**Fireproofing The Future: Next-Gen Wildfire Prediction**

A DISSERTATION SUBMITTED TO

SVKM’S NMIMS (DEEMED TO BE UNIVERSITY)

IN PARTIAL FULFILLMENT FOR THE DEGREE OF

**MASTERS OF SCIENCE**

**IN**

**STATISTICS AND DATA SCIENCE**

BY

**Tanvi Rane 86062300013**

**Advait Raut 86062300022**

**Shreya Sane 86062300009**

**Parth Sante 86062300067**

**Neha T Mathew 86062300018**

UNDER THE SUPERVISION OF

Dr Pradnya Khandeparkar

NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS AND ANALYTICS

SVKM’s Narsee Monjee Institute of Management Studies

(Deemed-To-Be-University)

V.L. Mehta Rd, Vile Parle (West), Mumbai – 400056

November 2023

INDEX

[1 Abstract: 3](#_Toc151582908)

[2 Introduction: 4](#_Toc151582909)

[3 Rationale: 5](#_Toc151582910)

[4 Objectives: 6](#_Toc151582911)

[5 Data Preparation 7](#_Toc151582912)

[5.1 About The Data: 7](#_Toc151582913)

[5.2 Data Cleaning: 7](#_Toc151582914)

[6 Methodology 9](#_Toc151582915)

[Models 9](#_Toc151582916)

[Model-based statistical analysis 11](#_Toc151582917)

[7 Results & Discussion 24](#_Toc151582918)

[8 Summary & Conclusion 26](#_Toc151582919)

[9 References 27](#_Toc151582920)

# Abstract:

In recent decades, the nature of wildfires has shifted, necessitating accurate predictive models on a national scale for efficient firefighting resource allocation. This change is notable in Mediterranean countries, where wildfires are frequent, concentrated mainly in summer. Due to seasonality, certain territories experience zero fires in some months and heightened occurrences in others. Zero-inflated negative binomial mixed models suit this data type, capturing both fire frequency and non-occurrence. Additionally, a parametric bootstrap method enhances predictions by estimating mean squared errors and constructing prediction intervals. Applying this statistical methodology and custom software, we model and predict wildfires in Spain from 2002 to 2015 by provinces and months. This approach aids in resource optimization and firefighting strategy planning, with broader implications for regions with similar seasonal patterns.

# Introduction:

Wildfires, often referred to as forest fires or bushfires, represent a formidable natural force that has shaped ecosystems and landscapes for millennia. These catastrophic events, characterized by uncontrolled and rapid combustion of vegetation, have significant ecological, environmental, and socio-economic impacts. As climate change intensifies and human activity encroaches upon wildland areas, the frequency and severity of wildfires are increasing worldwide. Understanding the dynamics, causes, and consequences of wildfires is paramount for both their mitigation and management.

Various countries have developed sophisticated statistical techniques and predictive models to anticipate and manage wildfires. These techniques often rely on factors like weather patterns, vegetation health, historical wildfire data, and human activity. While the methods vary from one nation to another, they share the common goal of providing timely warnings and information to prevent or mitigate the devastating effects of wildfires. In this paper, we are interested in digging deeper into how Spain predicts and manages wildfires in their country.

Spain is one of the European countries particularly vulnerable to wildfires due to its climate, terrain, and extensive forested areas. The combination of dry, hot summers and a Mediterranean climate makes many regions of Spain susceptible to wildfires. These fires pose significant threats to both human settlements and natural ecosystems, causing substantial ecological and economic damage. In recent years, Spain has experienced an increase in the frequency and severity of wildfires, which has prompted the development of advanced statistical models for wildfire prediction and management.

To understand and model the occurrence of wildfires, it is crucial to explore probability distributions. One such distribution is the negative binomial distribution. The negative binomial distribution is a [discrete probability distribution](https://en.wikipedia.org/wiki/Discrete_probability_distribution) that models the number of failures in a sequence of independent and identically distributed [Bernoulli trials](https://en.wikipedia.org/wiki/Bernoulli_trial) before a specified (non-random) number of successes (denoted ) occurs.

It has a [probability mass function](https://en.wikipedia.org/wiki/Probability_mass_function):



where *r* is a real, positive number.

In the context of wildfires, we will investigate the role of the zero-inflated negative binomial distribution, which combines the negative binomial distribution with the logit distribution to aid in predicting their occurrence and managing their impact. Zero-inflated models are particularly useful for count data that exhibit overdispersion and excess zeros. We will explore how this distribution plays a crucial role in capturing the complex dynamics of wildfires, especially in Spain, where these natural disasters pose significant threats to both human settlements and natural ecosystems

# Rationale:

Wildfire prediction is crucial globally because it allows for proactive resource allocation and emergency preparedness. Accurate predictions enable authorities to mobilize firefighting resources strategically, reducing the impact on ecosystems, human lives, and property. Predictive models at different spatial and temporal scales are needed to support the planning of a new firefighting system that prioritizes the places and times of greatest risk.

In the case of Spain, where wildfires are prevalent, prediction is especially vital due to the country's unique geographical and climatic conditions. Spain's susceptibility to seasonal wildfires demands precise forecasting to optimize resource distribution, enhance firefighting effectiveness, and minimize the socio-economic and environmental consequences associated with these events. Additionally, Spain's diverse landscape and varying wildfire patterns underscore the need for localized predictions to address the specific challenges posed by different regions and seasons.

