Eötvös Loránd University

Faculty of Science

Modeling the impact of wildfire-related air pollution on mortality

Thesis

Author:

Noémi Takács

Applied Mathematics MSc
specialisation in stochastics

Supervisor:

dr. András Zempléni

associate professor
Department of Probability Theory and Statistics

Budapest, 2025

Acknowledgment

My thesis could not have been written without my supervisor, dr. András Zempléni, who, in addition to recommending the topic, helped me with many useful suggestions and comments.

I would also like to thank my family and my partner, as I can always count on their support and love during my years at university and in life.

Noémi Takács

Contents

1 Introduction

In recent years, we have heard more and more about devastating wildfires around the world, mainly due to climate change, among other factors such as human activity. When a disaster occurs, it can have a number of serious consequences depending on its spatial and temporal extent. Both flora and fauna (and other forms of life) can suffer major damage, but now I address the problem primarily due to its impact on humanity. Of course, material assets can also be affected by such a disaster, e.g. a fire near a settlement can easily spread to residences. However, obviously the most important thing is our health. Pollution in the air due to wildfires can also have direct and indirect negative effects on the human organism [3]. In my thesis, I examine the relationship between air pollution from wildfires and mortality.

At the beginning of the work, I discuss the theoretical background of the modeling method I use. I write about regressions and the generalized linear model, respectively, and specifically in the case when we have count data. In this case, a common phenomenon is the overdispersion of the data, where a solution may be to fit a quasi-likelihood model. Since the data are consecutive in temporal order, they should be treated as a time series, taking into account the correlation between adjacent observations. This brings us to the Quasi-Poisson time series regression.

In the third section, I present the characteristics of wildfire smoke. Which are the fine particles, which chemical compounds are present in the smoke, what are the factors that determine its composition, and how does it differ from, for example, air pollution from factories? Depending on their size and the weather conditions, the emitted substances can travel long distances and remain airborne even for weeks. These fine particles can also enter the human body, easily causing health issues. The respiratory system is primarily at risk, but cardiovascular problems can also develop. From minor to fatal, a wide range of diseases and symptoms can be associated with wildfire-related air pollution.

In the second part, I outline the modeling context. Some of the tasks were done in Python (e.g. the data preprocessing), while other parts were implemented in the programming language R (such as the modeling). I describe the area under investigation, the variables that I use for the studies and the different approaches that I have considered. These relate to the variables describing the smoke: examining several components together or only one type of substance, or considering a moving average or some lagged values individually.

Finally, I summarize my results, based on the different specifications. I discuss the findings for the entire population under study, then present the differences obtained from comparing the sexes, and also the age-specific studies.

2 Theoretical foundations

The books [13], [9], [14] and [12] were used to compose subsection 2.1, 2.1.1, 2.1.2, 2.1.2.1 and 2.1.2.2. For the parts 2.2 and 2.2.1, I used my university lecture notes and article [10].

2.1 Generalized Linear Models

To determine the mortality risk for a given time or duration, we also need information from other variables. For this problem, we can apply regression analysis, which aims to estimate the relationship between a dependent or target variable ($\mathbf{Y}$) and one or more independent or explanatory variables ($\mathbf{X_1}, \ldots, \mathbf{X_k}$), based on observed data. Hence, our observations are a set of input-output pairs $\{(\mathbf{X_i}, Y_i)\}_{i=1}^n$ in $\mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^k$ and $\mathbb{Y} \subseteq \mathbb{R}$. From now on, $\mathbf{X_j}$ denotes the j-th explanatory variable (column vector) and $\mathbf{X_i}$ the i-th observation (row vector).

Regression is distinguished between parametric and non-parametric. In the former, a finite set of parameters is used to define the model, while in the latter, the goal is to extract patterns and trends from the data, rather than determining functions from specific forms (such as smoothing techniques). In this thesis, I deal with parametric regression.

Definition 1 (Function form of the model)**.** *Parametric regression models can be written in the following form*

$$Y_i = f(\mathbf{X_i}, \beta) + \varepsilon_i,$$

where $\beta \in \mathbb{R}^d$ is the parameter vector, $f = E\left[Y|\mathbf{X}, \beta\right] : \mathbb{R}^{k+d} \to \mathbb{R}$ is the regression function, and ε_i is the noise term, which is independent of the variables and has an expected value of zero.

Definition 2 (Prediction)**.** *The value $\hat{Y_i} = f\left(\mathbf{X_i}, \hat{\beta}\right)$ is called the estimation of the dependent variable or the prediction given by $\mathbf{X_i}$, where $\hat{\beta}$ is an estimation of the parameter vector.*

Definition 3 (Residual)**.** *The difference between the dependent variable and its estimated value is called the residual: $\hat{\varepsilon}_i = Y_i - \hat{Y_i}$.*

Linear models are the most fundamental type of regression. In this case, the target variable is approximated by a linear combination of explanatory variables, i.e. the model takes the following form:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \mathrm{X}_{ij} + \varepsilon_i \quad \text{or briefly} \quad \mathbf{Y} = \mathrm{X}\beta + \varepsilon,$$

Hence,

$$\mathrm{E}[Y_i|\mathbf{X_i}] = \mu_i = \sum_{j=1}^{k} \beta_j \mathrm{X}_{ij} \quad \text{or briefly} \quad \mu = \mathrm{X}\beta.$$

One of the most common methods used to estimate coefficients is Ordinary Least Squares (OLS), which is based on the idea of minimizing the sum of squares of the residuals. However, this technique assumes that the error term and the response variable are continuous and normally distributed. This means that the method can be useful in other cases, but the properties of the estimates are not optimal, and the tests on them do not work well.

Generalized Linear Models (GLMs) are an extension of the linear regression framework, allowing the dependent variable to follow a distribution from the exponential family other than the Gaussian. This model class keeps the linearity of the coefficients, but enables to model a nonlinear relationship between the target variable and the explanatory variables. A generalized linear model has three main components:

- **Random component**: A distribution from the exponential family from which the response variable $\mathbf{Y}$ is derived.
- **Systematic component**: (Linear predictor) A linear combination of the independent variables, $\eta_i = \mathbf{X_i}\beta$.
- **Link function**: A monotone and differentiable function g that connects the linear predictor and the expected value of the dependent variable, $\mu_i = \mathrm{E}[Y_i|\mathbf{X_i}]$, that is $\eta_i = g(\mu_i)$.

In classical linear regression, the random component is the normal distribution and the link function is the identity function ($\eta = \mu$).

2.1.1 Parameter estimation

For GLM fitting, we use the maximum likelihood estimation (MLE) of the parameters. The traditional method for calculating this estimation in this model class is the Iteratively Reweighted Least Squares (IRLS) algorithm (used, for example, by the statistical software R).

To build the algorithm, we first need some properties of the exponential distribution family. A standard formula for the single-parameter density function:

$$f(Y_i, \theta_i) = \exp\{Y_i\theta_i - b(\theta_i) + c(Y_i)\}, \tag{1}$$

where θ_i is the canonical parameter or link function, here specifically $\theta_i = \eta_i = g(\mu_i)$, $b(\theta_i)$ is called the cumulant function and $c(Y_i)$ is the normalization term. The likelihood function:

$$\mathcal{L} = \prod_{i=1}^{n} \exp\{Y_i\theta_i - b(\theta_i) + c(Y_i)\} = \exp\left\{\sum_{i=1}^{n} (Y_i\theta_i - b(\theta_i) + c(Y_i))\right\},$$

and the log-likelihood function:

$$\log \mathcal{L} = \sum_{i=1}^{n} (Y_i\theta_i - b(\theta_i) + c(Y_i)) . \tag{2}$$

Claim 1. *Given a density function from the exponential family in the form 1. Suppose that f is differentiable at least twice with respect to each θ_i. Then,*

- $\mathrm{E}_{(\theta_i)}[Y_i] = \mu_i = b'(\theta_i)$
- $\mathrm{Var}_{(\theta_i)}[Y_i] = b''(\theta_i)$.

Proof. Expected value:
We know that

$$1 = \int_{\mathbb{R}} f(y_i, \theta_i)\, dy_i = \int_{\mathbb{R}} e^{y_i\theta_i - b(\theta_i) + c(y_i)}\, dy_i = e^{-b(\theta_i)} \cdot \int_{\mathbb{R}} e^{y_i\theta_i + c(y_i)}\, dy_i.$$

From this,

$$e^{b(\theta_i)} = \int_{\mathbb{R}} e^{y_i\theta_i + c(y_i)}\, dy_i.$$

Differentiated by θ_i:

$$e^{b(\theta_i)} \cdot b'(\theta_i) = \int_{\mathbb{R}} y_i \cdot e^{y_i\theta_i + c(y_i)}\, dy_i.$$

Reordering the equation, we got

$$b'(\theta_i) = \int_{\mathbb{R}} y_i \cdot e^{y_i\theta_i - b(\theta_i) + c(y_i)}\, dy_i. = \int_{\mathbb{R}} y_i \cdot f(y_i, \theta_i)\, dy_i,$$

which is $\mathrm{E}_{(\theta_i)}[Y_i]$, the expected value of Y_i with parameter θ_i.
Variance:
Now, let the second equation differentiate twice:

$$e^{b(\theta_i)} \cdot b'(\theta_i) \cdot b'(\theta_i) + e^{b(\theta_i)} \cdot b''(\theta_i) = \int_{\mathbb{R}} y_i \cdot y_i \cdot e^{y_i\theta_i + c(y_i)}\, dy_i.$$

Rearranging:

$$(b'(\theta_i))^2 + b''(\theta_i) = \int_{\mathbb{R}} (y_i)^2 \cdot e^{y_i\theta_i - b(\theta_i) + c(y_i)}\, dy_i = \mathrm{E}_{(\theta_i)}[(Y_i)^2].$$

Based on the previous calculation,

$$(b'(\theta_i))^2 = \left(\mathrm{E}_{(\theta_i)}[Y_i]\right)^2,$$

so we obtain that

$$b''(\theta_i) = \mathrm{E}_{(\theta_i)}[(Y_i)^2] - \left(\mathrm{E}_{(\theta_i)}[Y_i]\right)^2,$$

which is $\mathrm{Var}_{(\theta_i)}[Y_i]$, the variance of Y_i with parameter θ_i. □

The IRLS algorithm provides a general framework for finding the maximum likelihood estimate. It can be applied to various GLMs such as Poisson or logistic regression, where the coefficients do not have a closed form and cannot be computed directly. It is often used for robust regression, where the influence of outliers can be handled by putting less weight on larger residuals. The algorithm is also useful for various L_1 minimization problems where the solution cannot be computed directly due to the non-differentiability of the absolute value function. In addition to its wide applicability, other advantages include ease of implementation and efficient computation.

The derivation of the IRLS algorithm is based on a modification of a two-term Taylor expansion of the log-likelihood function of 1. The steps of the algorithm are shown in Table 1.

1. Initialize estimates of μ_i and η_i $(i = 1, \ldots, n)$ with a specific link function.

2.1 In the k-th step, compute working weights $\mathrm{W}^{(k)} = diag(w_i^{(k)})$, where

$$w_i^{(k)} = \frac{1}{\mathrm{Var}\left(\mu_i^{(k)}\right)\left(g'(\mu_i^{(k)})\right)^2}.$$

2.2 Compute working response $\mathbf{z}^{(k)}$, where

$$z_i^{(k)} = \eta_i^{(k)} + \left(Y_i - \mu_i^{(k)}\right) g'\left(\mu_i^{(k)}\right),$$

a one-term Taylor linearization of the log-likelihood function.

3. Calculate the updated value of the regression coefficients $\beta^{(k+1)}$ by the weighted least squares on $\mathbf{z} \sim \mathrm{X}$ with weights W:

$$\beta^{(k+1)} = \left(\mathrm{X}^\top \mathrm{W}^{(k)} \mathrm{X}\right)^{-1} \mathrm{X}^\top \mathrm{W}^{(k)} \mathbf{z}^{(k)}.$$

4. Estimate the new η and μ vectors.

5. Repeat steps 2, 3 and 4 until the change in the values of the parameters is below a selected threshold or tolerance parameter.

Table 1: The IRLS algorithm

Claim 2. *The IRLS algorithm gives the ML estimates of the parameter β for generalized linear models.*

Proof. (Sketch) From the log-likelihood function (2) the score function is obtained by the chain rule $\forall j \in [k]$:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial (\log \mathcal{L})_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial (\log \mathcal{L})_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

The derivatives per term:

$$\frac{\partial(\log\mathcal{L})_i}{\partial\theta_i} = Y_i - b'(\theta_i) = Y_i - \mu_i,$$

where the first bullet of claim 1 was used.

$$\frac{\partial\theta_i}{\partial\mu_i} = \left(\frac{\partial\mu_i}{\partial\theta_i}\right)^{-1} = \mathrm{Var}^{-1}(Y_i),$$

using again claim 1. The rest two terms:

$$\frac{\partial\mu_i}{\partial\eta_i} = -\frac{1}{g'(\eta_i)^2}, \quad \frac{\partial\eta_i}{\partial\beta_j} = \mathrm{X}_{ij}.$$

In summary:

$$\frac{\partial\log\mathcal{L}}{\partial\beta_j} = (Y_i - \mu_i)\cdot\mathrm{Var}^{-1}(Y_i)\cdot\left(-\frac{1}{g'(\eta_i)^2}\right)\cdot\mathrm{X}_{ij}.$$

Let $\mathrm{W}^{-1} = \mathrm{Var}(Y_i)g'(\eta_i)^2$, then the MLE method aims to solve the following equation with respect to β_j:

$$\sum_{i=1}^{n}\mathrm{W}(Y_i - \mu_i)\mathrm{X}_{ij} = 0,$$

which leads to a linear weighted least squares problem. In the algorithm z_i is the first-order linearized form of the link function η_i. Iteration is necessary because, in general $\mathbf{z}$ and W depend on β. □

Hypothesis testing

Once we have calculated the ML estimates of the parameters, we can test their significance, i.e., whether the parameters are equal to 0 or not. A usual method is to perform a Wald-type test, which takes the following form:

$$H_0 : \beta_j = 0,$$
$$H_1 : \beta_j \neq 0.$$

To test the hypothesis, we need the asymptotic distribution of the estimate.

Claim 3. *Let $\hat{\beta}$ denote the ML estimate of the parameter vector. Then*

$$\hat{\beta} \sim N_k(\beta, \mathcal{I}^{-1}(\hat{\beta})),$$

where $\mathcal{I}(\hat{\beta})$ is the Fisher information.

Fisher information is equal to the variance of the score function, which is the negative expectation of the second derivative of the log-likelihood function. Therefore,

$$\mathcal{I}(\beta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\beta^2}\log\mathcal{L}(\beta)\;\middle|\;\beta\right].$$

In practice, this is approximated by the amount of information observed, that is

$$\mathcal{I}(\beta) \approx \mathcal{I}_n(\hat{\beta}) = -\frac{\partial^2}{\partial \beta^2} \log \mathcal{L}(\hat{\beta}).$$

The test statistic is the estimated coefficient divided by its standard error:

$$\mathbf{z} = \frac{\hat{\beta}}{SE(\hat{\beta})},$$

which follows standard normal distribution for H_0 and sufficiently large n. From here, we can easily calculate the critical values and confidence intervals. If p-values are lower than the significance level, it means that the explanatory variable is significantly related to the target variable.

2.1.2 Modeling count data

If we want to model count data, such as the number of deaths in a given period, a common method is Poisson regression. The Poisson distribution, which is discrete with a single parameter λ, can be defined as:

$$\mathbb{P}(Y_i = y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \tag{3}$$

where $Y_i \in \mathbb{N}$ is the count variable. The Poisson distribution has the property that its mean and variance are equal to its parameter, so $\mathrm{E}[Y_i] = \mathrm{Var}[Y_i] = \lambda_i$.

Rewriting equation 3 as

$$\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \exp\left\{ y_i \cdot \log(\lambda_i) - \lambda_i + \log\left(\frac{1}{y_i!}\right) \right\}$$

we can immediately see that in expression 1 for the Poisson distribution $\theta_i = \log(\lambda_i)$, so $\lambda_i = e^{\theta_i}$, $b(\theta_i) = \lambda_i$ and $c(y_i) = \log\left(\frac{1}{y_i!}\right)$. Using claim 1, we find that

$$\mathrm{E}[Y_i|\mathbf{X_i}] = \mu_i = b'(\theta_i) = \frac{d}{d\theta_i}\lambda_i = \left(e^{\theta_i}\right)' = e^{\theta_i} = \lambda_i \tag{4}$$

and

$$\mathrm{Var}[Y_i|\mathbf{X_i}] = b''(\theta_i) = \frac{d^2}{d\theta_i^2}\lambda_i = \left(e^{\theta_i}\right)'' = e^{\theta_i} = \lambda_i. \tag{5}$$

A typical choice of the link function in this case is the log-link function because it guarantees that the target variable cannot be negative. The log-link function:

$$g(\mu_i) = \log(\mu_i) = \mathbf{X_i}\beta = \eta_i \tag{6}$$

or rewritten to express the expected value, μ_i:

$$\mu_i = \exp(\mathbf{X_i}\beta) = f(\mathbf{X_i}, \beta). \tag{7}$$

This means in terms of parameter interpretation that if the value of X_j, the j-th explanatory variable changes by one unit, it means an $\exp(\beta_j)$ multiplier change in the response variable in the same direction.

