Institut de Recerca en Economia Aplicada Regional i Pública
*Research Institute of Applied Economics*

Document de Treball 2025/22 1/34 pág.
*Working Paper 2025/22 1/34 pag.*

# Integrating Road Safety and Environmental Impact via Telematics: Modeling Traffic Accident Risk Using Vehicle Emissions

Juan Sebastian Yanez, Montserrat Guillen, Paulina Roszkowsk and Jens Perch Nielsen

UBIREA Institut de Recerca en Economia Aplicada Regional i Pública

**Abstract**

Private vehicles harm public health by contributing to air pollution and traffic accidents, the leading cause of death among young adults. Despite these risks, drivers often ignore speed limits, while society increasingly prioritizes environmental protection. This tension between personal habits and collective responsibility highlights the urgent need for strategies to promote safer driving practices. Therefore, this paper introduces a novel approach to evaluating road crash risk using air pollutants as exposure measures, so drivers are simultaneously encouraged to reduce their environmental footprint and mitigate their road crash risk. We use a rich dataset of over 1,500 at-fault crash-related claims recorded over two years provided by an insurance company, merged with detailed telematics driving data for individual vehicles. We show that available emission factor models enable the integration of emission-based exposure measures to model road crash risk. Then, we provide empirical evidence that incorporating behavioral telematics data makes pollutant-driven models as efficient as traditional distance-driven ones. Our proposition has the potential to enhance road safety and reduce air pollution by directly linking environmentally conscious driving practices with reducing road crash risks.

**JEL Classification:** G22, G52

**Keywords:** air pollution; crash risk; public health; road accident; telematics; transportation

**Authors:**

**Juan Sebastián Yáñez,** Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Barcelona, Spain. yanez.juansebastian@gmail.com

**Montserrat Guillén,** Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Barcelona, Spain – mguillen@ub.edu

**Paulina Roszkowska**, Bayes Business School, City, University of London, London, United Kingdom. paulina.roszkowska@city.ac.uk

**Jens Perch Nielsen,** Bayes Business School, City, University of London, London, United Kingdom. Jens.Nielsen.1@city.ac.uk

**Acknowledgements:** Juan Sebastian Yanez gratefully acknowledges financial support from the Fonds de Recherche du Québec Nature et Technologie (FRQNT, grant B3X-346658). We thank the support received by the Spanish Ministry of Science, Agencia Estatal de Investigación PID2024-157911NB-I00/AEI/10.13039/501100011033 and ICREA Academia. We acknowledge support from the Departament de Recerca i Universitats, the Departament d'Acció Climàtica, Alimentació i Agenda Rural, and the Fons Climàtic of the Generalitat de Catalunya (2023 CLIMA 00012).

# 1 Introduction

Private vehicles pose a growing threat to public health. They significantly contribute to accident-related deaths, creating serious socioeconomic challenges at both individual and national levels. According to the WHO (2023), road traffic injuries are the leading cause of death among children and young adults aged 5–29 years, highlighting their devastating personal impact. These injuries disproportionately affect working-age individuals as well. For instance, the European Commission reported that in 2023, 46% of road fatalities in the EU occurred among those aged 15–49 (European Commission (2024)). On a broader scale, road traffic crashes impose substantial economic costs, estimated at approximately between 1% and 3% of the gross domestic product in most countries (WHO (2023)). With traffic accidents ranking as the top cause of death among young adults, private vehicles' impact is undeniable. Yet, despite extensive efforts and campaigns to promote road safety, risky driving behaviors persist, intensifying the associated dangers.