# Objectives:

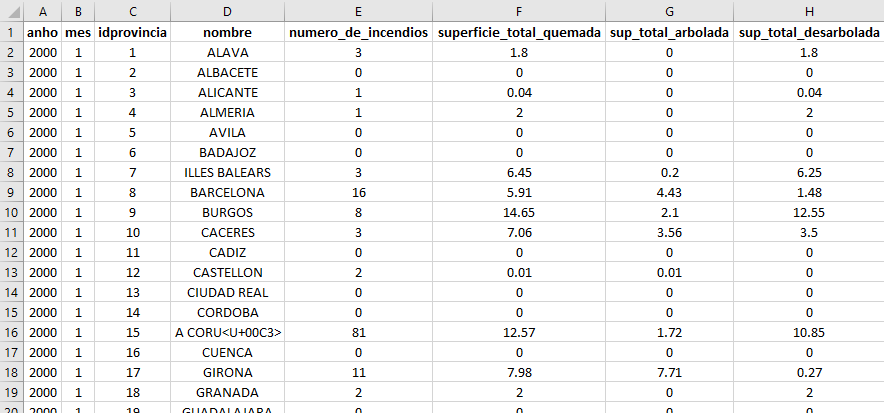
1. To study the application of zero-inflated negative binomial mixed model in wildfire prediction
2. To build a model for the prediction of wildfires in Spain
3. To analyse the advantage of using zero-inflated negative binomial mixed models over the Poisson mixed model. (overdispersion)

# Data Preparation

## About The Data:

The following meteorological raw data provided by the Spain Meteorological Agency (AEMET) and data related to unemployment rates, provided by the Spanish Statistical Office (INE) have been taken from github.

Spain Raw Dataset:



1. There are a total of 58 variables in our raw dataset.
2. There are a total of 9600 rows.

## Data Cleaning:

The raw data has many blank spaces so we need to do data cleaning. There are many variables for which most of the data are -9999 or NA.

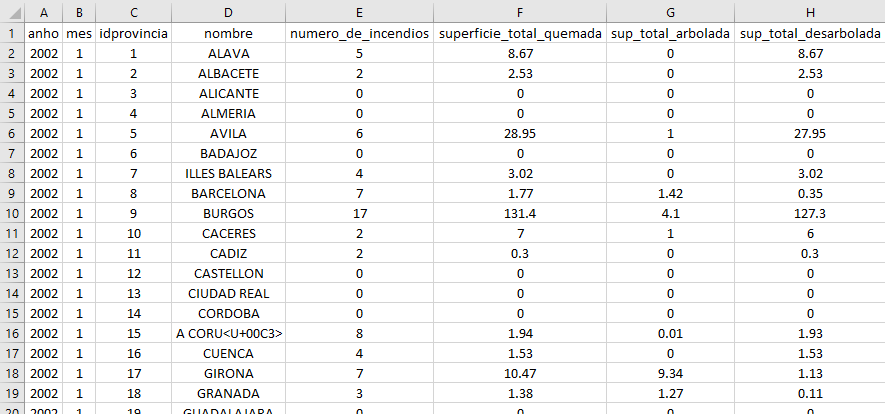
1. When there are many -9999 or NA (more than 250), these variables are deleted by using the following R code:

datosv2 = datos %>% select(-c(evap, glo, inso, n\_cub, n\_des, n\_nub, nv\_0050, nv\_0100, nv\_1000,p\_sol, ts\_10, ts\_20, ts\_50, w\_rec))

1. When there are few -9999 or NA, they are substituted by the median value of the remaining values and then converted into the percentage (rate) of days.
2. As we only have the unemployment rate from 2002 onwards, we will only consider data from 2002 onwards (anho=year)

datosv2 = datosv2[datosv2$anho>2001,]

Spain Prepared Dataset:



1. There are a total of 44 variables in our newly prepared dataset.
2. There are a total of 8400 rows.

The dependent variable 𝑦ijk counts the number of Spanish wildfires in year 𝑖, month 𝑗 and province 𝑘. Therefore, there were 𝐷 = 𝐼𝐽𝐾 = 8400 domains, defined by the crosses of years (𝐼= 14), months (𝐽 = 12)and provinces (𝐾 = 50).

# Methodology

## Models

Now we study the zero-inflated negative binomial mixed model applied in the data analysis of this study.



· *Yijk:* no. of wildfires in *ith* year, *j*th month and *k*th province, where *i* ∈ I = {1*,* …*, I*}, *j* ∈ J = {1*,* …*, J*} and *k* ∈ K = {1*,* …*, K*}. Let *D= IJK* be the total number of *𝑦*-values.

• Let 𝑧𝑖𝑗𝑘, 𝒙1,𝑖𝑗𝑘 = (𝑥1,𝑖𝑗𝑘1, …, 𝑥1,𝑖𝑗𝑘𝑞1 ) and 𝒙2,𝑖𝑗𝑘 = (𝑥2,𝑖𝑗𝑘1, …, 𝑥2,𝑖𝑗𝑘𝑞2), 𝑖 ∈ I, 𝑗 ∈ J, 𝑘 ∈ K, be latent (non observable) variables and 1 × 𝑞1 and 1 × 𝑞2 row vectors containing some explanatory variables, respectively.