Using equations 4 and 6 we can rewrite equation 3 as

$$\mathbb{P}(Y_i|\mathbf{X_i}) = \frac{(\mathrm{e}^{\mathbf{X_i}\beta})^{y_i} \cdot \mathrm{e}^{-\mathrm{e}^{\mathbf{X_i}\beta}}}{y_i!}.$$

From this the likelihood function is:

$$\mathcal{L} = \prod_{i=1}^{n} \frac{(\mathrm{e}^{\mathbf{X_i}\beta})^{y_i} \cdot \mathrm{e}^{-\mathrm{e}^{\mathbf{X_i}\beta}}}{y_i!} = \mathrm{e}^{\sum_{i=1}^{n} \mathbf{X_i}\beta y_i} + \mathrm{e}^{-\sum_{i=1}^{n} \mathrm{e}^{\mathbf{X_i}\beta}} + \prod_{i=1}^{n} \frac{1}{y_i!},$$

and the log-likelihood is:

$$\log \mathcal{L} = \sum_{i=1}^{n} \left(\mathbf{X_i}\beta y_i - \mathrm{e}^{\mathbf{X_i}\beta} + \log \frac{1}{y_i!} \right).$$

Taking the derivatives respect to β, we got

$$\frac{\partial \log \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} \left(\mathbf{X_i}^\top y_i - \mathbf{X_i}^\top \mathrm{e}^{\mathbf{X_i}\beta} \right) = \sum_{i=1}^{n} \mathbf{X_i}^\top \left(y_i - \mathrm{e}^{\mathbf{X_i}\beta} \right).$$

Therefore the MLE of β is the solution of

$$\sum_{i=1}^{n} \mathbf{X_i}^\top \left(y_i - \mathrm{e}^{\mathbf{X_i}\beta} \right) = 0, \tag{8}$$

which can be derived by the IRLS method.

As we have seen, the reliability of a MLE can be determined by the Fisher information. For a Poisson distributed sample with n observation it is

$$\mathcal{I}_n(\beta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \beta^2} \sum_{i=1}^{n} \left(\mathbf{X_i}\beta Y_i - \mathrm{e}^{\mathbf{X_i}\beta} + \log \frac{1}{Y_i!} \right) \right] = -\mathrm{E}\left[-\sum_{i=1}^{n} \mathbf{X_i}^\top \mathbf{X_i} \mathrm{e}^{\mathbf{X_i}\beta} \right] = \sum_{i=1}^{n} \mathbf{X_i}^\top \mathbf{X_i} \mu_i.$$

Rewritten in matrix form:

$$\mathcal{I}_n(\beta) = \mathrm{X}^\top \mathrm{M} \mathrm{X}, \tag{9}$$

where $\mathrm{M} = diag(\mu_1, \ldots, \mu_n)$.

2.1.2.1 Presence of overdispersion

In real datasets, the equality of mean and variance, called the equidispersion criterion, is often not satisfied, and overdispersion occurs. This means that $\sigma_i^2 > \mu_i$ (of course, the reverse is also possible, i.e. underdispersion, but it is less common). Overdispersion can cause the model to underestimate standard errors, making hypothesis testing on the significance of the independent variables unreliable. Biased p-values may even lead to misinterpretation of the effect of a non-significant predictor. However, if we are interested only in the estimated values of the target variable and not in the effects of the individual explanatory variables, then there may not be a problem with overdispersion.

A model using a generalized Poisson distribution ([16]) can be applied when the expected value and the variance are the same, but also for underdispersed and overdispersed count data. This distribution is defined with an additional parameter compared to the Poisson, its probability mass function is of the following form:

$$f(y_i, \lambda_i, \delta) = \frac{\lambda_i(\lambda_i + \delta y_i)^{y_i - 1} \mathrm{e}^{-\lambda_i - \delta y_i}}{y_i!}, \tag{10}$$

where $\max\{-1, -\lambda_i/4\} < \delta < 1$. The expected value and the variance of such a random variable are

$$\begin{aligned} \mathrm{E}[Y_i] &= \frac{\lambda_i}{1-\delta}, \\ \mathrm{Var}[Y_i] &= \frac{\lambda_i}{(1-\delta)^3} = \frac{1}{(1-\delta)^2}\mathrm{E}[Y_i] = \varphi \mathrm{E}[Y_i], \end{aligned}$$

where $\varphi = \frac{1}{(1-\delta)^2}$ can be interpreted as a dispersion factor. When $\delta = 0$, $\varphi = 1$, the data is equidispersed, and from 10 the Poisson distribution can be obtained. However, in the case of $\delta < 0$, we got $\varphi < 1$, which means underdispersion. Similarly if $\delta > 0$ and thus $\varphi > 1$, then the data is overdispersed.

Definition 4 (Pearson χ^2 dispersion statistic)**.** *For Poisson models, an indicator of overdispersion is the Pearson χ^2 dispersion statistic, which is the squared residuals weighted by the model variance and divided by the residual degrees of freedom:*

$$\hat{\phi} = \frac{\chi^2}{df}, \quad \chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{\mathrm{Var}(\hat{Y}_i)},$$

where df is the residual degrees of freedom and $\mathrm{Var}(\hat{Y}_i) = \mathrm{Var}[Y_i|\mathbf{X_i}]$.

If $\hat{\phi} = 1$, then there is no extra variance, a value greater than 1 means over- and less than 1 indicates underdispersion. However, whether the deviation from 1 is significant depends on its size and the number of sample elements. One can perform a chi-squared test to check this. Under the null hypothesis, which assumes that the data are equidispersed, the Pearson chi-square statistic χ^2 from definition 4 follows chi-squared distribution with $n - k$ degrees of freedom, where k is the number of explanatory variables.

There are also some other alternative statistical tests, such as the score test. It supposes that the variance can be written as $\mathrm{Var}[Y_i|\mathbf{X_i}] = \mu_i + \sigma^2\mu_i^2$ (like in the negative binomial distribution), and tests whether $\sigma^2 = 0$ in the following form:

$$\begin{aligned} H_0 &= \text{no overdispersion in the model} \\ H_1 &= \text{the model is overdispersed.} \end{aligned}$$

The test statistic can be calculated as

$$z = \frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2 - Y_i}{\sqrt{2}\sum_{i=1}^{n}\mu_i}.$$

It is assumed that z is t-distributed with $n-1$ degrees of freedom. If the p-value is less than a prefixed significance level (usually 0.05), then we have to reject the null hypothesis, as the present amount of overdispersion is not negligible. This test is post hoc, since we first have to build a Poisson model and only then can we predict the vector μ and calculate the value of z.

2.1.2.2 Quasi-likelihood model

When we have overdispersed data, quasi-likelihood methods can be used, which assume some other relationship between the mean and the variance. The two most common types of assumptions on them are $\mathrm{Var}[Y_i] = \phi\mu_i$ and $\mathrm{Var}[Y_i] = \mu_i + \sigma^2\mu_i^2$, where ϕ and σ^2 are unknown scale parameters.

Suppose that $\mathrm{Var}[Y_i] = V(\mu_i, \phi)$, a function of the mean, and a dispersion parameter, which is assumed to be constant, hence it does not depend on the β parameter vector. Introduce a notation for the following function, for a single observation:

$$U(\mu, y) = \frac{y - \mu}{V(\mu, \phi)}.$$

This function has the following properties, which are common with the derivative of a log-likelihood function:

- $\mathrm{E}[U] = 0$
- $\mathrm{Var}[U] = \dfrac{1}{V(\mu, \phi)}$
- $\mathrm{E}\left[\dfrac{\partial U}{\partial \mu}\right] = \dfrac{1}{V(\mu, \phi)}.$

Based on these characteristics, if it exists, the following integral should behave to some extent like a log-likelihood function. Then

$$Q(\mu, y) = \int_y^\mu \frac{y - t}{V(t, \phi)}\, dt$$

is called the quasi-likelihood, or more correctly the log quasi-likelihood function. If the observations are independent, then the quasi-likelihood for the complete dataset is

$$Q(\mu, \mathbf{Y}) = \sum_{i=1}^{n} Q_i(\mu_i, Y_i).$$

In quasi-likelihood models, the β parameter vector would obtained by differentiating $Q(\mu, \mathbf{Y})$ and solving the equation $U(\hat{\beta}) = 0$, where

$$U(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{Y_i - \mu_i}{V(\mu_i, \phi)}. \tag{11}$$

It can be written in matrix form, like $U(\beta) = \mathrm{D}^\top \mathrm{V}^{-1}(\mathbf{Y} - \mu)$, where

$$\mathrm{V} = diag(V(\mu_1, \phi), \ldots, V(\mu_n, \phi))$$

and D is a $n \times d$ matrix, which contains the derivatives of μ with respect to β, i.e. $\mathrm{D}_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$. With these notations, the covariance matrix of $U(\beta)$ is

$$\mathcal{I}_Q(\beta) = \mathrm{D}^\top \mathrm{V}^{-1} \mathrm{D}, \tag{12}$$

which is also the negative expected value of $\partial U(\beta)/\partial \beta$. For quasi-likelihood functions, the $\mathcal{I}_Q(\beta)$ matrix is the analogue of Fisher information for ordinary likelihood functions.

The standard formulation of the variance function under the quasi-likelihood theory is $V(\mu, \phi) = \phi \mathrm{Var}[\mu]$, proportional to the mean, which is $\phi\mu$ in the Quasi-Poisson model. In this case, the moment estimator of ϕ is the previously mentioned Pearson dispersion statistic χ^2. One can also observe that the value of ϕ does not affect the estimation of the regression coefficients from equation 11.

Rewriting equation 11 to the Quasi-Poisson case, where $\mu_i = \mathrm{e}^{\mathbf{X_i}\beta}$ and $V(\mu, \phi) = \phi\mu$, we get

$$U(\beta) = \sum_{i=1}^{n} \mathbf{X_i}^\top \mathrm{e}^{\mathbf{X_i}\beta} \frac{Y_i - \mathrm{e}^{\mathbf{X_i}\beta}}{\phi \mathrm{e}^{\mathbf{X_i}\beta}} = 0, \quad \text{from}$$

$$\sum_{i=1}^{n} \mathbf{X_i}^\top (Y_i - \mathrm{e}^{\mathbf{X_i}\beta}) = 0.$$

In the end, we get back the equation 8, hence the estimated parameters are the same in the case of the Poisson and Quasi-Poisson model.

If the values of the estimated parameters are not affected by the non-completion of the equidispersion criterion, what changes in the Quasi-Poisson model? Let us compare the information that the data carry about β. We have to calculate equation 12, where $\mathrm{D} = \mathrm{MX}$ with $\mathrm{M} = diag(\mu_1, \ldots \mu_n)$ and $\mathrm{V} = diag(\phi\mu_1, \ldots, \phi\mu_n) = \phi\mathrm{M}$. Hence,

$$\mathcal{I}_{Qn}(\beta) = \frac{1}{\phi} \mathrm{X}^\top \mathrm{M} \mathrm{M}^{-1} \mathrm{MX} = \frac{1}{\phi} \mathrm{X}^\top \mathrm{MX}. \tag{13}$$

It can be seen that the Fisher information in the Poisson (9) and Quasi-Poisson (13) cases differs by a constant multiplier, which is the inverse of the dispersion parameter. If $\phi > 1$, i.e., there is overdispersion, then $\mathcal{I}_{Qn}(\beta) <= \mathcal{I}_n(\beta)$ element-wise. Since the MLE vector $\hat{\beta}$ has a multivariate normal distribution with mean β and covariance matrix $\mathcal{I}_n^{-1}(\hat{\beta})$, then in the Quasi-Poisson model the covariance matrix of the estimated parameter vector is not less elementally than in the Poisson model. Thus, the Quasi-Poisson model corrects the underestimation of standard errors. This means that the test statistic for the Wald test will also be corrected, it will be smaller and follows a t-distribution with $n - k$ degrees of freedom rather than standard normal. Therefore, if the overdispersion is not handled, the test statistic may exceed the critical value, but after correction it will be smaller than the critical value, and so a different conclusion is reached.

When the dispersion parameter is not equal to 1, the equations 1 and 2 are modified as follows.

$$f(Y_i, \theta_i, \phi) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi)\right\}, \quad l = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi).$$

Therefore, the relation for the expected value is unchanged, but the relation for variance is also modified:

$$\mathrm{E}_{(\theta_i)}[Y_i] = \mu_i = b'(\theta_i), \quad \mathrm{Var}_{(\theta_i)}[Y_i] = \phi b''(\theta_i) = \phi \mathrm{Var}_{(\theta_i)}[\mu_i].$$

When fitting the Quasi-Poisson model, the IRLS algorithm can also be used, but the weights are scaled according to the dispersion: $\mathrm{W}_{disp} = \phi\mathrm{W}$. This induces a new step in the iterations, estimate the dispersion parameter with the current model, and then re-estimate the regression coefficients with $\hat{\phi}$. Scaling actually adjusts the standard errors of the model to the value that would have been calculated if the dispersion statistic had originally been equal to 1.

Quasi-Poisson regression follows the GLM structure as it uses a linear predictor and a link function, the same as Poisson regression. The mean and variance of the response variable is also structured, where the dispersion parameter should be estimated (it is equal to 1 for Poisson regression). But beyond that, we cannot say that it is strictly a GLM. Quasi-Poisson assumes just a mean-variance relationship, not a full probability distribution. It also implies that the parameters are estimated using a quasi-likelihood method rather than a maximum likelihood estimation like the standard GLMs, and the techniques based on likelihood, such as Akaike Information Criterion (AIC), are invalid.

A summary of the similarities and differences between the Poisson and Quasi-Poisson regression results is shown in Table 2.

	Poisson	**Quasi-Poisson**
Parameter estimate	$\hat{\beta}$	$\hat{\beta}$
Standard error	$SE(\hat{\beta})$	$\sqrt{\hat{\phi}}SE(\hat{\beta})$
Wald test statistic	$z = \dfrac{\hat{\beta}}{SE(\hat{\beta})}$	$t = \dfrac{\hat{\beta}}{\sqrt{\phi}SE(\hat{\beta})}$
Confidence interval for β	$\hat{\beta} \pm z_{1-\alpha/2}SE(\hat{\beta})$	$\hat{\beta} \pm t_{n-k,1-\alpha/2}\sqrt{\hat{\phi}}SE(\hat{\beta})$

Table 2: Comparison between Poisson and Quasi-Poisson model results

2.2 Time series

Definition 5 (Time series). *The $Y_1, Y_2, \ldots, Y_t, \ldots$ sequence of random variables is called time series if its index parameter can be interpreted as time.*

Definition 6 (Autocorrelation). *The $r(t,s) = corr(Y_t, Y_s)$ correlations of the time series values at different time points are called the autocorrelation function.*

Definition 7 (Weak stationarity). *The time series $\{Y_t\}$ is called weakly stationary if the expected value of Y_t is constant and $cov(Y_t, Y_s) = R(t-s)$, so that the covariance of two values depend only on the difference of the reference times. In other words, the second centered moments are shift invariant.*

Time-series data can often be decomposed into different components that indicate patterns within a given time series. The three main components are as follows:

- **Trend**: It describes the long-term movement of data, the slow change over time.
- **Seasonality**: It shows the periodicity of the data, i.e. the short-term, cyclical oscillations.
- **Noise**: It stands for random variability.

Two types of decomposition are distinguished, depending on whether we take the sum or the product of the components:

- Additive model: $Y_t = T_t + S_t + \varepsilon_t$,
- Multiplicative model: $Y_t = T_t \cdot S_t \cdot \varepsilon_t$.

2.2.1 Time series regression

To predict an output series, it is often useful to use other explanatory variables that are also time-dependent. The general formula of a time series regression (TSR) model at time t, as follows:

$$Y_t = \beta_0 + \sum_{j=1}^{k} (\beta_j \mathrm{X}_{tj}) + \varepsilon_t,$$

where Y is the outcome, $\mathbf{X_j}$ ($j = 1, \ldots, k$) are the explanatory variables, which can be also lagged values of some predictors, β_0 is the intercept, β_i ($i = 1, \ldots, k$) are the regression coefficients and ε denotes the residual vector. The most common TSR model used for count data is the Poisson model, which can be written as

$$Y_t \sim Poisson(\lambda_t),$$

$$\log(\lambda_t) = \beta_0 + \sum_{j=1}^{k} (\beta_j \mathrm{X}_{tj}) + f(t),$$

where $f(t)$ is a smooth function of time t, whose role is to manage seasonality and long-term trend effects. In practice, this model is usually not fully valid due to overdispersion.