At the same time, there is a shift in societal priorities, as more people grow increasingly committed to protecting the environment. While everyone agrees on the general goal, individuals do not always see their immediate impact on the environment. The land transportation sector is a significant contributor to air pollution, with private vehicles playing a substantial role in global carbon dioxide ($CO_2$) emissions, the greenhouse gas that drives climate change. In 2022, the domestic transport sector accounted for approximately 23.77% of total $CO_2$ emissions in the EU (EEA (2024)). Beyond $CO_2$, this sector also emits other harmful pollutants identified by the World Health Organization (WHO), such as nitrogen oxides ($NO_x$) and carbon monoxide (CO) (WHO (2021)). Urban areas, where motor vehicle density is highest, face the most severe impacts. For example, in 2022, 97% of the EU's urban population was exposed to fine particulate matter levels exceeding the updated WHO guidelines issued in 2021 (EEA (2021)). The increasing popularity of Electric Vehicles (EVs) confers an alternative to reduce emissions; however, currently, in many parts of the world, they are outnumbered by their Internal Combustion Engine (ICE) counterparts. For example, in all 27 EU Member States in 2023, Electric vehicles only accounted for 22.7% of new car registrations (EEA (2024)).

This tension between personal habits (driving behavior) and collective responsibility (care for the environment) highlights the urgent need to explore practical ways to encourage safer driving practices. Although there has been ample academic discussion about the risk factors that lead to road accidents and the effect of driving on emissions, to the best of our knowledge, no study merges road safety with air pollution emission levels.

In this paper, we marry these two pivotal drivers of public health: road safety and air pollution. We achieve this goal by modeling road accident risk as a function of particle emissions, replacing the distance or the observation period (often referred to as 'risk exposure' in academic literature). This is possible by including and adapting telematics data gathered through On-Board Devices (OBD) and smartphone applications and providing information regarding driving behavior and road conditions during specific trips. Typical examples are the kilometers driven (itself an exposure measure), the type of road (urban, rural, etc..), the speed, harsh braking events, and accelerations. In the pollutant emission estimations research, telematics data proved useful because they provide information regarding two of the three key factors that affect vehicle emissions: driving behavior and road conditions (the other factor being vehicle characteristics). In accident prevention research, telematics data have allowed researchers to identify behavioral risk factors such as the number of harsh accelerations and nighttime driving. Moreover, using telematics data researchers evaluated the exposure to risk with more precise longitudinal data such as the distance driven. We acknowledge our approach is not well adapted for EVs; nevertheless, we believe it is relevant to focus on ICE vehicles for two reasons: they represent the majority of cars in many parts of the world, such as the EU (see EEA (2024)) and there is evidence of a higher frequency of accidents caused by EVs (see McDonnell *et al.* (2024)).

Our modeling is based on two stages. First, we use a macroscopic model adapted to our data to estimate the level of emissions. Hereto, we benefit from having access to an extensive data set containing two years' worth of telematics data from 17,405 drivers and the number of at-fault accidents. Specifically, we convert the telematics data into particle emission estimations. Then, we propose several models (simple to more complex) to predict the risk of road accidents. In the model that is our ultimate proposition, we replace typical risk exposure measurements, such as the kilometers driven or the observation period, with the amount of pollutant particles emitted.

In the next step, we implement our models in a case study and provide numerical results. We show that decreasing air pollution emissions is compatible with reducing road accident risk. We provide empirical evidence that modeling the number of accidents as a function of the emission of air pollutants is as efficient as modeling them as a function of the distance: the models perform similarly from the precision of estimation perspective. The significant advantage of our model, however, is that we estimate road crash risk using air pollution emissions, which allows us to directly promote safer driving behavior using the argument of reducing drivers' environmental footprint.

In this paper, we contribute to academic literature two-fold. First, we take part in the discussion about the risk factors affecting road accidents. With the advent of telematics data collection, researchers have identified key risk factors affecting accident risk. The most prominent factor is the risk exposure, usually evaluated in terms of time or distance (see, for example Shen *et al.* (2020)). Speed (illustrated using different measures such as the average speed driven, sudden accelerations, and exceeding speed limits) has also been proven to increase the risk of road accidents (see, for example, Ma *et al.* (2018)). At the same time, both the speed and distance driven influence emissions (see Fafoutellis *et al.* (2020)). Against this backdrop, we provide empirical evidence that pollutant emissions are another determinant of road safety. Second, we complement road accident risk evaluation research by showing that car crashes can be modeled using pollutant emissions described in terms of particles emitted. This aligns with societal expectations of doing more to promote sustainable behaviors that help mitigate climate change. This insight also has a practical implication in the public policy space: government bodies and organizations can use our results as a motivation to promote road safety, which benefits public health.

