysis but may be hard, as usually not all factors influencing the dependent variable are observable. When one does not include an important control variable, its effects are reflected on the partial effects of other factors, making these last ones incorrect.

Stating the difference between explanatory variables and control variables may be hard. Even if some variables can be considered control variables in all occasions, all explanatory variables are eventually control variables when trying to explain the partial effects of another variable. For panel data, the variables that determine the difference between observations (in the case of this study: municipality and year) are always control variables as, even if they can explain part of the variance in the data, their main goal is to distinguish between observations.

# Data Exploration

## Dependent Variable

Three datasets containing information regarding the dependent variable were retrieved from the official Statistics website by DGPJ on the 4th of March of 2021. One containing information about the number of domestic violence occurrences nationwide, one with this data split by districts and a last one with the data split by municipalities. All of them contained information regarding three categories (as explained in 1.Introduction): **domestic violence against spouse or analogous (DVASA)**, **domestic violence against minors (DVAM)** and **others**. Finally, for all three datasets, the data was collected for the period between 2008 and 2019, due to data availability.

The Portuguese Criminal Code provides for and punishes the crime of domestic violence. Domestic violence assumes the nature of a public crime, which means that the criminal procedure is not dependent on a complaint by the victim, just a complaint or knowledge of the crime is enough for the Public Ministry to promote the process. Keeping this in mind, the registered number of occurrences of domestic violence in Portugal do not depend only on self-report by the victim. However, as it is a crime that commonly takes place in the privacy of a home, many cases may depend on self-report. The dataset obtained focuses on data registered by police authorities and, according to (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001), may suffer from underreporting, as it depends on self-report to some extent.

The data recorded for Portugal as a whole is a discrete time series. For each category there is a set of 12 observations recorded at uniformly spaced time values, in this case, years. This remains true for the data regarding districts and municipalities, except that for the first case there is one time series per category and per district, and for the second case there is one time series per category and per municipality.

The evolution of the number of domestic violence occurrences in Portugal can be seen in figures 1.1 and 1.2. It becomes clear by the analysis of these figures and of the descriptive statistics on Table 4.1 that domestic violence against spouse or analogous is the most prominent category out of the three. One can see that between 2008 and 2019, the yearly average of domestic violence occurrences was 27394. Considering the same period, the yearly average for the DVASA category was 22977,8, a value that clearly shows how relevant this category is for the total domestic violence occurrences. The remaining categories have less significant yearly averages.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DVASA** | **DVAM** | **Others** | **Total Occurrences** |
| **Std** | 1226,32 | 75,22 | 435,55 | 1612,36 |
| **Minimum** | 20394,00 | 430,00 | 3083,00 | 24157,00 |
| **Mean** | 22977,80 | 537,92 | 3879,17 | 27394,90 |
| **Maximum** | 25129,00 | 680,00 | 4651,00 | 30340,00 |
| **Q3** | 23382,80 | 599,00 | 4039,00 | 27877,00 |
| **Median** | 22851,50 | 515,50 | 3800,00 | 27155,00 |
| **Q1** | 22457,50 | 484,00 | 3647,75 | 26683,50 |

Table 4.1. Descriptive Statistics for the Dependent Variable (Portugal)

When comparing the evolution of total occurrences in Portugal (Figure 1.1) with the evolution of occurrences for DVASA (Figure 1.3), one can detect the same patterns. By calculating the difference in the number of occurrences for subsequent years, it is possible to notice that DVASA occurrences almost always justify over half of the growth or decrease in the number of total occurrences. This is only not true for the transition between 2014 and 2015, when the number of DVASA occurrences increased by 35 but the number of total domestic violence occurrences decreased by 48 due to a decrement in the other categories. Figure 4.1 shows exactly this.



Figure 4.1. Absolute Change in Total and DVASA Occurrences (Portugal)

Since the datasets only have 12 years worth of data, it would not be possible to perform a time series regression, as there would not be enough degrees of freedom to provide powerful estimates. Keeping this in mind, a panel data regression will be performed with data regarding the years and municipalities. For this purpose, the dataset containing information about domestic violence occurrences by municipality must be analyzed.