Compared to a basic regression, time series usually do not satisfy the independence of (adjacent) observations, i.e. the data are autocorrelated, which means that the IRLS algorithm described earlier cannot be applied in the same form (the covariance matrices will not be diagonal). Time series models can be categorized into two classes:

- **Observation-driven model**: The distribution of Y_t is determined by the past observations i.e. $Y_t = f(Y_{t-1}, \ldots, Y_1) + \varepsilon_t$.
- **Parameter-driven model**: Autocorrelation occurs by a latent process: it is assumed that (in the log-linear model) $\log(\mu_t) = \eta_t = \eta(\varepsilon_t, Y_{t-1}, \ldots, Y_1)$.

Here, the latter type is used.

3 Air pollution from wildfires

Particulate matter (PM) is a mixture of airborne solids and aerosols. The particles are divided into groups according to their size: coarse particles are up to 10 μm (PM_{10}), fine particles are under 2.5 μm ($PM_{2.5}$) and ultrafine particles are 1 μm or less in diameter (PM_1). Wildfires produce proportionately more fine and ultrafine particulates compared to coarse PM. Fine and ultrafine particles settle out of the atmosphere more slowly and can therefore travel further from the point of emission than the coarse ones. However, this is true not only for air but also for the human body, as they can enter deeper into the lungs and can cause health issues. [2]

3.1 Composition of the smoke

It is difficult to determine the composition of wildfire smoke [2] because it depends on many factors, such as the type of vegetation, its moisture content, weather conditions, etc., but also the phase of the burning determines the emergent chemical species. All of these make the content complex and varies over time. There is also a difference compared to the smoke from intentional burnings, which are carried out mainly during colder and wetter periods.

Unlike the combustion of fossil fuels, which are often used in industry, forest fire smoke contains many substances that are only produced when biomass is burned. These include, for example, combustion by-products of cellulose. But there is also a difference in the quality of the carbon present in the smoke: while fossil material produces it in elementary form when it burns, most of the particles after wood combustion are organic carbon.

The study of ozone production can also play an important role. Volatile organic chemicals in forest fire smoke can mix with anthropogenic nitrogen oxides in urban areas to produce ozone. In addition, much of the nitrogen from the smoke is converted into a stable compound (peroxyacetyl nitrate), which decomposes to produce ozone. All this means that, depending on the strength and direction of the wind, local ozone concentrations can increase away from the point of ignition.

3.2 Spread and behaviour of the emitted substances

The speed, length and duration of smoke transport depend on many factors [15]. One of the determining variables is the composition of the smoke, as some substances can react with molecules in the atmosphere. In addition, the size of the particles also affects the dispersion. Larger particles often only travel a few kilometres at most and then settle on the ground surface. Small particles, however, can travel high up into the stratosphere (at least 10 km), where they can remain for weeks or months, and horizontally they can travel up to 1000 km [3]. Another important factor is the weather. Wind speed, of course, but also humidity or temperature can shape the smoke cloud.

However, it is not only the weather that has an effect on the spread of smoke, but also the other way round. For example, heavy smoke delays the formation of precipitation. This means that particles stay in the atmosphere longer and travel further because they are not washed out of the air by rain. Smoke aerosols can also interact with solar and UV radiation, which affects cloud formation. Furthermore, reduced cloud cover leads to reduced humidity.

3.3 Potential diseases

To provide reliable results on the health effects of wildfire smoke, large-scale studies are needed, both spatially and temporally. Since the composition of smoke is dependent on many factors, its impacts can vary from area to area and can also be influenced by non-static variables such as weather at a given location. Furthermore, it is possible that a disease may not appear immediately at the time of the event or afterwards but may be progressively revealed over a long period of time. However, what is certain is that the quantity of fine particles in the air increases.

Particulate matter from bush and forest fires can generate more free radicals and more oxidative stress in the lungs than particles in urban environments, due to the occurrence of more polar organic species [2]. It can also induce an increase in specific symptoms, such as asthma. During a wildfire event, the risk of developing respiratory diseases is potentially increased. Several studies, including Liu et al. [11], have shown an association between wildfire smoke and the number of hospital visits for respiratory symptoms, proportional to the intensity of smoke waves.

Chen et al. have studied the association between wildfire-related $PM_{2.5}$ and mortality across different parts of the Earth [3]. Particles from wildfires has been shown to contribute to all-cause, respiratory and cardiovascular mortality. The association can be proved at global level, but the magnitude of the change in mortality risk differs between countries and regions. Chowdhury and others have published a Europe-wide study on the relationship between $PM_{2.5}$ exposure due to wildfires and chronic diseases [4]. It was found that wildfire-related particulate matter is more toxic to short-term health than those from other sources. According to age-categorized estimates, older people are primarily at risk.

4 Modelling

4.1 Study area

For my own studies, I chose Greece, one of the most fire-prone countries in Europe. Greece is located in the South-Southeast of the continent, between the latitudes $34°48'N$ and $41°46'N$ and the longitude ranges from $19°22'E$ to $29°39'E$. The country covers a total area of $132,049$ km^2 and has the 11th longest national coastline in the world at $13,676$ km, in large part due to its numerous islands and indented coastline.

In terms of topography, 80% of the country is mountainous, with the highest elevation at $2,917$ m and the lowest at -6 m. In 2021, 26% of the territory was covered by forest, 23% by arable land, 42% by grassland, 5% by settlements, and the remaining 4% by wetland and other types of land [5].

Greece has a Mediterranean climate, which means long periods of sunny days during the whole year, wet and mild to cool winters and dry, hot summers, mainly near the coast and on the islands. Toward the interior of the mainland, the climate is strongly influenced by the topography. The hottest time of the year is the second half of July and the first half of August, when the average daily maximum temperature ranges from 29 $°C$ to 35 $°C$. However, during heatwaves, temperatures above 40 $°C$ are becoming more common, occasionally exceeding also 45 $°C$. [8]

The official fire season in Greece runs from May to October. Every year, it takes a lot of resources to fight wildfires, but there is an increasing trend in both frequency and intensity due to climate change. Hot, dry and windy weather, vegetation and topography all contribute to the development of fires. Figure 1 shows the total area burnt by wildfires in recent years (data are from the European Forest Fire Information System (EFFIS)[1]).

[1] https://forest-fire.emergency.copernicus.eu/apps/effis.statistics/estimates/GRC

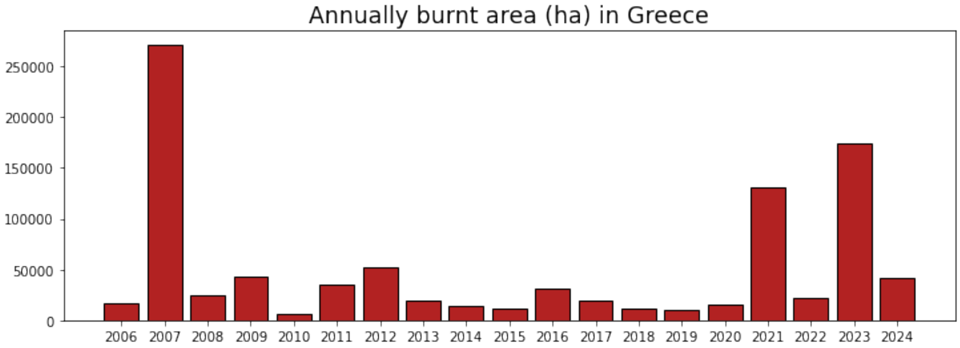

Figure 1: Annual burnt area statistics for Greece

Air pollution is a major environmental problem in Greece, determined by many factors. In addition to smoke from frequent wildfires, air quality is affected by pollutants from cities and industrial regions, dust from Africa and the Middle East, and also sea salt and anthropogenic pollution from major Mediterranean cities carried by air masses [1]. The main indicators of poor air quality are ozone and PM_{10}, with a mean annual 75.34% and 29.36% of the urban population exposed to higher concentrations of these pollutants than the EU standard (Table 3) between 2018 and 2022 [7].

Pollutant	**Concentration**	**Averaging period**	**Permitted exceedences each year**
Ozone	120 $\mu g/m^3$	maximum daily 8 hour mean	25 days averaged over 3 years
PM_{10}	50 $\mu g/m^3$	24 hours	35 days

Table 3: EU air quality standards

4.2 Data

The database variables used for modeling are available on different sites and in different formats. After preprocessing, the data have the following resolution and extent:

- **Temporally**:
  - Extent: 2015 - 2019
  - Resolution: weekly, from the 20th to the 42nd each year
- **Spatially**:
  - Resolution: NUTS 3 regions (52 members)

The number of deaths as the target variable is further categorized as follows: by 5-year age groups (the last group is 90 years and over, plus the total values) and by sex (total, female and male). Figure 2 shows the values of mortality rates by age group. Not surprisingly, this indicator increases as we look at older people. We can see it against the genders in Figure 3. In general, women have lower mortality rates, but the outliers are similar in quantity and magnitude to those for men.

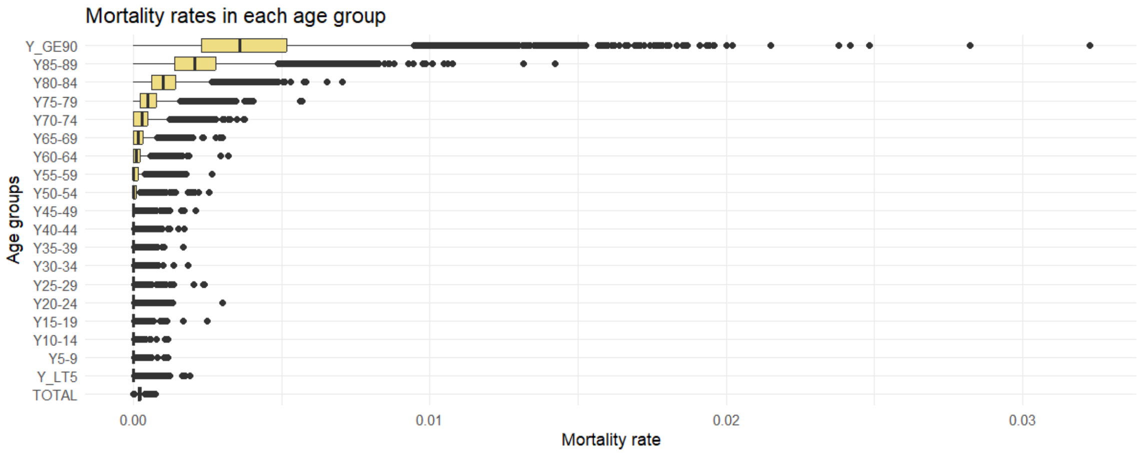

Figure 2: Boxplots of the mortality rates by age groups

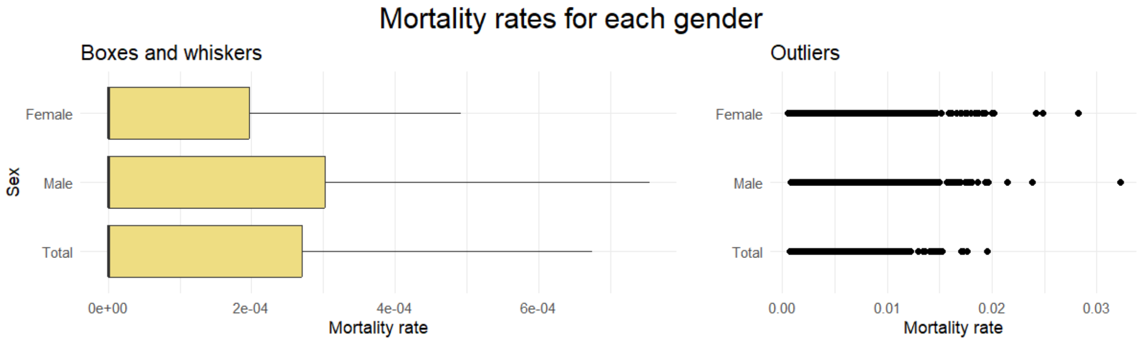

Figure 3: Boxplots of the mortality rates by genders

The start of the study period was determined by the availability of data, as previous data on deaths are highly incomplete. The last year is 2019, as the first official coronavirus infection in Greece was in February 2020, so the mortality data for the summer and the next few years were significantly affected by the pandemic and cannot be disjointly separated from the effect of smoke. Within a year, the weeks studied were chosen to coincide with the Greek fire season. Thus, the effect of temperature is also mostly limited to heat (which can also cause health problems) and is not distorted by, for example, influenza epidemics associated with cold winter weather.

The NUTS (Nomenclature of Territorial Units for Statistics), which was developed by the EU, classifies each country into three levels. NUTS 3 is for the small regions, the finest division. Figure 4 shows the Greek areas.

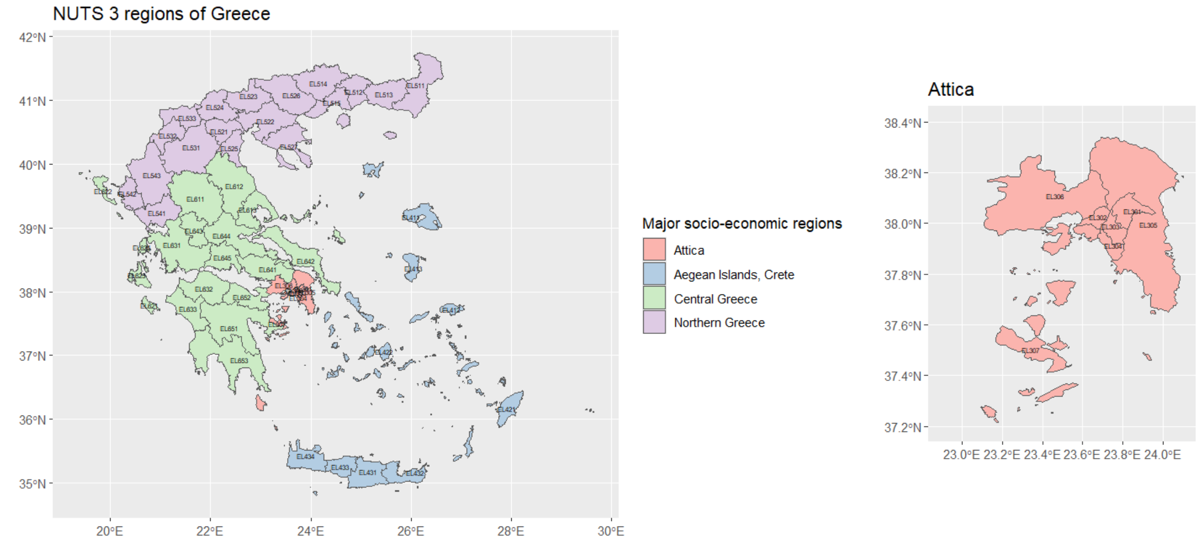

Figure 4: NUTS 3 regions of Greece, grouped into NUTS 1 by colour

Demography

Data on the number of deaths in Greece come from a Eurostat database[2] and have the same resolutions and categories as the final table described above. Here, minimal formatting of the data was needed to make it a more manageable layout.

To calculate the mortality rate, I also downloaded the population data from Eurostat[3]. It is also detailed by age group, sex, and NUTS 3 regions, but this is yearly data (measured on 1 January).

Weather

The weather data, that is, mean temperature ($^\circ C$), maximum temperature ($^\circ C$) and relative humidity (%), are from Copernicus Climate Change Service[4], which are measured near the surface, at a height of 2 m. The raw variables are gridded with horizontal resolution $0.1^\circ \times 0.1^\circ$, and in temporal terms, they are daily values.

During preprocessing, I first calculated the respective weekly averages and maxima for each grid point. Then I determined from the grid points which region of Greece (defined as polygons) they fall into, and again took the spatial means and maxima. However, for two

[2] https://ec.europa.eu/eurostat/databrowser/view/demo_r_mweek3__custom_15383641/default/table?lang=en

[3] https://ec.europa.eu/eurostat/databrowser/view/demo_r_pjangrp3/default/table?lang=en&category=demo.demopreg

[4] https://cds.climate.copernicus.eu/datasets/insitu-gridded-observations-europe?tab=overview

regions, Western and Southern Athens, no observations fell within their boundaries. Thus, on each date, I assigned to them the weighted average of the nearest 5 and 4 observations (see Figure 5), respectively, where the weights are the reciprocal of the distance between the location of the observation and the center of the region, normalized so that the sum of the weights is equal to 1.