· Zijk = 1; when there is no occurrence of wildfire in *ith* year, *j*th month and *k*th

province.

= 0; when there is wildfire in *ith* year, *j*th month and *k*th

province.

· Zijk ~ Ber (Pijk), where Pijk: Probability of no occurrence of wildfire in in *ith year*, *j*th month and *k*th province.

· P (Yijk = 0| Zijk = 1) = 1, since there will be no occurrence of wildfire, Yijk will always be zero.

· So, the variable of interest would be Yijk | Zijk=0,

Since the overdispersion can be overcome using a NB mixed model,

Y𝑖𝑗𝑘|Z𝑖𝑗𝑘=0 ∼ NB (r, µijk),

P (Yij𝑘 = t| Z𝑖𝑗𝑘 = 0) = ^t ^r , t ∈ N ∪ {0}

where, Pijk ∈ (0,1), r > 0, µijk > 0.

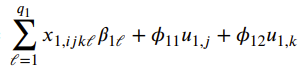
In addition, Pijkand µijk depend on the explanatory variables given by,

1. 𝒙1, ijk and 𝒙2, ijk,

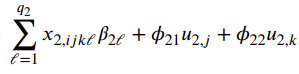
2. on the regression parameters 𝜷1 = (𝛽11, …, 𝛽1𝑞1) ′ and 𝜷2 = (𝛽21, …, 𝛽2𝑞2) ′

3. and on the standard deviation parameters 𝜙11 > 0, 𝜙12 > 0, 𝜙21 > 0 and 𝜙22 > 0 by means of the link functions.

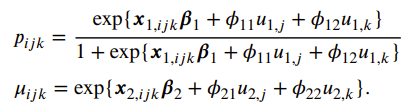
logit (Pijk)= log= 𝒙1,𝑖𝑗𝑘𝜷1 + 𝜙11𝑢1,𝑗 + 𝜙12𝑢1,𝑘



log(µijk) = 𝒙2,𝑖𝑗𝑘𝜷2 + 𝜙21𝑢2,𝑗 + 𝜙22𝑢2,𝑘



Conversely, for 𝑖 ∈ I, 𝑗 ∈ J, 𝑘 ∈ K, we have



To complete the definition of the AZINB11 model, we assume that the vectors (Yijk, Zijk) ′, 𝑖 ∈ I, 𝑗 ∈ J, 𝑘 ∈ K, are independent conditional on **𝒖**.

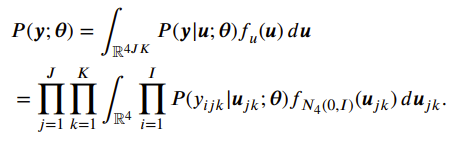
Let 𝜽1 = (𝜷′1, 𝜙11, 𝜙12) ′, 𝜽2 = (𝜷′2, 𝜙21, 𝜙22) ′, 𝜽 = (𝜽′1, 𝜽′2) ′ be the vectors of model parameters and define 𝜉𝑖𝑗𝑘 = 𝐼 {0} (y𝑖𝑗𝑘), 𝑖 ∈ I, 𝑗 ∈ J, 𝑘 ∈ K. From the properties of the NB, it holds that

P (Y𝑖𝑗𝑘|𝒖𝑗𝑘; 𝜽) = 𝜉ijk [𝑝𝑖𝑗𝑘 + (1 − 𝑝𝑖𝑗𝑘) exp {𝑟 log 𝑟 − 𝑟 log (𝑟 + 𝜇𝑖)}] + (1 − 𝜉𝑖𝑗𝑘) [(1 − 𝑝𝑖𝑗𝑘) exp {𝑦𝑖𝑗𝑘 log 𝜇𝑖𝑗𝑘 − (𝑦𝑖𝑗𝑘 + 𝑟) log (𝑟 + 𝜇𝑖𝑗𝑘) + 𝑟 log 𝑟+ log } ]

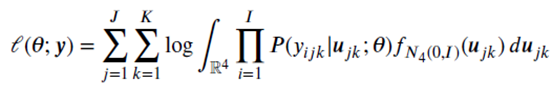
From the assumption of independence, we have

𝑃 (𝒚|𝒖; 𝜽) =P (𝒚jk|𝒖jk; 𝜽) ; P (𝒚jk|𝒖jk; 𝜽) =P (𝒚ijk|𝒖jk; 𝜽)

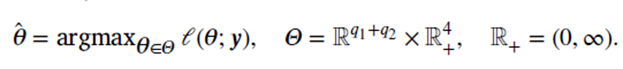
Therefore, the likelihood function of the AZINB11 model is



The log-likelihood function of the AZINB11 model is



Given 𝒚, the maximum likelihood (ML) estimator of 𝜽 is



Further, we use ML-Laplace approximation algorithm to maximize 𝓁(𝜽; 𝒚) and to calculate the ML estimators of the model parameters.

## Model-based statistical analysis

We apply the AZINB11 mixed model outlined to the fire data. The model incorporates month-dependent random effects (u1,j and u2,j for j ∈ J) and province-dependent random effects (u1,k and u2,k for k ∈ K). The model, consisting of two submodels, includes a BE mixed model with q1 = 6 covariables and an NB mixed model with q2 = 18 covariables.