## 2 Background

Understanding the environmental and safety impacts of land transportation is crucial in addressing contemporary challenges in public health. The research areas encompassing the prediction of pollutant emissions and fuel consumption, as well as the assessment of road accident risks, motivate and provide the foundation of our investigation. Indeed, several authors have hinted at the importance of considering these two crucial problems simultaneously. For instance, Segui-Gomez *et al.* (2011) advocated reducing the distance drivers traveled to avoid injurious accidents while also hinting at the possibility of reducing air pollution. Meanwhile, in the ambient air pollution survey provided by Anenberg *et al.* (2016), the authors promoted the development of assessment tools that address air pollution and other related health risks (including vehicle accidents). Our access to extensive telematics data allows us to bridge this gap by modeling road accident risk as a function of particle emissions.

### 2.1 Pollutant Emissions and Fuel Consumption Predictions

Over the last 30 years, air pollution caused by private vehicles has been studied using different statistical methods and datasets. Researchers have focused on two distinct objectives: predicting the amount of fuel per kilometer (petrol or diesel) or the amount of pollutant emissions per kilometer (CO2, NOx, CO, or hydrocarbons (HC)). Whichever variable is studied, be it particles or fuel, models can be more or less complex depending on the level of detail of the input data and the scale at which the measurements are made.

The literature on predicting pollutant emissions and fuel consumption is divided into two main categories: macroscopic and microscopic Nejadkoorki *et al.* (2008); Nocera *et al.* (2017). Macroscopic models estimate emissions (i.e., the mass of pollutants) or the fuel consumed in units per distance. The data is generally based on aggregated values over large areas and long periods. For instance, government agencies such as the United Kingdom's National Atmospheric Emissions Inventory (NAEI) and the European Environment Agency (EEA) provide big datasets for nationwide emissions. There are also models for estimating emissions in large-scale areas based on the mean travelling speed, such as the United States Environmental Protection Agency's Mobile Source Emissions Factor (MOBILE), the California Air Resources Board's EMission FACtor (EMFAC) and the Calculations of Emissions from Road Transport (COPERT) (see Ntziachristos *et al.* (2009)). In addition, Smit *et al.* (2008a) and Smit *et al.* (2010) list other models that incorporate the traffic situations as an input variable, for example, the Handbook Emission Factors for Road Transport (HBEFA) by De Haan & Keller (2004), the Assessment and Reliability of Transport Emission Models and Inventory Systems (ARTEMIS) by Andre *et al.* (2004) or Traffic Emissions and Energetics (TEE) by Negreti (1996).

As for microscopic models, the goal is to measure the emissions on a much smaller scale for specific vehicles, trips, or even driving cycles (acceleration, decelerations, etc.). They vary in complexity and often require a significant amount of detailed input data, requiring expertise at an engineering level to develop. Some examples include the Comprehensive Modal Emission Model (CMEM) as in Scora *et al.* (2006), the Virginia Tech Microscopic energy and emission model (VT Micro) developed in Ahn *et al.* (2002) and Rakha *et al.* (2004), or the Passenger car and the Heavy-duty Emission Model (PHEM) from the Graz University of Technology (see Hausberger *et al.* (2013)). The limitation of these models is that they are complex to implement due to requiring extensive, difficult-to-access data, particularly regarding the vehicle parameters that can generally only be acquired directly from the manufacturer. In addition to the models that use engineering-level data, scholars proposed multivariate regressions that use factors based on the vehicle class and driving patterns (speed, driving cycles, etc.), such as the Mobile Emissions Assessment System for Urban and Regional Evaluation (MEASURE) by Fomunung *et al.* (2000) or the VERSIT+ model by Smit *et al.* (2007).