Portugal is divided into **18 districts** and **2 autonomous regions**. Each of these are subdivided into municipalities. Currently, Portugal has **308 municipalities**. The municipality data retrieved from the official Statistics website by DGPJ measured the three domestic violence categories for the 308 Portuguese municipalities and for an extra N.E. one, meaning not specified (*não especificado* in Portuguese). Since it would not be possible to find the explanatory variables values for this special case, this extra municipality was eliminated from the dataset. Furthermore, 12 of the 308 municipalities did not have values for all the categories. Corvo, the smallest island in the Autonomous Region of the Azores, only had data for the DVASA category. The remaining 11 municipalities (Pampilhosa da Serra, Golegã, Ribeira de Pena, Vila de Rei, Barrancos, Vila Viçosa, Penela, Alcoutim, Alfândega da Fé, São Roque do Pico e Aguiar da Beira) were missing data for the DVAM category.

The statistical confidentiality principle is declared in Diário da República (the Portuguese official gazette) in Law nº22/2008, the law that legislates about the National Statistical System. This principle, referred to in article 6 of the mentioned law, aims to safeguard citizens' privacy and guarantee trust in the Statistical System. Therefore, in the retrieved datasets, numbers below 3 are not presented, being symbolized as missing values.

Keeping this in mind, the number of missing values was calculated for each category. As one can see from Table 4.2, there were a total of 4356 missing values among the three categories. The majority of these can be found in the second and third categories. The missing values for DVASA represent only around 2.3% of the total missing values in the dependent variable dataset.

|  |  |
| --- | --- |
| Category | Number of Missing Values |
| DVASA | 100 |
| DVAM | 2862 |
| Others | 1394 |
| Total | 4356 |

Table 4.2. Missing Values by Category (Municipalities)

As mentioned before and seen on Figure 4.1, DVASA is the most prominent category in the total domestic violence occurrences in Portugal. Adding this to the facts that it is also the category with the least missing values (Table 4.2) and that it is the only category measured for all 308 municipalities, one can conclude that this is the best dependent variable for this study.

In order to better understand the missing values and to find the best way to impute them, the difference between the national values for each year and the sum of the values for each municipality in each year was calculated (including the values for N.E.). This can be seen on Table 4.3. One can see that the number of missing values for the municipalities in each year is always very close to the number of occurrences reported on the national level but unreported on a municipal level. Keeping this in mind, these missing values were replaced by the value 1 as I find it better to keep information about municipalities with low occurrences than to remove them from the study.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
| Missing Values | 21 | 16 | 6 | 7 | 5 | 7 | 9 | 8 | 4 | 6 | 4 | 7 |
| Difference | 27 | 21 | 8 | 7 | 6 | 10 | 7 | 12 | 4 | 10 | 3 | 10 |

Table 4.3. Missing Values for DVASA in Municipalities and Difference Between National Total and Municipality Total

## Explanatory Variables

According to (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001) high parity is a risk factor for domestic violence. This means that **individuals belonging to families with more children are more prone to suffer assaults**. As a way to include this factor in the present study, the **synthetic fertility index** was considered as an explanatory variable. This index is the average number of children born for each woman in fertile age (between 15 and 49 years). In order for the generation renewal to be assured, the synthetic fertility index must be at 2,1. The data regarding this variable was retrieved from the Pordata website on the 15th of April of 2021.

# Methodology

Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology.

# Results and Discussion

Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion.

# Conclusions

Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text.

# Limitations and Recommendations for Future Works

Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works.

# Bibliography

APAV - Associação Portuguesa de Apoio à Vítima. (2020). *Vítimas de Homicído Relatório APAV 2019.* Lisbon: Associação Portuguesa de Apoio à Vítima.

Correia, A. M. (2019, March 9). Nem mais uma: os números da violência doméstica que nos envergonham. *Expresso*.

Dantas, M. (2019, December 29). Número de mortes por violência doméstica sobe para 35 só este ano. *Público*.

# Appendix (optional)

# Annexes (optional)

1. (Brasil, Alves, & Soares, Dados 2017, 2018) [↑](#footnote-ref-1)
2. (Brasil, Alves, & Soares, Dados 2018, 2019) [↑](#footnote-ref-2)
3. (Soares, Branco, & Alves, 2020) [↑](#footnote-ref-3)
4. (APAV - Associação Portuguesa de Apoio à Vítima, 2020) [↑](#footnote-ref-4)
5. (Ellsberg, Heise, Peña, Agurto, & Winkvist, 2001) [↑](#footnote-ref-5)
6. (APAV - Associação Portuguesa de Apoio à Vítima, 2018) [↑](#footnote-ref-6)
7. (APAV - Associação Portuguesa de Apoio à Vítima, 2018) [↑](#footnote-ref-7)
8. (Aizer, 2010) [↑](#footnote-ref-8)