* The covariables for the BE mixed model (BE-covariables) are:

x1,1 = intercept, x1,2 = hr, x1,3 = np.300, x1,4 = ta.max, x1,5 = year 3.2, x1,6 = year 3.3.

* The NB mixed model covariables (NB-covariables) consist of:

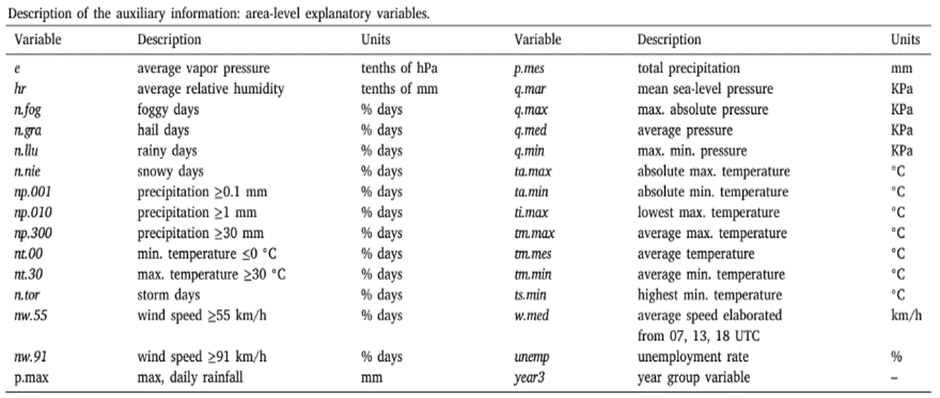
x2,1 = intercept, x2,2 = e, x2,3 = hr, x2,4 = n.llu, x2,5 = n.nie, x2,6 = np.300,

x2,7 = nt.00, x2,8 = nw.55, x2,9 = nw.91, x2,10 = q.mar, x2,11 = q.max, x2,12 = q.min,

x2,13 = ta.max, x2,14 = ta.min, x2,15 = tm.mes, x2,16 = tm.min, x2,17 = year 3.2,

x2,18 = year 3.3.

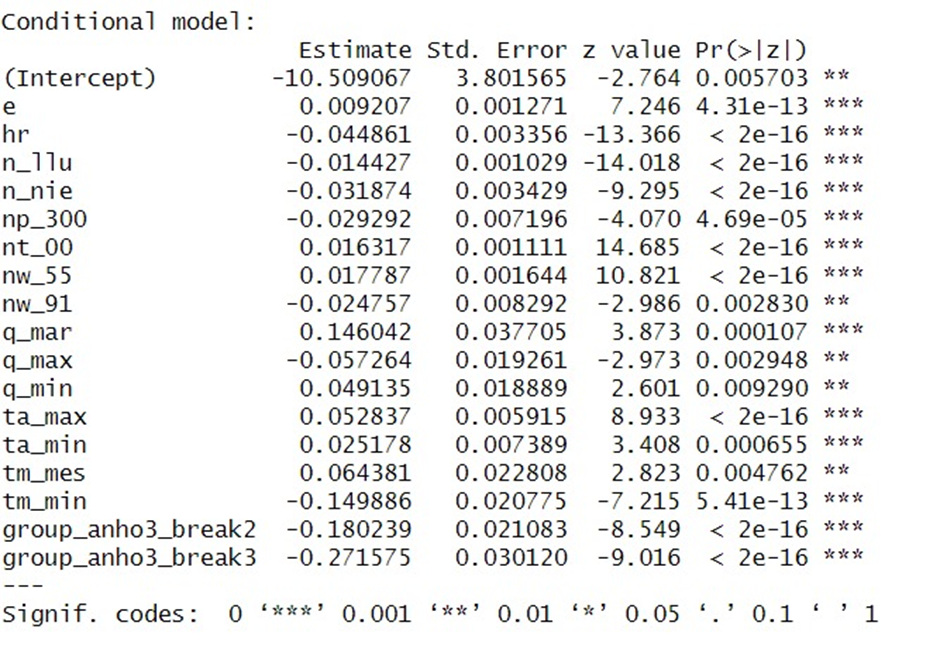
The AZINB11 model is fitted to the data spanning the years 2002–2014, with the exclusion of 2015 to evaluate the performance of future predictions based on the model. The table below summarizes the information about the area-level auxiliary variables.

  
  
  
  
  
**For NB mixed model:**

We test the hypotheses 𝐻0 ∶ 𝛽2,l = 0 vs 𝐻1 ∶ 𝛽2,l **≠** 0; 𝑙 ∈ {1, …, 18}

For checking the significance of the ML regressor parameters

Following two tables present estimates of the ML regression parameters 𝛽2, asymptotic standard errors (SE) and 𝑝-values.