More recently, the increasing availability of telematics data through GPS (Global Positioning System), OBDs, and smartphone applications made obtaining trip-specific and driver-specific information possible. Trip-specific data include factors such as the type of road and driving conditions. Driver-specific data are related to the driver's aggressiveness and speed (e.g., violating the speed limit). These data allowed researchers to identify driving behavior that impacts fuel consumption (i.e., eco-driving). A non-comprehensive list includes models that make use of smartphone data, such as Pirayre *et al.* (2022), Yao *et al.* (2020) and Meseguer *et al.* (2017) (for a general overview of current methods see Fafoutellis *et al.* (2020)). Other researchers have focused on predicting emissions directly. For instance, Rodriguez *et al.* (2023) and Thibault *et al.* (2016) studied behavioral profiles based on geographical data from smartphones, while Sentoff *et al.* (2015) and Huboyo *et al.* (2017) developed models using data directly from OBDs.

### 2.2 Road Accident Risk

Studying the factors that influence road accident risk presents several challenges, primarily due to the relative infrequency of road accidents compared to more common driving events, such as acceleration and braking. Worldwide, drivers are involved in at-fault accidents once every ten years on average. Then, several external factors, such as the weather, driver-specific behavior, or acceleration, must be considered.

Several scholars have investigated the impact of weather on accident risk. For instance, El-Basyouny *et al.* (2014) employed a multivariate Poisson model to examine how time and weather conditions influence accidents. Similarly, Naik *et al.* (2016) utilized highly detailed data from weather stations across the United States to assess its effect on road safety. In Finland, Malin *et al.* (2019) proposed a Palm probability framework to study the relationship between weather and accident risk. More recently, Becker *et al.* (2000) applied a Generalized Additive Model (GAM) to analyze the combined effects of weather and congestion on accident risk.

Road type and congestion levels are other critical external factors when analyzing accident risk. Ahn *et al.* (2004) applied a logistic distribution to examine the impact of traffic conditions on road accidents. Yu *et al.* (2013) employed data on freeway congestion levels in a multivariate model to predict accident risk. More recently, Guo *et al.* (2021) used a random forest approach to identify risk factors for road accidents and understand their relationship with traffic flow and risky driving behavior. Wang *et al.* (2013) and Retallack *et al.* (2019) provide a comprehensive literature review of this stream of research.

New research on driver-specific behavior has been made possible by the development and wider adoption of telematics. Both Ma *et al.* (2018) and Quddus (2002) found that harsh braking and rapid accelerations increase accident rates. Researchers have also conducted location-based analyses. For instance, Stipancic *et al.* (2018) demonstrated that areas with higher frequencies of acceleration and braking events are associated with an increased risk of road accidents. In actuarial science research, significant efforts have been made to develop risk profiles based on driver behavior, using advanced methodologies and telematics data. For example, Guillen *et al.* (2019) showed that specific behavioral data captured by OBDs increase the expected number of accidents (e.g., limit violations and driving in urban areas). Wüthrich (2017) designed speed-acceleration heatmaps as a tool to identify driver risk profiles. Gao & Wüthrich (2019) employed a convolutional neural network to classify telematics data, offering detailed insights into varying levels of driving risk. Similarly, using telematics data Huang & Meng (2019) applied logistic regression and machine learning techniques to telematics data to classify drivers into distinct risk categories. Across these and other studies, acceleration has been consistently identified as a critical factor in road accident analysis.