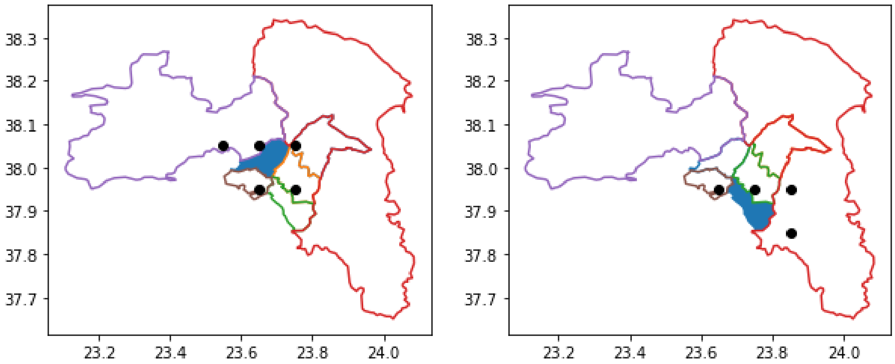

Figure 5: The blue regions are Western and Southern Athens, and the black points are the locations of the nearest weather observational points. (There is a sea to the southwest, so there are no observations.)

Another problem with the relative humidity observations was that there were a significant number of missing values (approximately 45,6 %). To handle this, I used linear interpolation in the plane for each date. For seven regions, all belonging to the Aegean Islands and Crete, unfortunately all the 2018 and 2019 data were missing, and due to their location, spatial interpolation may not be the best way to fill them. Hence, I fitted the following linear model to the regions concerned (the coefficients were estimated with Ordinary Least Squares):

$$\begin{aligned}\text{Relative humidity} = & -0.6 \cdot \text{mean temp.} - 1.182 \cdot \text{max temp.} - 0.159 \cdot \text{year} \\ & + \text{region} + \text{week} + 433.625 \cdot 1(\text{intercept}),\end{aligned}$$

where variables region and number of the week are factors, while year and temperatures are numeric values. Figure 6 shows the coefficients of the regions, where the base category is EL301 (North Athens). Similarly, the coefficients for the week factor are in Figure 7, where week 20 is the base. I predicted the missing relative humidity values with this model. The whole model output (coefficients, std. error, t values, and p-values) can be seen in the Appendix (Table 11).

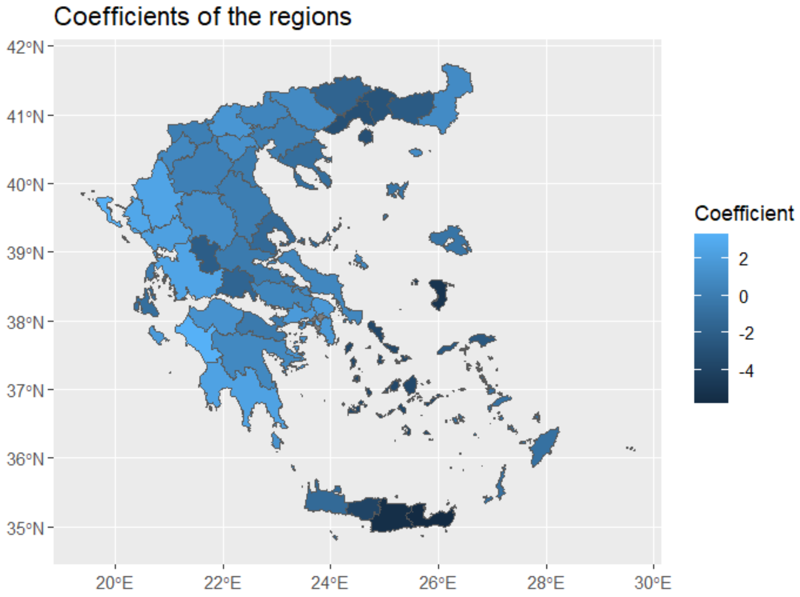

Figure 6: Coefficients of the region factor in the humidity model

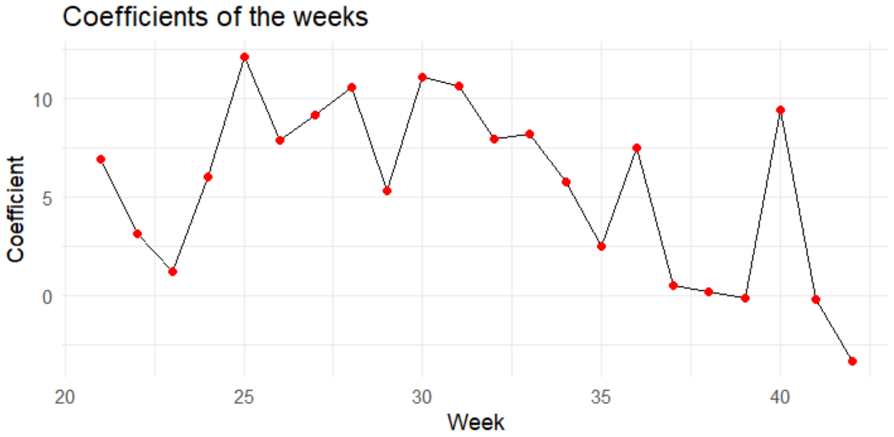

Figure 7: Coefficients of the week factor in the humidity model

Burning emissions

Data on air emissions from wildfires come from the Copernicus Atmosphere Monitoring Service (CAMS)[5]. The information is provided by the CAMS Global Fire Assimilation System, which is based on satellite observations of the fire radiation power (FRP). In my thesis I use the following variables, the flux of carbon dioxide (CO_2), organic carbon, nitrogen oxides (NO_x), fine particulate matter ($PM_{2.5}$) and total particulate matter, each in $\frac{kg}{m^2s}$. Similarly to weather variables, in their raw state these data are also available at $0.1° \times 0.1°$ grid points, with daily resolution.

[5] https://ads.atmosphere.copernicus.eu/datasets/cams-global-fire-emissions-gfas?tab=overview

The emissions data from biomass burning were preprocessed as temperature variables. First, I calculated the weekly averages for each grid point, then assigned the values to the NUTS 3 regions, and again took the regional averages. Since the grid point locations are the same as the data from Climate Copernicus, the problem of Western and Southern Athens is also present here. Since the observation points are close to the areas and the effect of smoke can appear in these regions (e.g. depending on the wind), I took an unweighted average. Another difference is that I considered 5-5 grid points for both regions (see figure 8), since there is measured data south of Southern Athens, just inland (weather was not observed at this point).

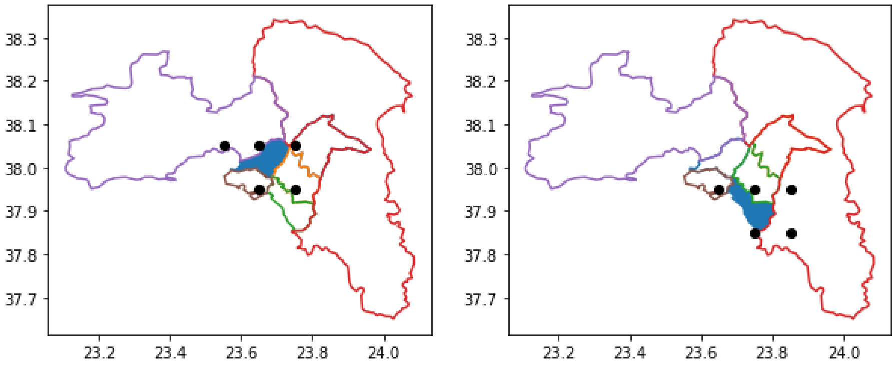

Figure 8: The blue regions are Western and Southern Athens, and the black points are the locations of the nearest emission observational points. (There is a sea to the southwest, so there are no observations.)

The distribution of the non-zero fluxes of the different pollutants is shown in Figures 9, 10, 11, 12 and 13. Overall, 13.04 % of the observations (year, week, region) have a positive emissions value, i.e. this is the fire exposure rate.

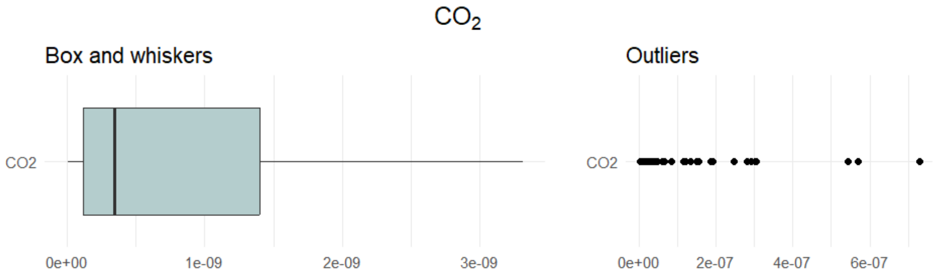

Figure 9: Boxplot of the non-zero valued CO_2 fluxes

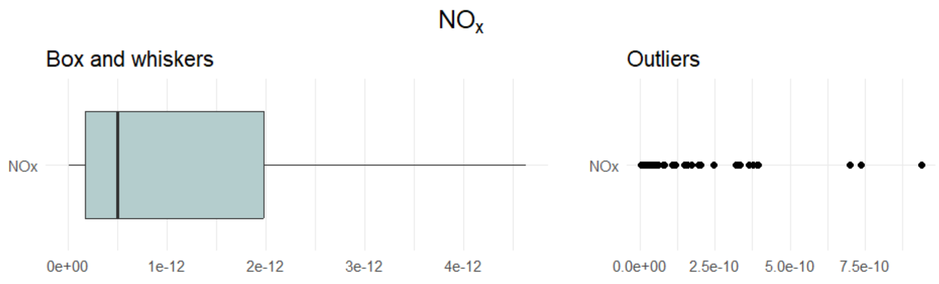

Figure 10: Boxplot of the non-zero valued NO_x fluxes

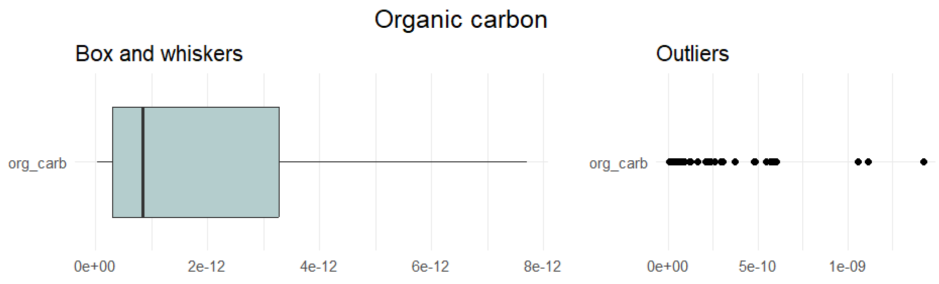

Figure 11: Boxplot of the non-zero valued organic carbon fluxes

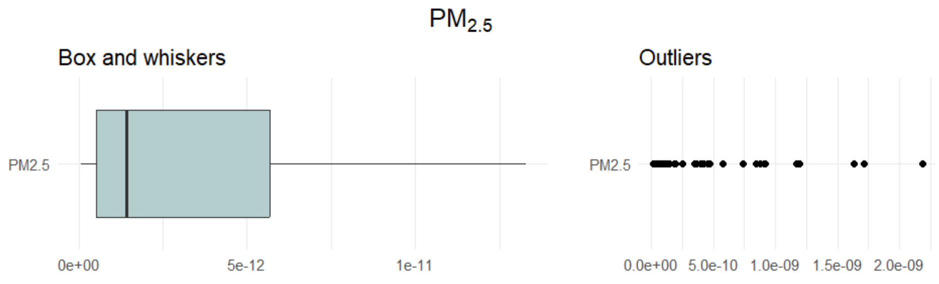

Figure 12: Boxplot of the non-zero valued $PM_{2.5}$ fluxes

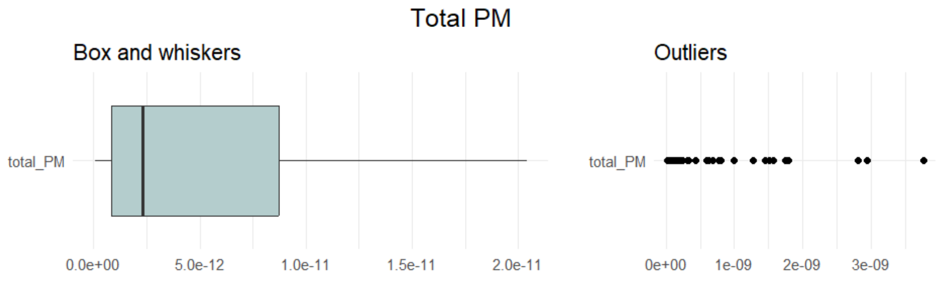

Figure 13: Boxplot of the non-zero valued total PM fluxes

4.3 Method

Based on the empirical correlations calculated from the dataset (see Figure 14), a negative coefficient for relative humidity and positive coefficients for temperatures are expected. This is logical, as high temperatures often cause illness and the air is typically drier. The correlation of combustion emissions with mortality rate is very low, which can be explained by the fact that in most cases the fluxes are 0. However, they are all positive, which is good, as this is what the thesis aims to measure.

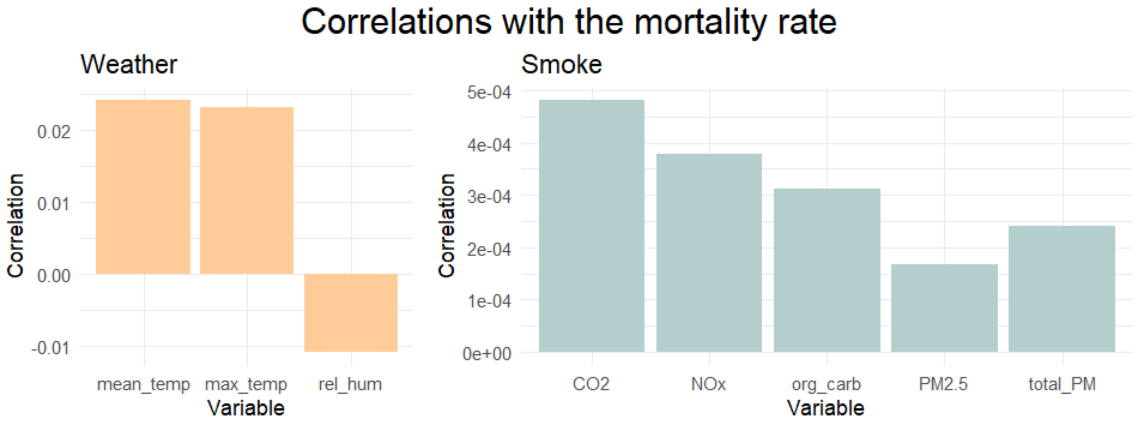

Figure 14: The correlations of the explanatory variables with the target

In each case, I performed the tests as follows: I fixed an area and a year, then fitted a model to the weekly data. Since I am not looking at the whole duration of the years, so that after the last week in October it will be about 6 months until the next observation, that is why I fitted a separate model for each year. I also changed the unit of measurement of pollutants from $\frac{kg}{m^2s}$ to $\frac{\mu g}{m^2s}$ to minimize the possibility of machine error. I have done three rounds of modeling, according to the attributes of the mortality data:

- Aggregated deaths, sex, and age groups are not distinguished.
- Gender-specific analysis, comparison between males and females.
- Age-specific analysis, comparison between the age groups.

In each model, I considered a 1-week shift of the target variable based on autocorrelations (see Figure 15) and computational demand.

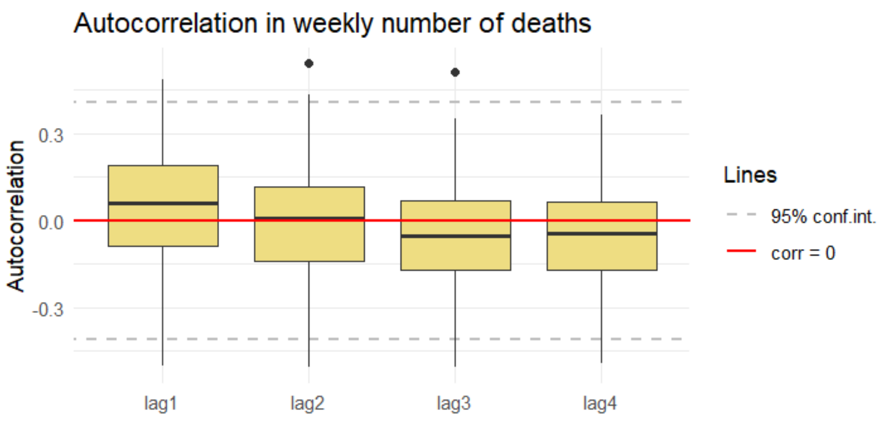

Figure 15: Autocorrelations in the aggregated deaths (fixed regions and years)

Because of the strong multicollinearity in the explanatory variables, I ended up using only the mean from the two temperatures (which represents the persistent heat) and relative humidity, as weather indicators. For smoke, I have included all the components first, but these variables are also highly correlated (see Figure 16), so in this case it is worth looking at their joint effect, as individual coefficients can be misleading. In order to observe the influence of the different pollutants alone, I also performed the modeling by using only one of the smoke variables at a time. For the representation of a given emission, I used two different approaches after calculating the 1- and 2-week lags: first, I used the 3-week moving average and then introduced the 0, 1-, and 2-week lags as separate variables.