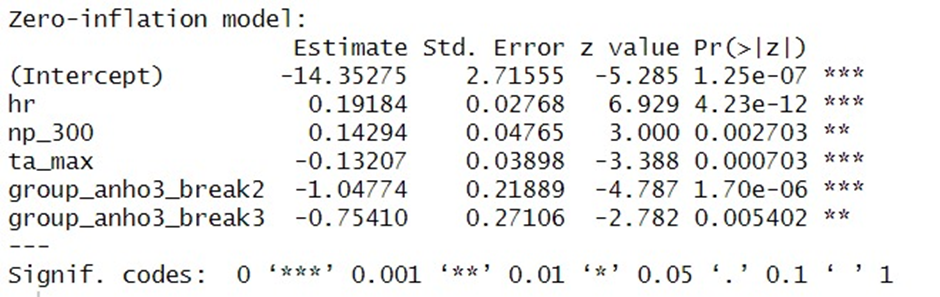
We can see that the variables such as an increase in humidity (*hr*), on general rainy days (*n.llu*) and on those whose precipitation is higher than 300 mm (*np.300*), on snowy days (*n.nie*) and on days with wind speed greater than or equal to 91 km/h (*nw.91*), contributed to reduce the number of wildfires, because their signs were significantly negative. The same happened with an increase in the maximum absolute pressure (*q.max*) and in the average minimum temperature (*tm.min*).

In addition, if a wildfire occurred in the time interval [2002, 2006], the intercept was negative value (−10.5090), which decreased by 0.1802 units when changing the event to [2007, 2012] and by 0.2716 when changing it to the last time interval, being significant with both modifications (*year3.2* and *year3.3*). Meanwhile, the variables contributing to the increase in the no. of wildfires are the ones with the positive estimate values, namely an increase in the average vapor pressure (*e*), on days with minimum temperatures below 0 ◦C (*nt.00*) and on days with wind speed greater than or equal to 55 km/h (*nw.55*)

* **For BE mixed model:**

We test the hypotheses 𝐻0 ∶ 𝛽1, l = 0 vs 𝐻1 ∶ 𝛽1, l **≠** 0; 𝑙 ∈ {1, …, 6}

Like the NB mixed model, the following table provides similar information displayed in the table above, viz estimates of ML regression parameter 𝛽1, standard errors and 𝑝-values.



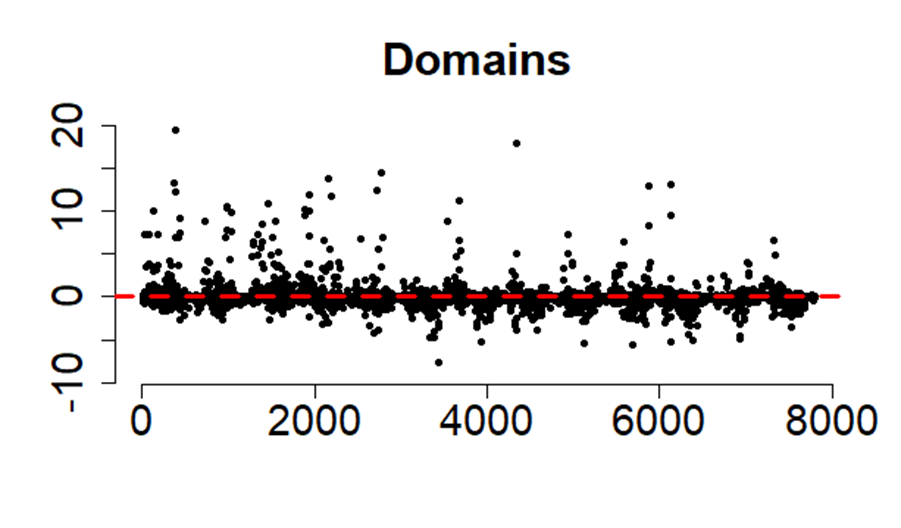
For the BE mixed model, above table provides interesting results related to the estimated zero inflated probability. Again, conditional on the model covariates, as the humidity (*hr*) and the percentage of days with precipitations greater

than 30 mm (*np.300*) increased, the zero-inflated probability also increased.

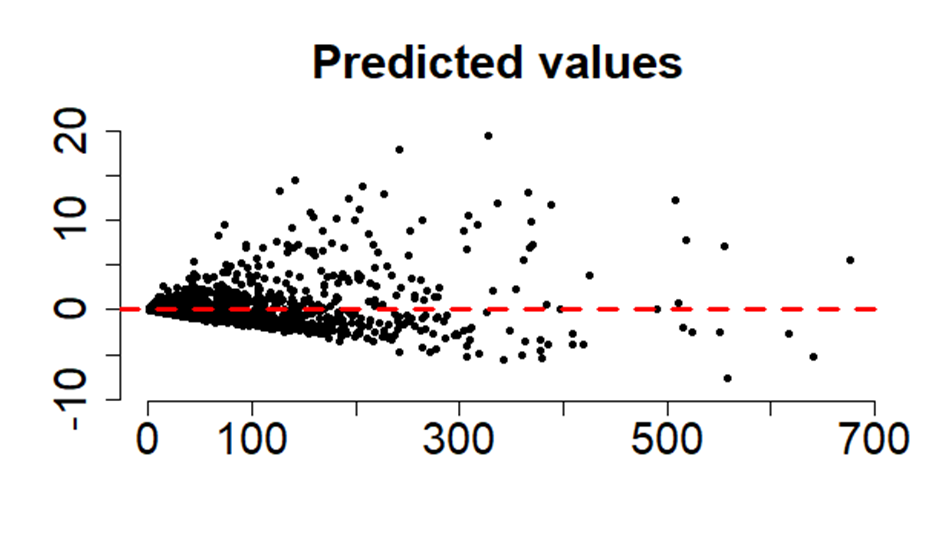
Not less important, an increase in the maximum temperature recorded (*ta.max*) decreased the zero-inflated probabilities.