Another critical factor in assessing road accident risk is the exposure measure, which reflects how long a driver is at risk of an accident. In the literature, the most commonly used exposure measures are the distance driven or the time (i.e., observation period). From a macroscopic perspective, time is often simplified due to the lack of detailed data about individual drivers. This limitation is prevalent in many contexts; for instance, the insurance industry typically lacks access to detailed telematics data from devices like OBD systems or cell phones. As a result, the coverage period is commonly used as a proxy for exposure (see Ohlsson *et al.* (2010) for a comprehensive overview). From a microscopic perspective, distance and time have been used as exposure measures. For example, Ayuso *et al.* (2014) and Ayuso *et al.* (2016a) examined the distance driven before an accident using a Weibull distribution. Similarly, Boucher *et al.* (2017) employed Generalized Additive Models (GAM) that incorporated both distance traveled and time as exposure metrics. Recognizing the critical importance of employing either of these two exposure measures, several studies have also explored the relationship between them (see Pei *et al.* (2013), Shen *et al.* (2020), and Chipman *et al.* (1992)).

## 3 Methodology

This paper's approach to road accident modeling is driven by the accessibility of telematics data, which has demonstrated a significant impact on both accident risk and emissions research. Figure 1 provides a graphical illustration of all available data used in accident risk modeling, our choice of data, and our modeling process.

Before specifying our model, it is essential to discuss the composition of a typical telematics dataset and its influence on models developed for estimating emissions and accident risk. The factors derived from telematics data can be broadly categorized into three groups: those that significantly influence both accidents and emissions and those that are impactful specifically in only one of these contexts. A key link between emissions and road accidents is that the factors primarily driving emissions, such as vehicle characteristics, road conditions, and driving behavior, are also significant in modeling road accidents.

In modeling emissions and road accidents, it is common to include an exposure measure, such as distance traveled, alongside other factors specific to estimating emissions or road accident risk. This broadly illustrates approaches (models) presented in the existing literature, which are strongly influenced by the availability of input data.

This paper builds upon the observed positive correlation between emissions and accident risk by leveraging risk factors that are shared between the two. Our original contribution lies in replacing traditional exposure measures, such as distance, with emission estimations in order to understand accident risk as a function of emissions, which is illustrated in the green part of Figure 1.

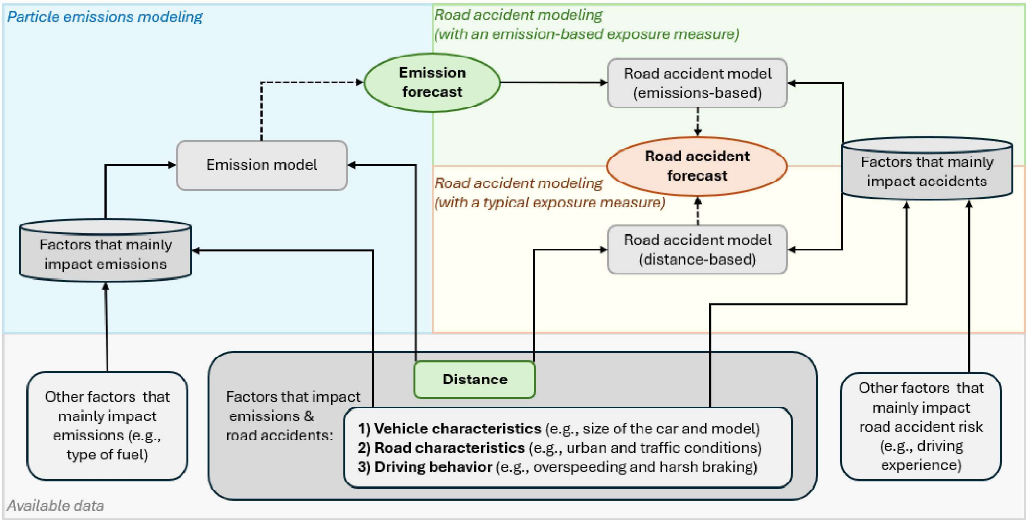


Figure 1: Emissions and road accident risk modeling.