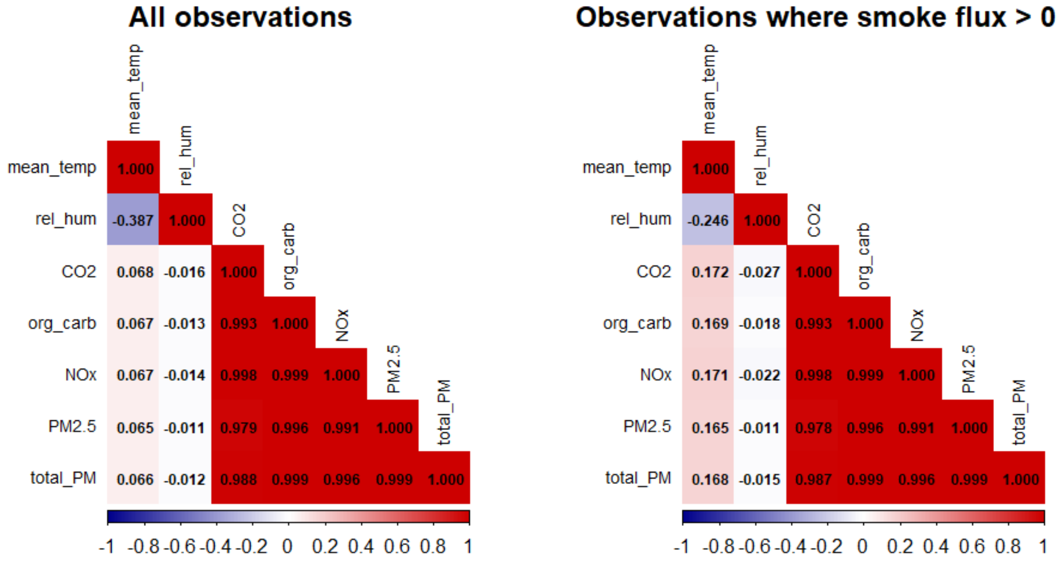

Figure 16: Correlation matrices of the explanatory variables with respect to all observations and fiery weeks

Gender and age-specific analyses were performed using only the moving average approach and PM$_{2.5}$ was used as the smoke variable. I applied one-hot encoding to the given factor variable and included the interaction of PM$_{2.5}$ and the sex/age factors as additional variables.

5 Results

5.1 Total population

As shown in Table 4, for the 3-week moving average, the correlations of combustion emissions with the number of deaths are higher than for the lagged ones individually, but the latter together provide more information than separately. I did not take into account the higher lags of the emission variables, because they have smaller correlations with the number of deaths (see Table 5), and also reduce the number of observations. Shifts could not be calculated for the first weeks of each year, and although there is a very small chance of a wildfire in April, it cannot be excluded.

	CO_2	NO_x	organic carbon	$PM_{2.5}$	total PM
lag 1	0.0512	0.0499	0.0488	0.0467	0.0479
lag 2	0.0502	0.0487	0.0475	0.0451	0.0465
3 week moving average	0.0752	0.0731	0.0713	0.0681	0.0700

Table 4: Correlations between deaths and different functions of the pollutants

	CO_2	NO_x	organic carbon	$PM_{2.5}$	total PM
lag 3	0.0415	0.0401	0.0389	0.0367	0.0380
lag 4	0.0350	0.0337	0.0326	0.0306	0.0317

Table 5: Correlations between deaths and the (unused) higher lags of the pollutants

To measure the significance of the variables, I ran a model using the glm function (which assumes the independence of observations), as this output provides more information than tsglm, which is used later. I used PM$_{2.5}$ from the smoke variables, first just the current value, then the moving average of 3 weeks, and finally the current and the lagged values of 1 and 2 weeks. The dispersion parameter for the Quasi-Poisson family was taken to be 1.15. The p-value for the mean temperature was always very small ($< 2 \cdot 10^{-16}$), but the relative humidity values were also below 0.05 on all three occasions (between 0.011 and 0.018). The situation is different for fine particles. On its own it did not seem significant with a p-value of 0.246, but when looking at the moving average I got a value of 0.023, which is below the usual significance level. Separately for the lagged variable, I also obtained values higher than 0.05, which are 0.332, 0.099, and 0.214. (The 1 week shift seems to be the most useful of these three variables.) The full outputs of the models can be found in the Appendix (Tables 12, 13 and 14).

5.1.1 All components of the smoke

In this case, as I mentioned before, I only considered the cumulative effect of the burning emissions. Table 6 contains the mean raw and mean transformed coefficients of the two weather variables in both models (using moving average and lagged values of the pollutant variables). As expected, higher temperatures increase the number of deaths and humidity has the opposite effect, with drier air proving more dangerous.

	raw coef. from moving average	exp coef. from moving average	raw coef. from lagged values	exp coef. from lagged values
Mean temperature	0.0184	1.0190	0.0170	1.0174
Relative humidity	-0.0005	0.9995	-0.0011	0.9990

Table 6: The mean coefficients of the weather variables in the all smoke components models

I calculated the multipliers associated with combustion emissions separately for each observation and then aggregated them to get the combined effect. If there were no fires in a given region and in a given year, then all fluxes are 0, so the model also gives zeros as coefficients. Leaving these cases out, Figures 17 and 18 show the estimated multipliers from the two models for all regions and all years. In summarizing the further results, I also do not consider the cases where the coefficient of the pollutant is exactly 0 (and hence the corresponding multiplier is 1).

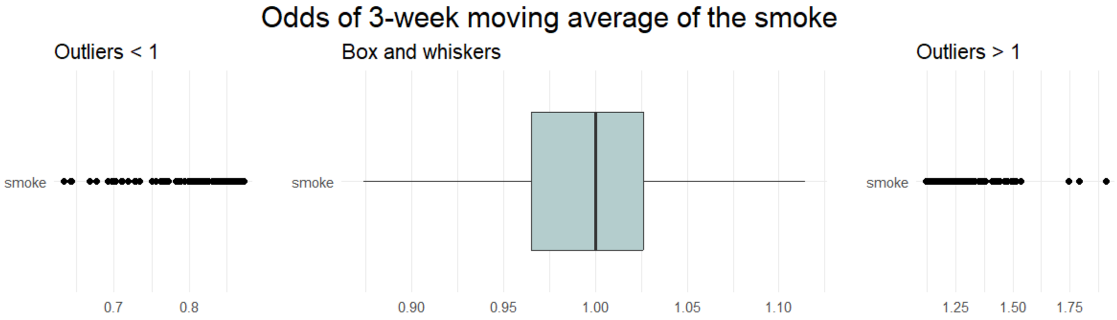

Figure 17: Boxplot of the odds calculated for each observations using 3-week moving averages of all pollutants

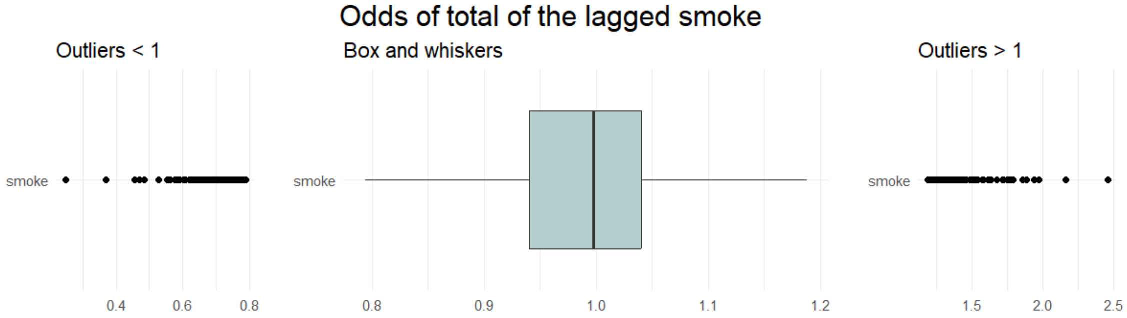

Figure 18: Boxplot of the odds calculated for each observations using lagged values of all pollutants

From the multipliers obtained, I calculated the estimated number of deaths attributable to wildfire smoke in each week in each region using the following formula.

$$\hat{D}_{smoke}(r, w) = D(r, w) - D(r, w)/odds(r, w),$$

where $D(r, w)$ is the number of deaths in region r on week w, and $odds(r, w)$ is the estimated multiplier of smoke in region r on week w (the odds in a given region are the same for a given year). The estimate using the moving average model is 673 deaths for the whole country and for the whole period studied, which is an average of 135/year. With the second approach these are 248 in total and 50 per year. There is a rather large difference between the two estimates, and the second one seems more uncertain from the figures. Here the range of results is much wider, and an outlier can very much distort the overall picture.

Figure 1 shows that the burnt area in 2023 is approximately 10 times the average burnt area in the years I have studied. Assume, for example, that the fluxes change by a factor of 10. Then, with the same coefficient, the odds increase to the power of 10. In the moving average model, the combined average multiplier for all smoke variables is 1.0047 (when not equal to 1). This would change to 1.048007, but if we average the new multipliers, we get that it increases by a factor of 2.2589. For the whole country and for the total 5 years, the number of deaths would increase from 673 to 1520 using this value, which means 304 cases for one year.

5.1.2 PM$_{2.5}$

Moving average

The average of the results of all 260 runs (52 regions, 5 years) using the moving average is shown in Table 7.

	Mean temperature	Relative humidity	$PM_{2.5}$
raw coefficient	0.01859	-0.00102	1.72522
exp coefficient	1.01911	0.99904	$2.735 \cdot 10^{143}$

Table 7: The mean coefficients of the model using 3-week moving average of $\text{PM}_{2.5}$

The mean of the coefficients could be biased by one or two outliers, which due to the exponential function contain even larger values. However, if we take the average of the coefficients of particulate matter and multiply by the average of the values of particulate matter and then take the exponential, we get a multiplier of 1.0068 in the estimate of the target variable. The same only by omitting the 0 values is 1.0313, which means that it would increase the number of deaths by 3.13 percent. The distribution of the odds calculated for each observation is shown in Figure 19.

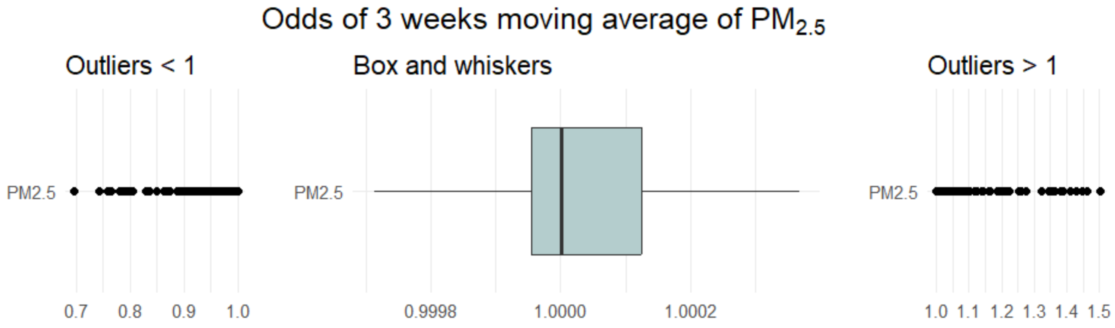

Figure 19: Boxplot of the odds calculated for each observation using 3-week moving average of $\text{PM}_{2.5}$

Sometimes a negative effect is obtained for the fine particles, but positive values are the majority, and the outliers are also larger in this direction. Overall, using these values, it is estimated that 562 people died from $\text{PM}_{2.5}$ emissions during the period under review, an average of 112 people/year.

Lagged values

In the second approach, the pollutant variables were the flux of fine particulate matter from the actual week and the 1 and 2 week lags. Table 8 shows the mean of the coefficients in the same form as before.

	Mean temp.	Rel. humidity	$PM_{2.5}$	$PM_{2.5}$ lag 1	$PM_{2.5}$ lag 2
raw coef.	0.01876	-0.00098	2.04878	1.04913	-1.69484
exp coef.	1.01929	0.99909	$6.934 \cdot 10^{87}$	$1.230 \cdot 10^{42}$	$3.467 \cdot 10^{58}$

Table 8: The mean coefficients of the model using 0-, 1- and 2-week lags of $\text{PM}_{2.5}$

With and without 0 for the mean coefficient and the mean flux, we obtain multipliers of 1.0081 and 1.0779 for the given week, 1.0041 and 1.0410 for a 1-week lag, and 0.9934 and 0.9322 for a 2-week lag. These suggest that, in general, $PM_{2.5}$ from a given week and 1 week earlier due to a wildfire will increase the number of deaths, while data 2 weeks apart may 'decrease' the number of deaths. Figure 20 shows the multipliers calculated with the observed values and the joint effect of the 3 variables.

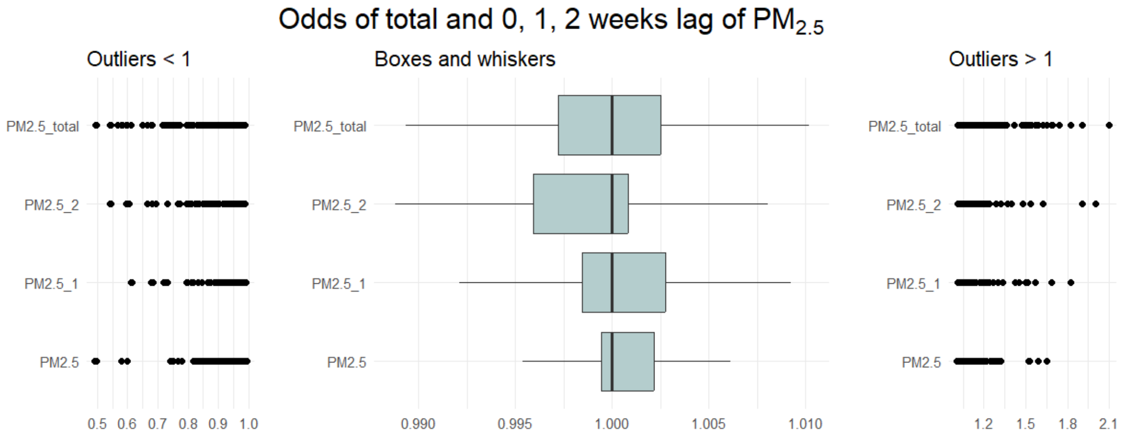

Figure 20: Boxplots of the odds calculated for each observation using lagged values of $PM_{2.5}$, excluding the odds of 1 (where the coefficients are 0)

Here, there are even larger outliers in both directions compared to the moving average, but especially in the positive direction, even above a multiplier of 2. In this case, the estimated total number of deaths during the period under review is 543, or an average of 109 per year. These estimates are similar to the moving average, but slightly lower.

5.1.3 Other pollutants

Moving average

The distributions of the multipliers for CO_2, NO_x, organic carbon and total particulate matter are shown in Figures 21, 22, 23 and 24. For each of these variables, there are larger outliers upwards, which are similar in magnitude, although the widths of the boxes are different. The carbon dioxide flux values are about 1000 times the other variables, and here we observe the widest range in all parts of the boxplot. The averages of the multipliers are, in order, 1.0045, 1.0015, 0.9997, and 1.0028, i.e. for organic carbons the value is less than 1, but for the other variables it is greater than 1.