**Outlier Detection:**

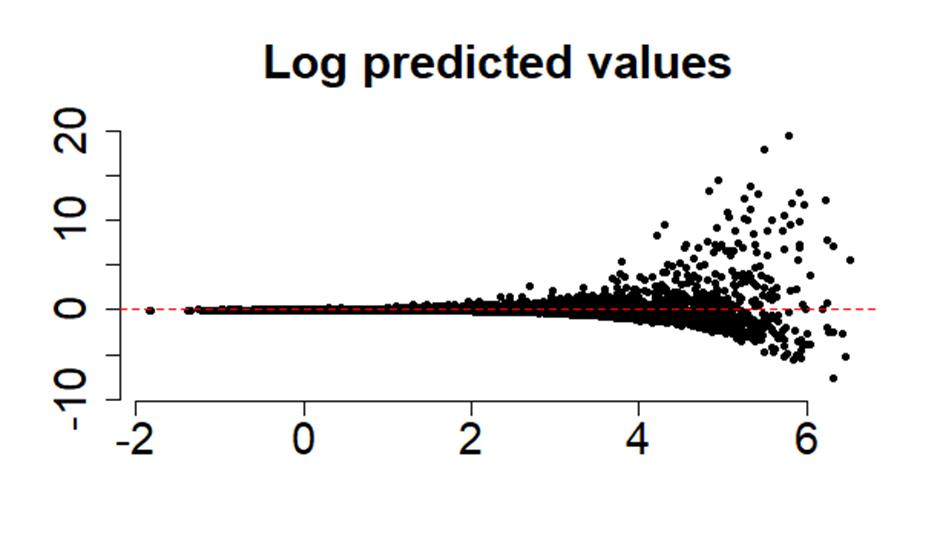
For validating the fitted model, we analysed the behaviour of model residuals.



In this graph, standardized residuals are plotted versus the domain indexes. The standardised residuals have fluctuated around zero. But there were more positive large residuals than negative ones. This is because there were some provinces where the number of observed wildfires in summer was extremely high, as we will see later.

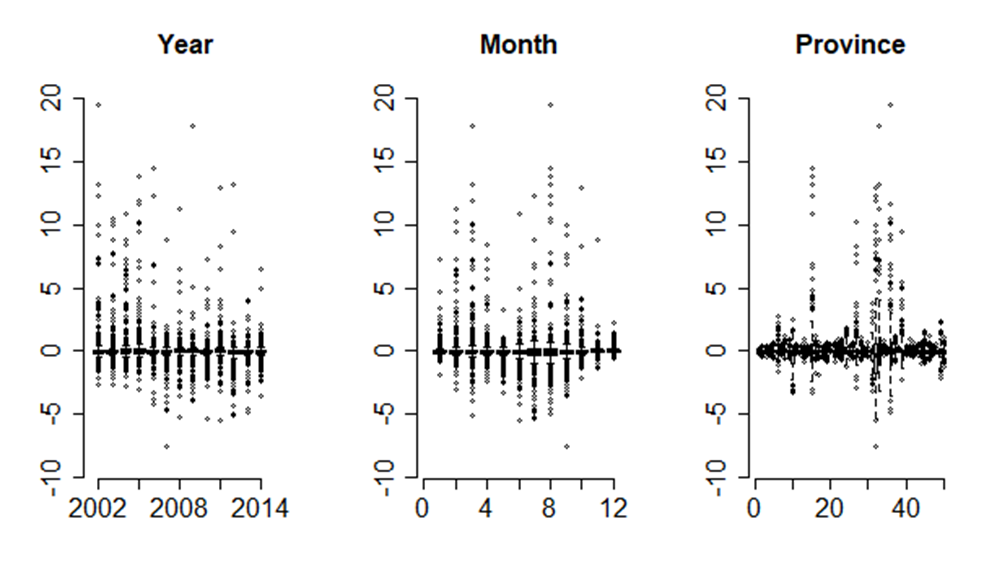


In this graph, standardized residuals are plotted versus predicted values. We saw something similar happened with a small percentage of domains with large predicted values exceeding the threshold of 400.



In this graph, standardized residuals are plotted versus log predicted values. We detect a conical pattern in the scatter plot maintaining a positive symmetry. As log predicted values increase, the variability of residuals also increases. This is in agreement with the overdispersion of the model.

To detect atypical data, we plot boxplots of the standardized residuals by year, month, and province.



In the first graph, we do not observe any pattern. The standardised residuals are evenly spread. As a result, this graph is not significant for outlier detection.

In the second graph, we do observe a slight pattern. Here, more atypical data falls in spring and summer. Hence, month seems more important than year.

In the third graph, the majority of standardized residuals are concentrated around a certain interval [-3,3]. It is clear that province is the most crucial factor in outlier detection.