After broadly discussing the impact of data on the modeling process, we now describe the statistical methods employed in our empirical analysis. First, we introduce the notation used for each variable in this paper. Then, we outline the distributions considered in our analysis.

### 3.1 Statistical Framework

Let us assume that $\mathcal{D}$ is a dataset containing data from $I$ drivers over $J$ years, where, for each observation $(i, j)$ (respectively identifying the driver and the year), the accident counts are recorded. In addition, suppose that for each driver, some characteristics regarding them and their vehicle are available as static covariates. Let us also assume that yearly telematics data regarding driving behavior and distance driving are recorded. Moreover, let us consider some telematics data as candidates to replace time as an exposure measure. Specifically, we select three distinct telematics-based exposure measures $k$: the distance driven ($k$=1), the the NOx ($k$=2), and the CO emissions (in grams) ($k$=3).

Based on our assumptions, for $k = 1, 2, 3$ and $i = 1, \ldots, I$ and $j = 1, \ldots, J$, let the following variables be:

- $N_{i,j}$; the number of at-fault accidents from driver $i$ at year $j$,
- $T_{i,j}^{(0)} = 1 \ \forall \ i, j$; the exposure in years for driver $i$ at year $j$,
- $T_{i,j}^{(k)}$ (for $k > 0$); the telematics-based exposure of type $k$ for driver $i$ at year $j$,
- $\mathbf{Z_{i,j}}$; the vector containing additional telematics data from driver $i$ at year $j$,

- $\mathbf{X_i}$; a vector containing non-telematics covariates from driver $i$.

Next, we identify the statistical framework of our accident frequency models, conditional to the telematics and static covariate vector. Specifically, let us consider a classical GLM structure, where the distributions are defined through a mean parameter, and three structures are considered independently of the model. For every exposure measure, they are presented in Table 1.

Table 1: Mean parameter $\left(\mu_{i,j}^{(k)}\right)$ structure for each exposure measure $k$

| Tag Number | $\mu_{i,j}^{(k)}$ formula | Description |
|---|---|---|
| (**0**) | $T_{i,j}^{(0)} \exp\left(\mathbf{X_i'}\boldsymbol{\beta}\right)$ | Benchmark model, uses time as exposure measure |
| (**I**) | $T_{i,j}^{(k)} \exp\left(\mathbf{X_i'}\boldsymbol{\beta}\right)$ | Telematics-based exposure as an offset |
| (**II**) | $\exp\left(\mathbf{X_i'}\boldsymbol{\beta} + \ln\left(T_{i,j}^{(k)}\right)\alpha^{(k)}\right)$ | Log of the telematics-based exposure as covariate |
| (**III**) | $\exp\left(\mathbf{X_i'}\boldsymbol{\beta} + \ln\left(T_{i,j}^{(k)}\right)\alpha^{(k)} + \mathbf{Z_{i,j}'}\gamma\right)$ | Model (**II**), with the other telematics covariates |

$\beta$, $\alpha^{(k)}$, and $\gamma$ are the parameter vectors that combine linearly to their respective covariates.

### 3.2 Distributions

The Poisson regression model is widely used for counting variables due to its simplicity and ability to account for exposure, either as an offset or as a covariate affecting the mean parameter ($\mu$). Consequently, we choose the Poisson model as our baseline and use it as a benchmark to evaluate more complex models.

The Negative Binomial regression is another widely used model for counting variables. Unlike the Poisson regression, it introduces an additional parameter, often denoted as ($\sigma$), which is typically set as a constant. This added complexity allows the model to overcome the equidispersion property of Poisson regression, (i.e., the expected value of the response variable is equal to its variance). The flexibility provided by the Negative Binomial model makes it particularly well-suited for overdispersed data, commonly observed in low-frequency events such as road accidents.