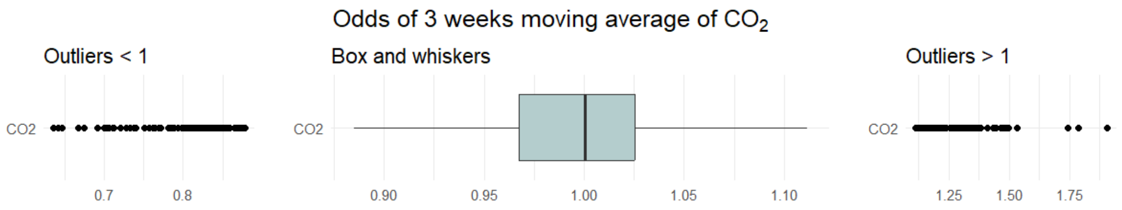

Figure 21: Boxplot of the odds calculated for each observation using 3-week moving average of CO_2

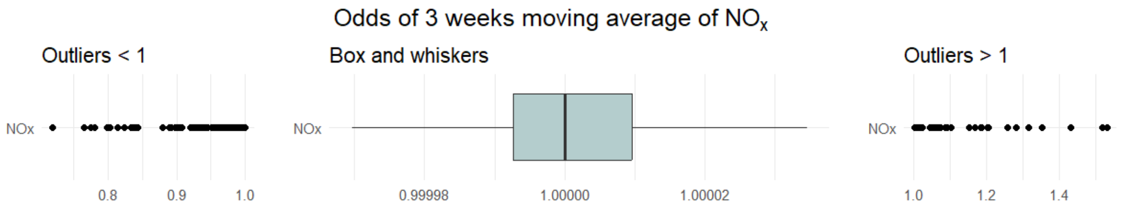

Figure 22: Boxplot of the odds calculated for each observation using 3-week moving average of NO_x

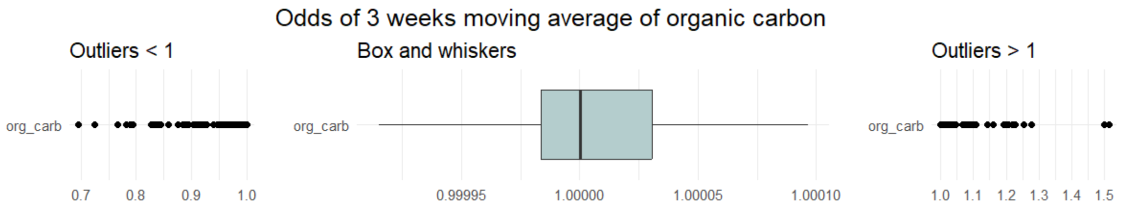

Figure 23: Boxplot of the odds calculated for each observation using 3-week moving average of organic carbon

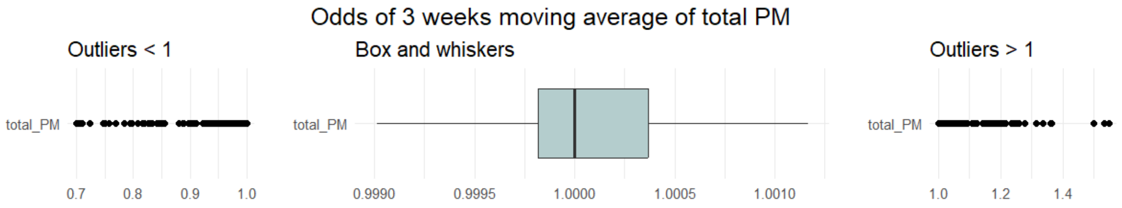

Figure 24: Boxplot of the odds calculated for each observation using 3-week moving average of total PM

Mortality models using different burning emissions, as can be seen in the multipliers, produce different estimates of the number of deaths (see Table 9). The aggregated result, including all 5 smoke-related variables, is shown in Figure 9. Based on these results, we can divide the variables into two groups: CO_2, $PM_{2.5}$, and total PM, which give higher estimates of mortality and are therefore perhaps more dangerous, and NO_x and organic carbon, which give much lower estimates. Different smoke components may have the same or different effects on the human body, so they cannot be considered as independent. Hence the number of deaths estimated from different pollutants cannot be added together.

	CO_2	NO_x	organic carbon	total PM
overall	714	203	144	498
yearly	143	41	29	100

Table 9: The estimated number of deaths caused by different pollutants

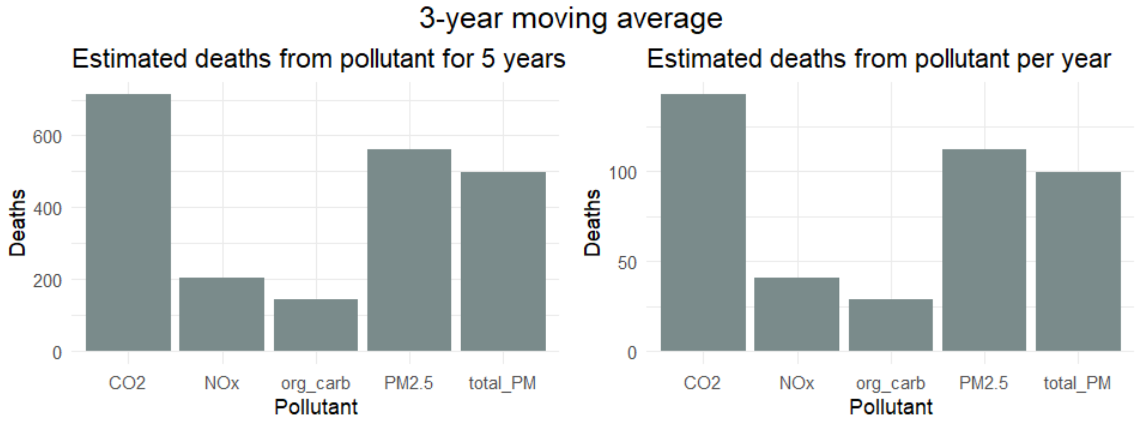

Figure 25: Estimation of overall and yearly deaths based on the different pollutants using moving averages

Lagged values

Similarly to $PM_{2.5}$, the boxplots for the other combustion emissions have higher outliers and also the boxes are wider when the current and 1 and 2 week lags are the explanatory variables for the smoke (on the Figures 26, 27, 28 and 29 are the joint effects). However, it is not surprising that the three variables together lead to higher uncertainty in this case. Table 10 contains the summary of the odds of the different lagged variables. In general, the means of the multipliers show a decreasing trend as the value of the shift increases. The exception to this is NOx, where the multipliers for the one-week lag had the lowest average, and the two-week lag was almost identical to the non-lagged variable.

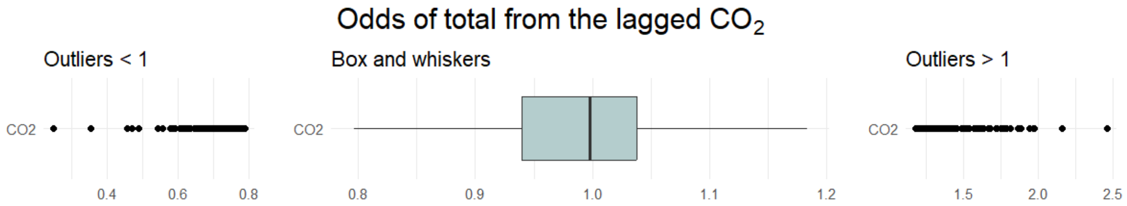

Figure 26: Boxplot of the odds calculated for each observation using lagged values of CO_2

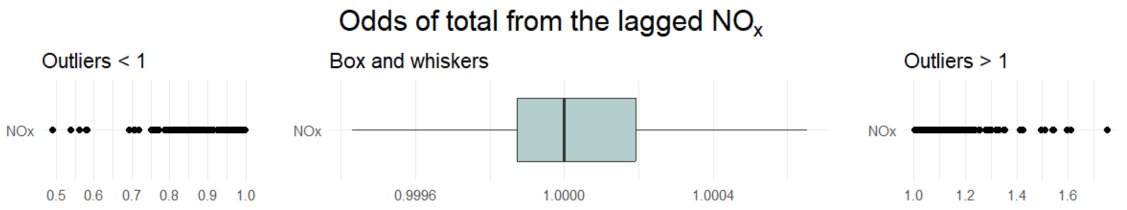

Figure 27: Boxplot of the odds calculated for each observation using lagged values of NO_x

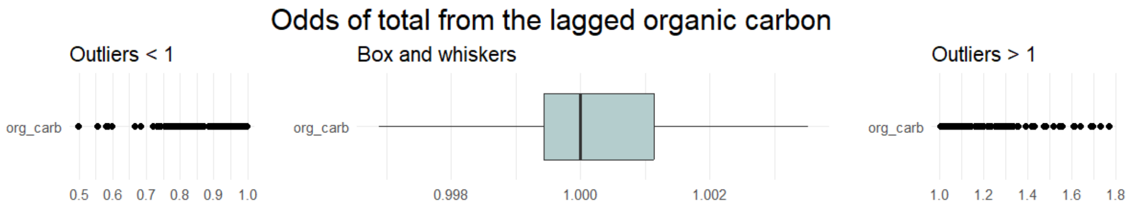

Figure 28: Boxplot of the odds calculated for each observation using lagged values of organic carbon

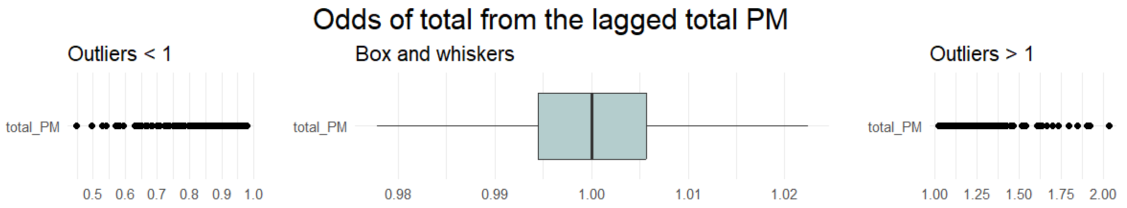

Figure 29: Boxplot of the odds calculated for each observation using lagged values of total PM

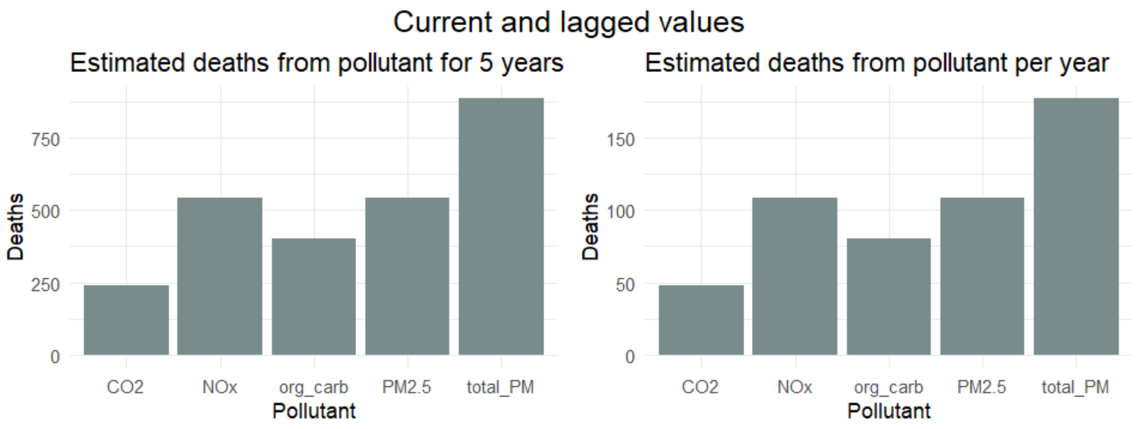

Figure 30: Estimation of overall and yearly deaths based on the different pollutants using current and lagged values

When estimating the number of deaths (see Figure 30), significant differences can be observed compared to the moving average method. Although the results for fine particles are quite similar, there are large differences for other variables. For example, carbon dioxide first gave the highest estimate and now has the lowest. However, a similarity between the two types of models is that the total PM gave a relatively high estimate in both cases, but the second time it outperforms all other variables.

	Min	1st. Qu	Median	Mean	3rd Qu.	Max
CO_2	0.2484	0.9756	1.0004	1.0136	1.0272	2.1593
CO_2 lag 1	0.5187	0.9611	0.9979	1.0030	1.0235	1.9758
CO_2 lag 2	0.3534	0.9510	0.9952	0.9953	1.0176	2.4633
NO_x	0.4889	1.0000	1.0000	1.0038	1.0002	1.6100
NO_x lag 1	0.5658	0.9999	1.0000	0.9991	1.0001	1.5062
NO_x lag 2	0.5957	0.9998	1.0000	1.0039	1.0001	1.7527
organic carbon	0.4990	0.9999	1.0000	1.0063	1.0016	1.6586
organic carbon lag 1	0.5629	0.9994	1.0000	1.0028	1.0008	1.7312
organic carbon lag 2	0.5872	0.9988	1.0000	1.0012	1.0003	1.7680
total PM	0.4489	0.9990	1.0000	1.0080	1.0048	1.7722
total PM lag 1	0.5289	0.9952	1.0000	1.0075	1.0038	2.0301
total PM lag 2	0.5395	0.9922	1.0000	1.0062	1.0024	1.9166

Table 10: Basic statistics of the multipliers of the different pollutants and lags

5.2 Comparison of sexes

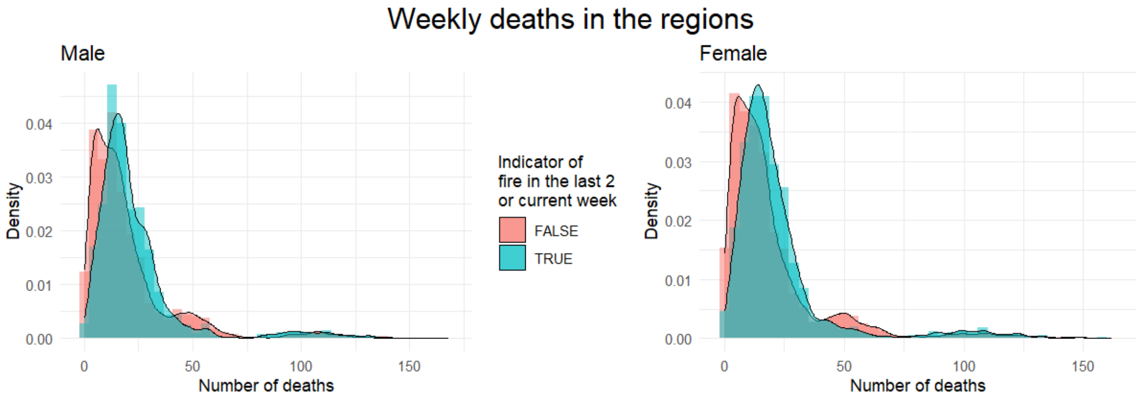

Figure 31: Histogram of weekly deaths in the regions by the indicator of whether there was a wildfire in that or the preceding two weeks, separately for the sexes.

As Figure 31 shows, a slight difference can be seen between the histogram or the empirical density function (of course the number of deaths can only be integers) whether there was a wildfire in that or the preceding two weeks. Technically, it means that the moving averages of the smoke variables are greater than or equal to 0. On both histograms we can see a peak around 50, which are not from fiery weeks. Most of these observations are from the Athenian regions, where the higher population means that more people tend to die in a week, and the urban land cover means that we cannot speak of wildfires.

As before, I have estimated the number of deaths caused by fine particles over the period studied. To compare the two sexes, I calculated the percentage of deaths that were due to the interaction of sex and $PM_{2.5}$, and then added the overall effect of the fine particles. The result was that interaction with women increased the number of deaths by approximately 0.02 %. Overall, I estimated that 0.33 % of male deaths and 0.35 % of female deaths were due to $PM_{2.5}$ emissions associated with wildfires. These suggest that the difference is not large, but that women's health was slightly more affected by air pollution from wildfires.

Several studies have been published in the literature on the effects of air pollution for males and females, particularly on the respiratory system. Although the results vary, more studies have shown stronger effects by women than by men ([6]). The difference may be due to a number of factors, including social (e.g., work, lifestyle) and biological (e.g. hormonal status) differences.

5.3 Comparison of age groups

The difficulty with age groups is that there are often 0 values for the number of deaths. While I first considered the whole population and then split the data into two parts by sex, in this section, I have divided the observations into 19 groups. Figure 32 shows the average number of weekly deaths per region depending on whether the 3-week moving average of emitted combustion substances is greater than or equal to 0.

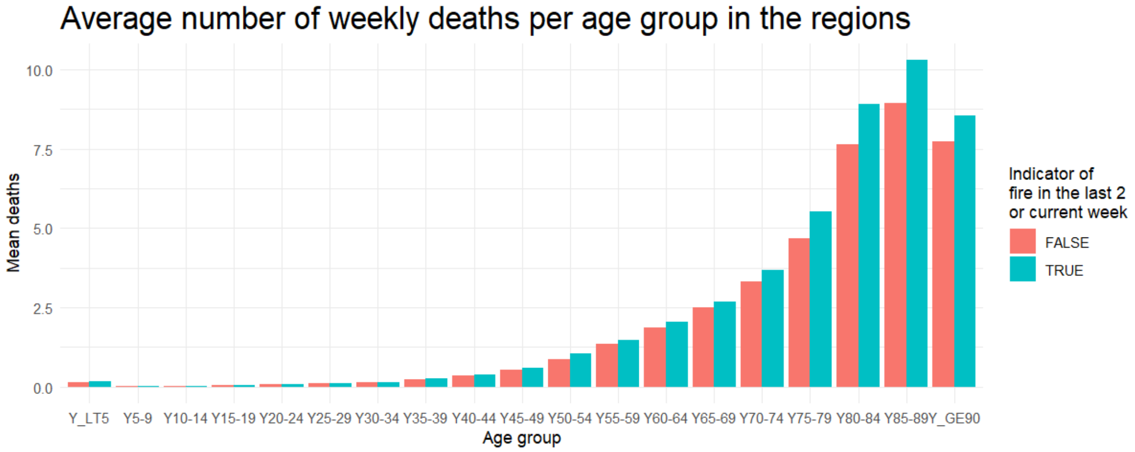

Figure 32: Barplot of weekly mean deaths in the regions by the indicator whether there was a wildfire in that or the preceding two weeks, separately for the age groups

In addition to the lower values for younger people, the difference between the two groups is very small, but the average is higher for all but five to nine year old children when there has been a fire. The differences are also more apparent for older people. The question is whether the difference is significant. When the values are very low, an increase of one unit is a big change in relative terms, but in this case it could easily be due to an accident, for example. Chance may play a large role here, as one such death may fall in a fire season or a non-fire season, which may be negligible for the population as a whole or divided just by the two genders.