There were six provinces with absolute standardized residuals greater than 3. These are our outliers which we will study later. These outliers were caused by the model’s underpredictions, because of the high number of unexpected wildfires reported and this produced some estimation problems that cannot be solved by incorporating new dummy variables, which will model whether or not the fire occurred in these provinces. These six provinces are:

1. A Coruña (18)

2. Lugo (5)

3. Ourense (22)

4. Pontevedra (17)

5. Asturias (14)

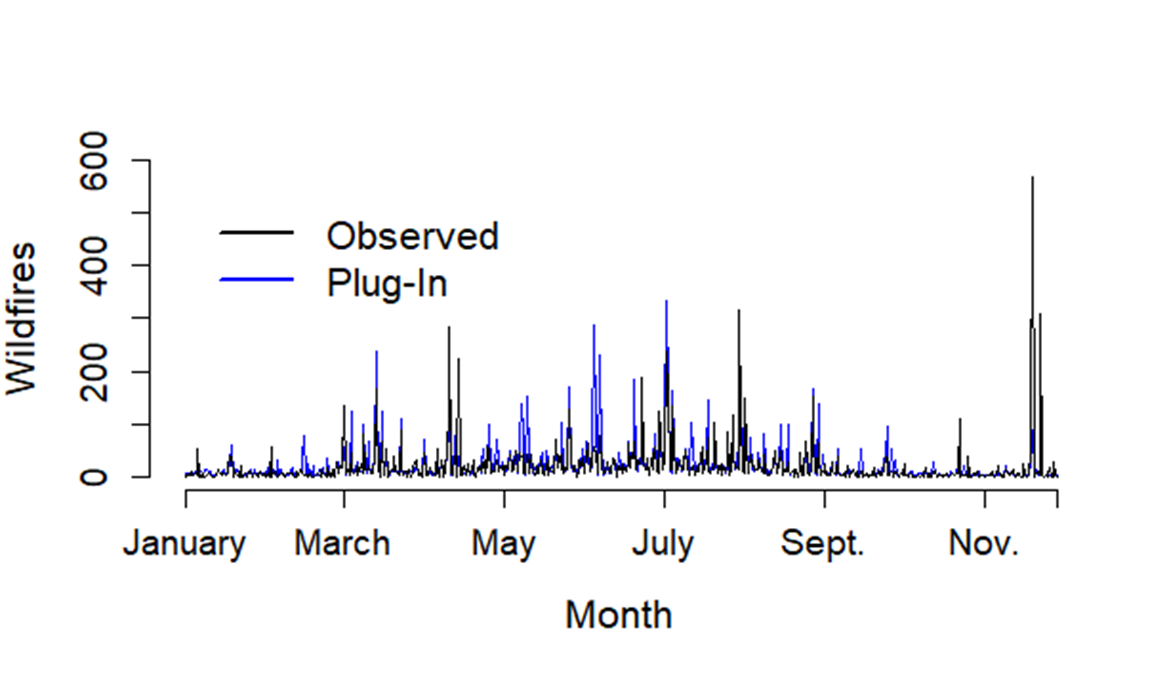
6. Cantabria (6)

**Therefore, we consider month + province randomness for the intercept and not year as it doesn’t show any pattern in its boxplot.**

**Wildfire forecasting:**

In this section, we will predict the number of wildfires in the year 2015. We will fit our AZINB11 model for the data from 2002 – 2014 and keep 2015 to check the performance of future model-based predictions.

First, we will see the observed and predicted values for all the months in 2015.



This graph plots the observed and predicted wildfires by provinces in the year 2015. As we already have the values of 2015, we can check the accuracy of our model. We can clearly see that our model represented the trend in the observed data with acceptable accuracy. However, there were some extreme values observed in December 2015 which our model couldn’t predict accurately. This is again due to the six north-western provinces registering an excessively high number of wildfires due to their unpredictable weather conditions.

We can also figure out that the majority of wildfires are seen in summer. The summer season in Spain is from July to September.

Observed and Predicted wildfires for the months of summer season:

JULY 2015:

AUGUST 2015:

SEPTEMBER 2015:

These are the observed and predicted values for the months of the summer season (July – September) 2015. We can say that our proposed model is fairly accurate in predicting the number of wildfires for the year 2015. But as discussed earlier, our model suffers from some outliers which are the six north-western conflicting provinces which refuse to follow our model. Due to their unpredictable climatic conditions, these provinces report an excessively high number of wildfires and also with very high variability. These six provinces are:

1. A Coruña

2. Lugo

3. Ourense

4. Pontevedra

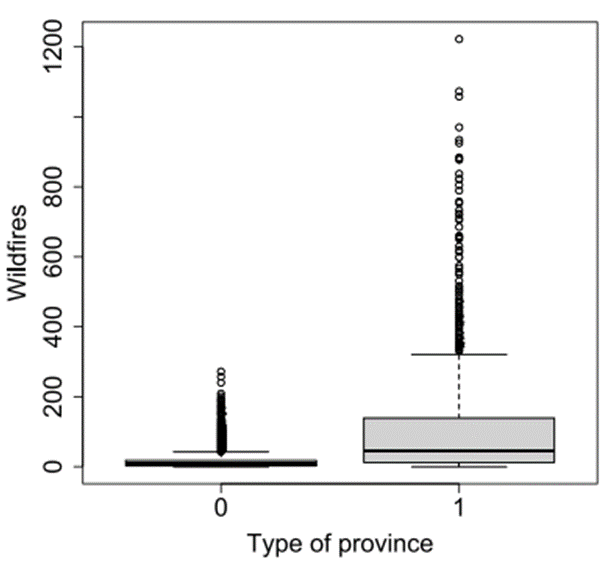
5. Asturias

6. Cantabria

# Results & Discussion

As we look through the Observed and predicted values for the months of summer (July – September) 2015,we can conclude that our proposed model is fairly accurate in predicting the number of wildfires for the year 2015. But as discussed earlier, our model suffers from some outliers which are the six north-western conflicting provinces which refuse to follow the model.

A data set of 8400 registers, 18+6=24 covariates, and 2+2=4 sources of variation for random effects was used to fit the chosen AZINB11 model to wildfire counts. As a result, we anticipate that there will be outliers in a specific percentage of domains where the model's predictions are not met. This issue mostly affects the six provinces in northwest Spain, where there are typically more fires and more unpredictability in the frequency of fires than throughout the nation. When we look at each of the six troublesome provinces independently, we find that 168 domains have seen more than 200 wildfires. A graphic representation of the variations in the number of wildfires across provinces maybe found in the box plot below.



Here 0 denotes it is not North West province

1 denotes it is a North West province

The research has demonstrated that the characteristics of forest fires are influenced by both spatial and temporal meteorological conditions. This finding aligns with the conclusions of Russo et al. (2017), who observed a strong connection between large summertime wildfire activity and preceding climatic conditions in the north-western sector of the Iberian Peninsula. Russo and colleagues predicted an increasing prevalence of extreme conditions in this region.

However, the study acknowledges that certain events may deviate from the predicted model, as noted by Boubeta et al. (2015) and Parente et al. (201fire8). An illustrative example is the winter fires of 2015 in the regions of Asturias and Cantabria. During this period, unusual southern winds, characterized as dry and hot, were observed in the weeks leading up to the fires—a phenomenon atypical for these colder months in this particular region (Calheiros et al., 2020). This deviation from the expected model highlights the complexity and variability of factors influencing fire behaviour, emphasizing the need for a nuanced understanding of the interplay between climatic conditions and wildfire occurrences. The suggested model proposes a division of Spain into two distinct fire behaviour zones. The first, situated in the northwest, experiences a higher frequency of wildfires due to structural reasons, as indicated by previous studies (Gómez-Vázquez et al., 2009; Marey-Pérez et al., 2010; López-Rodríguez et al., 2021). These reasons are often linked to forest management practices, leading to the development of various prediction models. In contrast to the rest of Spain, this region exhibits unique fire characteristics.

Notably, the predictions for 2015 revealed a discrepancy between the anticipated number of fires projected by the model and the actual observed occurrences. This aligns with findings from previous models, suggesting that arsonists strategically choose optimal conditions, creating a "window of opportunity" to initiate fires. This perspective is consistent with the argument put forth by Russo et al. (2017) that periods of preceding drought are essential for fire presence.

Furthermore, other researchers, such as Marcos et al. (2015) and Ying et al. (2021), have drawn similar conclusions through the analysis of ROC curves. Their results highlight the significance of a combination of low relative humidity and high air temperature in the occurrence of large fires, a pattern observed in both the Mediterranean region and China, mirroring the outcomes of our study.

# Summary & Conclusion

In light of the challenges posed by escalating wildfires in Mediterranean countries, the imperative for a paradigm shift in firefighting strategies becomes increasingly evident. The collaborative approach and methodology validated in Spain offer a beacon of hope in navigating the complexities of wildfire management. The successful implementation of zero-inflated negative binomial mixed models at the provincial level not only refines our understanding of fire behaviour but also provides a scalable framework adaptable to diverse climatological and socio-economic conditions.

The application of this methodology not only addresses the limitations of traditional planning systems but also optimizes the allocation of firefighting resources. By sidestepping the pitfalls associated with excessive zero values, the provincial focus allows for the integration of auxiliary variables from government databases, offering a holistic perspective crucial for effective decision-making.

The forecasting tool's ability to retrodict with 95% confidence empowers authorities to proactively implement prevention tasks. This not only enhances the resilience of communities but also ensures a more efficient and strategic deployment of resources during wildfire crises. The commitment to continuous improvement, as evidenced by the ongoing exploration of additional auxiliary variables, underscores the adaptability and forward-thinking nature of this approach.

As we look to the future, the refinement of predictive capabilities for extreme events emerges as a key research avenue. The incorporation of land use and socio-economic variables promises to enrich the model's predictive power, providing a more nuanced understanding of the underlying factors influencing wildfire occurrences. This extended conclusion underscores the transformative potential of the presented methodology, urging a collective re-evaluation of firefighting approaches in the face of escalating challenges in the Mediterranean region. The journey towards effective wildfire management demands sustained collaboration, innovative methodologies, and a commitment to staying ahead of the curve in anticipating and mitigating the impacts of these formidable natural disasters.

# References

Bugallo, María, et al. “Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain.” *Journal of Environmental Management*, vol. 328, 2023, p. 116788, https://doi.org/10.1016/j.jenvman.2022.116788.

https://github.com/mbugallo/aZINB11Fires.

https://en.wikipedia.org/wiki/Negative\_binomial\_distribution

<https://en.wikipedia.org/wiki/Zero-inflated_model>