To test the difference, I used a two-sample one-sided Poisson test, which is analogous to the two-sample one-sided Student's t-test, but while the former is applied to count data, the latter is for normally distributed data. In all cases, my alternative hypothesis is that where there was a fire in the last two or the current week, the average mortality is higher than in smoke-free periods. The tests for each age were bootstrapped with 500 resamples to estimate the distribution of p-values. I set the significance level at the usual 0.05. Here, values lower than this mean that the fire period values are significantly larger. The results are shown in Figure 33. We can say that something similar was to be expected. Between roughly 5 and 34 years, the p-values, although not uniformly distributed, show that there is no significant difference between the two means. For those under 5 years and between 35 and 49 years, most values are below 0.05, with the interquartile range filling this interval, but there are outliers up to 1. For those aged between 50 and 74 years, the p-values were clearly close to 0, but even here outliers are visible. And for those 75 years and older, very small p-values were obtained in all cases. These show that the youngest and, from about 40 years of age, the older people are at greater risk from forest fire-related air pollution.

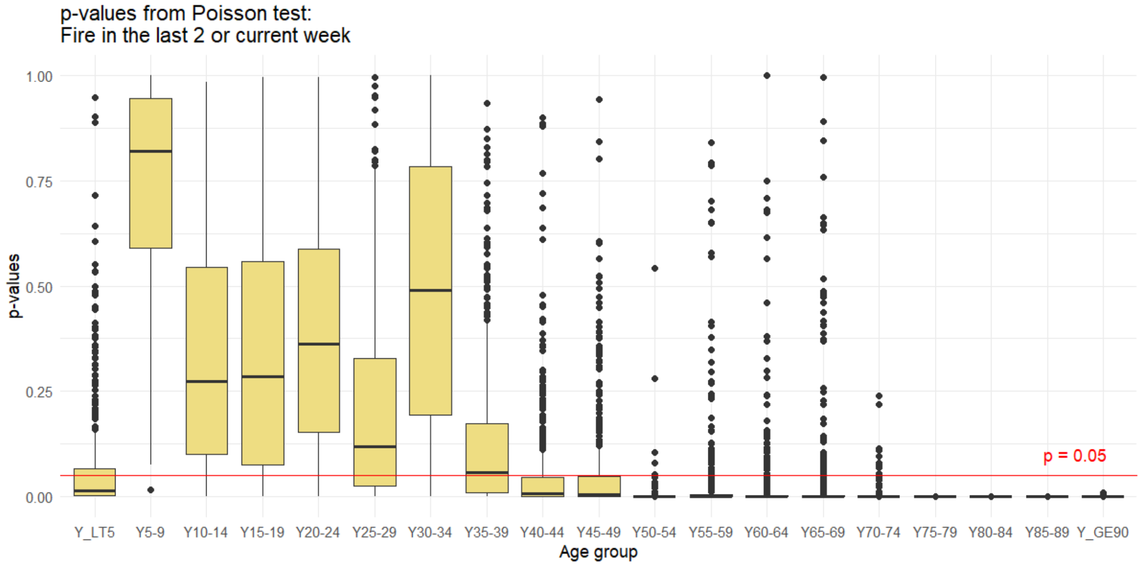

Figure 33: p-values from the Poisson test, whether mortality is higher when burning emissions are released into the air

Although Poisson tests give logical results, the regression did not seem reliable. In my opinion, a different data structure would be needed, e.g. filter out accidents and other sudden deaths, create fewer groups (wider age intervals), and consider a coarser temporal resolution such that two or three weeks.

6 Summary

In my thesis, I explored a topical issue that is perhaps increasingly worth paying attention to. Climate change affects us all and have impacts on many aspects of life. It is also a major cause of the increasing frequency of wildfires. Of the many potential sources of damage, I have been looking at air pollution and, more specifically, its impact on mortality.

From a mathematical point of view, the study of this topic involves modeling count data (number of deaths). A common method is Poisson regression, which is a generalized linear model. However, for real data, the equidispersion condition of the Poisson regression is often not met, and we can observe overdispersion in the data. In this case, an appropriate method is the Quasi-Poisson model. Furthermore, since I have examined consecutive weekly data, they can be considered as a time series, i.e. adjacent observations may be correlated.

For my studies, I chose Greece, a country with a Mediterranean climate, where wildfires are relatively not rare. For data availability and to avoid the coronavirus epidemic, I looked at the 5-year period from 2015 to 2019, each year covering the weeks approximately from May to October.

The models included five smoke-related variables: carbon dioxide, organic carbon, nitrogen oxides, fine particles, and total particulate matter. I also added 2 weather variables, mean temperature, and relative humidity. Because of the strong correlation, I conducted several tests, first using all smoke variables but only looking at their cumulative effect, and then including only one type of emission at a time. These data were available as fluxes, so I tried two methods: first, I took 3-week moving averages, then I took the respective weekly and one- and two-weekly lags as separate variables. In terms of deaths, I conducted three rounds of analyses: for the whole population, then sex, and finally age specifically.

My results show that in Greece, wildfire smoke is responsible for an average of 100-150 deaths a year. In some cases, higher and lower values have been found. Of the variables, nitrogen oxides and organic carbon showed smaller effects on mortality, while carbon dioxide and the two types of particulate matter had larger effects. Of course, there is a lot of uncertainty in the modeling, as wildfires are also associated with hot and dry weather, and other unobserved causes of death may be behind them. Gender-specific studies have shown that although both sexes experience an increase in deaths from this type of air pollution, women are more affected. This is in agreement with most of the findings in the literature on the effect of air pollution. Finally, I performed a Poisson test on bootstrap samples for age groups, since frequent 0 values due to many groups make the modeling highly unstable. Based on the p-values, people under 5 years of age and those who are older are at higher risk.

A development option in modeling is to use the actual amount in the air instead of or in addition to the fluxes of the smoke variables. There is a publicly available database, at the resolution I use, of the air composition, but it needs to be estimated how much comes from wildfires. For this purpose, 3-dimensional chemical transport models can be used, which are able to model the spread of the emitted substances under given weather conditions. It is possible, that strong wind or rain can cause particles to disappear quickly over a region or that they can remain there for several weeks or be blown in from other surrounding areas. Thus, with the amount of particles in the air at the time, the effect of wildfire smoke on mortality might be even more precisely detectable.

Another research question could also be to look at gender and age together. There are studies ([6]) on air pollution in general that show that, for example, in children, it has a stronger effect on boys in early life and on girls in late childhood. In adults and older age, women have generally been shown to be more at risk.

References

[1] *Atmosphere monitoring service.* (n.d.). Copernicus. Retrieved February 23, 2025, from https://atmosphere.copernicus.eu/greece

[2] Black, C., Tesfaigzi, Y., Bassein, J. A., & Miller, L. A. (2017). Wildfire smoke exposure and human health: Significant gaps in research for a growing public health issue. *Environmental Toxicology and Pharmacology*, *55*, 186–195. https://doi.org/10.1016/j.etap.2017.08.022

[3] Chen, G., Guo, Y., Yue, X., Tong, S., Gasparrini, A., Bell, M. L., Armstrong, B., Schwartz, J., Jaakkola, J. J. K., Zanobetti, A., Lavigne, E., Nascimento Saldiva, P. H., Kan, H., Royé, D., Milojevic, A., Overcenco, A., Urban, A., Schneider, A., Entezari, A., ... Li, S. (2021). Mortality risk attributable to wildfire-related $PM_{2.5}$ pollution: A global time series study in 749 locations. *The Lancet Planetary Health*, *5*(9), e579–e587. https://doi.org/10.1016/S2542-5196(21)00200-X

[4] Chowdhury, S., Hänninen, R., Sofiev, M., & Aunan, K. (2024). Fires as a source of annual ambient $PM_{2.5}$ exposure and chronic health impacts in europe. *Science of The Total Environment*, *922*, 171314. https://doi.org/10.1016/j.scitotenv.2024.171314

[5] *Climate and energy in the eu.* (2021). European Environment Agency. Retrieved February 21, 2025, from https://climate-energy.eea.europa.eu/countries/greece

[6] Clougherty, J. E. (2010). A growing role for gender analysis in air pollution epidemiology. *Environmental Health Perspectives*, *118*(2), 167–176. https://doi.org/10.1289/ehp.0900994

[7] *Greece – air pollution country fact sheet 2024.* (2024). European Environment Agency. Retrieved February 23, 2025, from https://www.eea.europa.eu/en/topics/in-depth/air-pollution/air-pollution-country-fact-sheets-2024/greece-air-pollution-country-fact-sheet-2024

[8] *Hellenic national meteorogical service.* (n.d.). Retrieved February 21, 2025, from https://emy.gr/en/the-climate-of-greece

[9] Hilbe, J. M. (2014). *Modeling count data.* Cambridge University Press. https://doi.org/10.1017/CBO9781139236065

[10] Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, *142*, 319–327. https://doi.org/https://doi.org/10.1016/j.envres.2015.06.040

[11] Liu, J. C., Wilson, A., Mickley, L. J., Dominici, F., Ebisu, K., Wang, Y., Sulprizio, M. P., Peng, R. D., Yue, X., Son, J.-Y., Anderson, G. B., & Bell, M. L. (2017). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. *Epidemiology*, *28*(1), 77–85. https://doi.org/10.1097/EDE.0000000000000556

[12] McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736

[13] Nguyen, M. (2020). *A guide on data analysis*. Bookdown. https://bookdown.org/mike/data_analysis/

[14] Roback, P., & Legler, J. (2021). *Beyond multiple linear regression: Applied generalized linear models and multilevel models in R* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9780429066665

[15] Sokolik, I. N., Soja, A. J., DeMott, P. J., & Winker, D. (2019). Progress and challenges in quantifying wildfire smoke emissions, their properties, transport, and atmospheric impacts. *J. Geophys. Res.*, *124*(23), 13005–13025. https://doi.org/10.1029/2018JD029878

[16] Yadav, B., Jeyaseelan, L., Jeyaseelan, V., Durairaj, J., George, S., Selvaraj, K., & Bangdiwala, S. I. (2021). Can generalized poisson model replace any other count data models? an evaluation. *Clinical Epidemiology and Global Health*, *11*, 100774. https://doi.org/10.1016/j.cegh.2021.100774

7 Appendix

Output of the relative humidity model from subsection 4.2:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	433.6252	154.2878	2.81	0.0050 **
geoEL303	3.0149	1.0198	2.96	0.0031 **
geoEL305	1.7814	1.0182	1.75	0.0803 .
geoEL306	1.9171	1.0407	1.84	0.0655 .
geoEL307	1.4194	1.0240	1.39	0.1658
geoEL411	-0.5136	1.0169	-0.51	0.6135
geoEL412	-2.4481	1.1763	-2.08	0.0375 *
geoEL413	-5.0642	1.1765	-4.30	0.0000 ***
geoEL421	-0.7579	1.1820	-0.64	0.5214
geoEL422	-3.6200	1.1755	-3.08	0.0021 **
geoEL431	-5.5035	1.1749	-4.68	0.0000 ***
geoEL432	-5.7777	1.1763	-4.91	0.0000 ***
geoEL433	-3.8922	1.1777	-3.30	0.0010 ***
geoEL434	-1.2561	1.0236	-1.23	0.2198
geoEL511	0.9716	1.0342	0.94	0.3476
geoEL512	-2.7485	1.0428	-2.64	0.0084 **
geoEL513	-2.2304	1.0281	-2.17	0.0301 *
geoEL514	-1.6102	1.1133	-1.45	0.1481
geoEL515	-3.0423	1.0271	-2.96	0.0031 **
geoEL521	1.2665	1.0775	1.18	0.2399
geoEL522	0.0861	1.0446	0.08	0.9343
geoEL523	0.5839	1.0516	0.56	0.5788
geoEL524	1.6547	1.0904	1.52	0.1292
geoEL525	-0.0626	1.0472	-0.06	0.9524
geoEL526	0.8867	1.0606	0.84	0.4032
geoEL527	-0.9209	1.0246	-0.90	0.3688
geoEL531	0.2577	1.0977	0.23	0.8144
geoEL532	0.7663	1.1302	0.68	0.4978
geoEL533	0.1517	1.1027	0.14	0.8906
geoEL541	2.2722	1.0418	2.18	0.0292 *
geoEL542	2.7086	1.0392	2.61	0.0092 **
geoEL543	2.5591	1.1365	2.25	0.0244 *
geoEL611	1.0320	1.0797	0.96	0.3392
geoEL612	0.0875	1.0436	0.08	0.9332
geoEL613	-1.2293	1.0310	-1.19	0.2332
geoEL621	1.9650	1.0162	1.93	0.0532 .
geoEL622	3.3341	1.0188	3.27	0.0011 **
geoEL623	-0.9145	1.0170	-0.90	0.3686
geoEL624	0.0571	1.0168	0.06	0.9552

geoEL631	2.5426	1.0390	2.45	0.0144	*
geoEL632	1.4212	1.0735	1.32	0.1856	
geoEL633	3.3145	1.0318	3.21	0.0013	**
geoEL641	0.5981	1.0367	0.58	0.5640	
geoEL642	0.7144	1.0232	0.70	0.4851	
geoEL643	-2.0913	1.0868	-1.92	0.0544	.
geoEL644	-0.1157	1.0530	-0.11	0.9125	
geoEL645	-1.5435	1.0841	-1.42	0.1546	
geoEL651	0.8131	1.0708	0.76	0.4477	
geoEL652	-0.1897	1.0534	-0.18	0.8571	
geoEL653	2.3982	1.0591	2.26	0.0236	*
year	-0.1591	0.0766	-2.08	0.0378	*
week21	6.9246	0.7156	9.68	0.0000	***
week22	3.1228	0.7324	4.26	0.0000	***
week23	1.2097	0.7557	1.60	0.1095	
week24	6.0488	0.8044	7.52	0.0000	***
week25	12.1468	0.8408	14.45	0.0000	***
week26	7.9279	0.8415	9.42	0.0000	***
week27	9.1913	0.8743	10.51	0.0000	***
week28	10.5779	0.8908	11.87	0.0000	***
week29	5.3399	0.8684	6.15	0.0000	***
week30	11.1042	0.9080	12.23	0.0000	***
week31	10.6908	0.9687	11.04	0.0000	***
week32	7.9731	0.9317	8.56	0.0000	***
week33	8.2358	0.8922	9.23	0.0000	***
week34	5.7980	0.8615	6.73	0.0000	***
week35	2.4890	0.8533	2.92	0.0035	**
week36	7.5433	0.8231	9.16	0.0000	***
week37	0.4984	0.7948	0.63	0.5307	
week38	0.1633	0.7518	0.22	0.8281	
week39	-0.1540	0.7112	-0.22	0.8286	
week40	9.4462	0.7088	13.33	0.0000	***
week41	-0.1670	0.7123	-0.23	0.8146	
week42	-3.3780	0.7136	-4.73	0.0000	***
tg	-0.6001	0.1072	-5.60	0.0000	***
tx	-1.1816	0.0797	-14.83	0.0000	***

Table 11: Model output for lm(relative humidity ∼ as.factor(region) + year + as.factor(week) + mean temperature + max temperature

Model outputs from subsection 5.1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.1981	0.0416	100.99	0.0000 ***
geoEL302	-0.2429	0.0148	-16.39	0.0000 ***
geoEL303	0.7513	0.0120	62.65	0.0000 ***
geoEL304	-0.0485	0.0140	-3.46	0.0005 ***
geoEL305	-0.3309	0.0151	-21.94	0.0000 ***
geoEL306	-1.4184	0.0221	-64.18	0.0000 ***
geoEL307	0.0408	0.0137	2.99	0.0028 **
geoEL411	-1.5599	0.0233	-66.99	0.0000 ***
geoEL412	-2.4136	0.0337	-71.57	0.0000 ***
geoEL413	-2.2717	0.0319	-71.21	0.0000 ***
geoEL421	-1.3199	0.0211	-62.66	0.0000 ***
geoEL422	-1.6220	0.0240	-67.69	0.0000 ***
geoEL431	-0.7184	0.0172	-41.68	0.0000 ***
geoEL432	-1.8655	0.0270	-68.98	0.0000 ***
geoEL433	-2.0269	0.0290	-69.89	0.0000 ***
geoEL434	-1.2680	0.0212	-59.73	0.0000 ***
geoEL511	-1.1094	0.0196	-56.66	0.0000 ***
geoEL512	-1.6497	0.0251	-65.77	0.0000 ***
geoEL513	-1.4406	0.0226	-63.65	0.0000 ***
geoEL514	-1.3639	0.0233	-58.57	0.0000 ***
geoEL515	-1.1342	0.0202	-56.10	0.0000 ***
geoEL521	-1.1575	0.0205	-56.47	0.0000 ***
geoEL522	0.6618	0.0121	54.52	0.0000 ***
geoEL523	-1.5559	0.0236	-65.81	0.0000 ***
geoEL524	-1.1058	0.0202	-54.67	0.0000 ***
geoEL525	-1.2898	0.0214	-60.17	0.0000 ***
geoEL526	-0.7300	0.0174	-42.00	0.0000 ***
geoEL527	-1.5552	0.0236	-65.78	0.0000 ***
geoEL531	-0.9165	0.0198	-46.26	0.0000 ***
geoEL532	-2.1877	0.0334	-65.45	0.0000 ***
geoEL533	-2.0977	0.0317	-66.14	0.0000 ***
geoEL541	-1.2211	0.0209	-58.47	0.0000 ***
geoEL542	-2.4049	0.0345	-69.69	0.0000 ***
geoEL543	-1.0724	0.0212	-50.62	0.0000 ***
geoEL611	-0.4900	0.0168	-29.23	0.0000 ***
geoEL612	-0.6102	0.0169	-36.20	0.0000 ***
geoEL613	-0.9202	0.0186	-49.38	0.0000 ***
geoEL621	-2.5848	0.0367	-70.34	0.0000 ***
geoEL622	-1.4840	0.0227	-65.45	0.0000 ***
geoEL623	-2.5068	0.0360	-69.63	0.0000 ***
geoEL624	-2.8256	0.0418	-67.57	0.0000 ***

geoEL631	-0.8275	0.0180	-45.87	0.0000	***
geoEL632	-0.5313	0.0169	-31.52	0.0000	***
geoEL633	-1.1130	0.0198	-56.27	0.0000	***
geoEL641	-1.4769	0.0228	-64.73	0.0000	***
geoEL642	-0.8537	0.0179	-47.72	0.0000	***
geoEL643	-3.1400	0.0511	-61.44	0.0000	***
geoEL644	-1.0276	0.0196	-52.48	0.0000	***
geoEL645	-2.4801	0.0372	-66.59	0.0000	***
geoEL651	-0.8846	0.0187	-47.26	0.0000	***
geoEL652	-1.1371	0.0204	-55.67	0.0000	***
geoEL653	-0.5654	0.0166	-34.15	0.0000	***
as.factor(week)21	-0.0499	0.0151	-3.30	0.0010	***
as.factor(week)22	-0.0398	0.0152	-2.61	0.0090	**
as.factor(week)23	-0.0722	0.0157	-4.61	0.0000	***
as.factor(week)24	-0.0773	0.0165	-4.69	0.0000	***
as.factor(week)25	-0.0492	0.0175	-2.81	0.0049	**
as.factor(week)26	-0.0851	0.0176	-4.84	0.0000	***
as.factor(week)27	-0.1103	0.0182	-6.07	0.0000	***
as.factor(week)28	-0.1244	0.0185	-6.72	0.0000	***
as.factor(week)29	-0.1667	0.0181	-9.20	0.0000	***
as.factor(week)30	-0.1253	0.0189	-6.63	0.0000	***
as.factor(week)31	-0.1321	0.0201	-6.58	0.0000	***
as.factor(week)32	-0.0907	0.0192	-4.74	0.0000	***
as.factor(week)33	-0.1268	0.0185	-6.87	0.0000	***
as.factor(week)34	-0.1634	0.0180	-9.10	0.0000	***
as.factor(week)35	-0.1622	0.0177	-9.17	0.0000	***
as.factor(week)36	-0.1446	0.0173	-8.38	0.0000	***
as.factor(week)37	-0.1593	0.0166	-9.59	0.0000	***
as.factor(week)38	-0.1351	0.0157	-8.58	0.0000	***
as.factor(week)39	-0.0795	0.0151	-5.28	0.0000	***
as.factor(week)40	-0.0227	0.0151	-1.51	0.1323	
as.factor(week)41	0.0198	0.0148	1.34	0.1812	
as.factor(week)42	0.0350	0.0148	2.37	0.0180	*
mean_temp	0.0257	0.0016	16.31	0.0000	***
rel_hum	-0.0006	0.0003	-2.37	0.0176	*
PM2.5	0.0286	0.0247	1.16	0.2460	

Table 12: Model output for glm(deaths $\sim$ region + as.factor(week) + mean temperature + relative humidity + $\mathrm{PM}_{2.5}$)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0977	0.0467	87.82	0.0000 ***
geoEL302	-0.2504	0.0155	-16.19	0.0000 ***
geoEL303	0.7566	0.0125	60.45	0.0000 ***
geoEL304	-0.0531	0.0146	-3.63	0.0003 ***
geoEL305	-0.3308	0.0157	-21.04	0.0000 ***
geoEL306	-1.4213	0.0231	-61.65	0.0000 ***
geoEL307	0.0470	0.0142	3.30	0.0010 ***
geoEL411	-1.5596	0.0243	-64.28	0.0000 ***
geoEL412	-2.4228	0.0353	-68.69	0.0000 ***
geoEL413	-2.2697	0.0332	-68.27	0.0000 ***
geoEL421	-1.3256	0.0220	-60.36	0.0000 ***
geoEL422	-1.6135	0.0249	-64.81	0.0000 ***
geoEL431	-0.7109	0.0179	-39.63	0.0000 ***
geoEL432	-1.8581	0.0282	-65.98	0.0000 ***
geoEL433	-2.0207	0.0302	-66.92	0.0000 ***
geoEL434	-1.2640	0.0221	-57.10	0.0000 ***
geoEL511	-1.1070	0.0204	-54.26	0.0000 ***
geoEL512	-1.6357	0.0261	-62.65	0.0000 ***
geoEL513	-1.4334	0.0236	-60.77	0.0000 ***
geoEL514	-1.3482	0.0243	-55.50	0.0000 ***
geoEL515	-1.1245	0.0211	-53.38	0.0000 ***
geoEL521	-1.1510	0.0214	-53.82	0.0000 ***
geoEL522	0.6657	0.0127	52.56	0.0000 ***
geoEL523	-1.5528	0.0247	-62.96	0.0000 ***
geoEL524	-1.0963	0.0211	-51.97	0.0000 ***
geoEL525	-1.2862	0.0224	-57.48	0.0000 ***
geoEL526	-0.7267	0.0181	-40.05	0.0000 ***
geoEL527	-1.5461	0.0246	-62.84	0.0000 ***
geoEL531	-0.9008	0.0207	-43.62	0.0000 ***
geoEL532	-2.1853	0.0351	-62.32	0.0000 ***
geoEL533	-2.0719	0.0329	-62.94	0.0000 ***
geoEL541	-1.2194	0.0218	-55.95	0.0000 ***
geoEL542	-2.4005	0.0360	-66.76	0.0000 ***
geoEL543	-1.0689	0.0222	-48.22	0.0000 ***
geoEL611	-0.4820	0.0175	-27.55	0.0000 ***
geoEL612	-0.6094	0.0176	-34.59	0.0000 ***
geoEL613	-0.9142	0.0194	-47.05	0.0000 ***
geoEL621	-2.5766	0.0381	-67.60	0.0000 ***
geoEL622	-1.4944	0.0237	-62.99	0.0000 ***
geoEL623	-2.4909	0.0373	-66.78	0.0000 ***
geoEL624	-2.8331	0.0438	-64.67	0.0000 ***
geoEL631	-0.8161	0.0188	-43.49	0.0000 ***

geoEL632	-0.5266	0.0176	-29.92	0.0000	***
geoEL633	-1.0987	0.0205	-53.52	0.0000	***
geoEL641	-1.4790	0.0238	-62.06	0.0000	***
geoEL642	-0.8539	0.0187	-45.76	0.0000	***
geoEL643	-3.1154	0.0530	-58.77	0.0000	***
geoEL644	-1.0202	0.0204	-49.97	0.0000	***
geoEL645	-2.4741	0.0389	-63.55	0.0000	***
geoEL651	-0.8718	0.0195	-44.77	0.0000	***
geoEL652	-1.1219	0.0212	-52.86	0.0000	***
geoEL653	-0.5650	0.0173	-32.69	0.0000	***
as.factor(week)23	-0.0354	0.0146	-2.42	0.0157	*
as.factor(week)24	-0.0448	0.0149	-3.00	0.0027	**
as.factor(week)25	-0.0199	0.0155	-1.28	0.2000	
as.factor(week)26	-0.0568	0.0156	-3.64	0.0003	***
as.factor(week)27	-0.0837	0.0161	-5.21	0.0000	***
as.factor(week)28	-0.0989	0.0163	-6.05	0.0000	***
as.factor(week)29	-0.1402	0.0161	-8.72	0.0000	***
as.factor(week)30	-0.1031	0.0167	-6.18	0.0000	***
as.factor(week)31	-0.1140	0.0177	-6.45	0.0000	***
as.factor(week)32	-0.0689	0.0168	-4.10	0.0000	***
as.factor(week)33	-0.1022	0.0163	-6.27	0.0000	***
as.factor(week)34	-0.1367	0.0160	-8.56	0.0000	***
as.factor(week)35	-0.1351	0.0158	-8.54	0.0000	***
as.factor(week)36	-0.1144	0.0155	-7.39	0.0000	***
as.factor(week)37	-0.1268	0.0152	-8.36	0.0000	***
as.factor(week)38	-0.0976	0.0149	-6.55	0.0000	***
as.factor(week)39	-0.0339	0.0152	-2.23	0.0258	*
as.factor(week)40	0.0246	0.0153	1.61	0.1070	
as.factor(week)41	0.0687	0.0154	4.47	0.0000	***
as.factor(week)42	0.0851	0.0156	5.44	0.0000	***
mean_temp	0.0287	0.0017	17.39	0.0000	***
rel_hum	-0.0007	0.0003	-2.52	0.0117	*
PM2.5_m	0.0936	0.0412	2.27	0.0233	*

Table 13: Model output for glm(deaths $\sim$ region + as.factor(week) + mean temperature + relative humidity + 3-week moving average of $\text{PM}_{2.5}$)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0972	0.0467	87.73	0.0000 ***
geoEL302	-0.2504	0.0155	-16.19	0.0000 ***
geoEL303	0.7566	0.0125	60.44	0.0000 ***
geoEL304	-0.0531	0.0146	-3.63	0.0003 ***
geoEL305	-0.3308	0.0157	-21.04	0.0000 ***
geoEL306	-1.4213	0.0231	-61.63	0.0000 ***
geoEL307	0.0470	0.0142	3.30	0.0010 ***
geoEL411	-1.5596	0.0243	-64.27	0.0000 ***
geoEL412	-2.4228	0.0353	-68.67	0.0000 ***
geoEL413	-2.2697	0.0333	-68.26	0.0000 ***
geoEL421	-1.3256	0.0220	-60.35	0.0000 ***
geoEL422	-1.6135	0.0249	-64.80	0.0000 ***
geoEL431	-0.7108	0.0179	-39.62	0.0000 ***
geoEL432	-1.8581	0.0282	-65.97	0.0000 ***
geoEL433	-2.0207	0.0302	-66.91	0.0000 ***
geoEL434	-1.2640	0.0221	-57.09	0.0000 ***
geoEL511	-1.1070	0.0204	-54.25	0.0000 ***
geoEL512	-1.6357	0.0261	-62.64	0.0000 ***
geoEL513	-1.4334	0.0236	-60.76	0.0000 ***
geoEL514	-1.3481	0.0243	-55.48	0.0000 ***
geoEL515	-1.1245	0.0211	-53.37	0.0000 ***
geoEL521	-1.1510	0.0214	-53.81	0.0000 ***
geoEL522	0.6658	0.0127	52.55	0.0000 ***
geoEL523	-1.5528	0.0247	-62.94	0.0000 ***
geoEL524	-1.0962	0.0211	-51.96	0.0000 ***
geoEL525	-1.2861	0.0224	-57.47	0.0000 ***
geoEL526	-0.7267	0.0181	-40.04	0.0000 ***
geoEL527	-1.5460	0.0246	-62.82	0.0000 ***
geoEL531	-0.9008	0.0207	-43.61	0.0000 ***
geoEL532	-2.1853	0.0351	-62.31	0.0000 ***
geoEL533	-2.0718	0.0329	-62.93	0.0000 ***
geoEL541	-1.2194	0.0218	-55.94	0.0000 ***
geoEL542	-2.4005	0.0360	-66.75	0.0000 ***
geoEL543	-1.0689	0.0222	-48.21	0.0000 ***
geoEL611	-0.4820	0.0175	-27.54	0.0000 ***
geoEL612	-0.6094	0.0176	-34.58	0.0000 ***
geoEL613	-0.9142	0.0194	-47.04	0.0000 ***
geoEL621	-2.5766	0.0381	-67.58	0.0000 ***
geoEL622	-1.4944	0.0237	-62.98	0.0000 ***
geoEL623	-2.4909	0.0373	-66.77	0.0000 ***
geoEL624	-2.8330	0.0438	-64.66	0.0000 ***
geoEL631	-0.8161	0.0188	-43.48	0.0000 ***

geoEL632	-0.5266	0.0176	-29.92	0.0000	***
geoEL633	-1.0987	0.0205	-53.51	0.0000	***
geoEL641	-1.4790	0.0238	-62.05	0.0000	***
geoEL642	-0.8539	0.0187	-45.76	0.0000	***
geoEL643	-3.1154	0.0530	-58.76	0.0000	***
geoEL644	-1.0202	0.0204	-49.96	0.0000	***
geoEL645	-2.4741	0.0389	-63.54	0.0000	***
geoEL651	-0.8717	0.0195	-44.76	0.0000	***
geoEL652	-1.1219	0.0212	-52.85	0.0000	***
geoEL653	-0.5650	0.0173	-32.68	0.0000	***
as.factor(week)23	-0.0354	0.0146	-2.42	0.0157	*
as.factor(week)24	-0.0448	0.0149	-3.00	0.0027	**
as.factor(week)25	-0.0200	0.0155	-1.28	0.1989	
as.factor(week)26	-0.0568	0.0156	-3.64	0.0003	***
as.factor(week)27	-0.0838	0.0161	-5.21	0.0000	***
as.factor(week)28	-0.0989	0.0164	-6.05	0.0000	***
as.factor(week)29	-0.1397	0.0161	-8.66	0.0000	***
as.factor(week)30	-0.1037	0.0167	-6.19	0.0000	***
as.factor(week)31	-0.1139	0.0177	-6.43	0.0000	***
as.factor(week)32	-0.0691	0.0168	-4.11	0.0000	***
as.factor(week)33	-0.1021	0.0163	-6.27	0.0000	***
as.factor(week)34	-0.1368	0.0160	-8.57	0.0000	***
as.factor(week)35	-0.1351	0.0158	-8.54	0.0000	***
as.factor(week)36	-0.1144	0.0155	-7.39	0.0000	***
as.factor(week)37	-0.1269	0.0152	-8.36	0.0000	***
as.factor(week)38	-0.0976	0.0149	-6.55	0.0000	***
as.factor(week)39	-0.0339	0.0152	-2.23	0.0259	*
as.factor(week)40	0.0246	0.0153	1.61	0.1072	
as.factor(week)41	0.0688	0.0154	4.47	0.0000	***
as.factor(week)42	0.0851	0.0156	5.44	0.0000	***
mean_temp	0.0287	0.0017	17.38	0.0000	***
rel_hum	-0.0007	0.0003	-2.50	0.0123	*
PM2.5	0.0239	0.0246	0.97	0.3325	
PM2.5_1	0.0393	0.0238	1.65	0.0991	.
PM2.5_2	0.0297	0.0239	1.24	0.2144	

Table 14: Model output for glm(deaths $\sim$ region + as.factor(week) + mean temperature + relative humidity + $\mathrm{PM}_{2.5}$ + lag 1 of $\mathrm{PM}_{2.5}$ + lag 2 of $\mathrm{PM}_{2.5}$

<table>
<tr><td colspan="4">I, Noémi Takács, confirm that during the writing of my thesis, I used the given AI-based tools to perform the tasks listed below:</td></tr>
<tr><td>Task</td><td>Tool</td><td>Where to use</td><td>Comment</td></tr>
<tr><td>Checking grammar, translating words</td><td>DeepL Translate</td><td>entire thesis</td><td></td></tr>
<tr><td colspan="4">I have not used any AI-based tools other than those listed.</td></tr>
</table>