In modeling road accident counts, the Zero-Inflated Poisson regression offers a method to address excessive observations with no events (see Tang *et al.* (2014)). Specifically, the response variable is assumed to be a random draw from one of two events: (i) with probability $\sigma \in [0, 1]$, we assume that the observation always equals 0; (ii) with probability $1 - \sigma$, we assume that the observation is drawn from a Poisson regression. The excessive zero component is represented by a Bernoulli variable, which is equal to 0 when we draw from the Poisson model and 1 otherwise. Given that the Bernouilli regression is based on the probability of the variable to be equal to 1 as a parameter (i.e., $\sigma \in [0,1]$), incorporating an exposure measure as an offset to $\sigma$ may violate the boundaries in which the parameter is defined (e.g. obtaining values where $\sigma > 1$). To address this, exposure should instead be included as a covariate in the model, allowing the probability of drawing from the Poisson distribution to appropriately depend on the exposure.

Let us now summarize the main components of each regression using the variables defined in Section 3.1.

#### 3.2.1 *Poisson*

The Poisson distribution is defined as per Equation (1),

$$\left(N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right) \sim \mathrm{PO}\left(\mu_{i,j}^{(k)}\right), \text{ for } T_{i,j}^{(k)} > 0, \tag{1}$$

where, $\mu_{i,j}^{(k)}$ is the mean parameter and can be defined with the structures showcased in Table 1).

The mean and the variance of our Poisson model are given by:

$$\begin{aligned}
\mathrm{E}\left[N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right] &= \mu_{i,j}^{(k)}, \\
\mathrm{Var}\left[N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right] &= \mu_{i,j}^{(k)}.
\end{aligned}$$

#### 3.2.2 *Negative Binomial*

The Negative Binomial distribution is defined as per Equation (2),

$$\left(N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right) \sim \mathrm{NB}\left(\mu_{i,j}^{(k)},\sigma^{(k)}\right), \text{ for } T_{i,j}^{(k)} > 0, \tag{2}$$

where $\mu_{i,j}^{(k)}$ and $\sigma^{(k)}$ are, respectively, the mean parameter and the dispersion parameter. The former can be defined with the structures showcased in Table 1; the latter is a constant.

The mean and the variance of our Negative Binomial model are given by:

$$\begin{aligned}
\mathrm{E}\left[N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right] &= \mu_{i,j}^{(k)}, \\
\mathrm{Var}\left[N_{i,j}|\mathbf{X}_i,\mathbf{Z}_{i,j},T_{i,j}^{(k)}\right] &= \mu_{i,j}^{(k)} + \sigma^{(k)}\left(\mu_{i,j}^{(k)}\right)^2.
\end{aligned}$$

### 3.2.3 *Zero-Inflated Poisson*

Two components define the Zero-Inflated Poisson distribution. We set $N_{i,j} = 0$ with probability $\sigma_{i,j}^{(k)}$. Then, we set a Poisson distribution with probability $1 - \sigma_{i,j}^{(k)}$. The Zero-Inflated Poisson distribution is defined as per Equation (3),

$$\left(N_{i,j}|\mathbf{X}_i, \mathbf{Z}_{i,j}, T_{i,j}^{(k)}\right) \sim \text{ZIP}\left(\mu_{i,j}^{(k)}, \sigma_{i,j}^{(k)}\right), \text{ for } T_{i,j}^{(k)} > 0, \tag{3}$$

where, $\mu_{i,j}^{(k)}$ and $\sigma_{i,j}^{(k)}$ are the mean parameter of the Poisson distribution and the probability of not drawing from the Poisson distribution, respectively. The former can be defined with the structures showcased in Table 1; the latter is a function of the exposure measure, as per Equation (4),

$$\sigma_{i,j}^{(k)} = \frac{\exp\left(\psi_0^{(k)} + \ln\left(T_{i,j}^{(k)}\right)\psi_1^{(k)}\right)}{1 + \exp\left(\psi_0^{(k)} + \ln\left(T_{i,j}^{(k)}\right)\psi_1^{(k)}\right)}, \text{ for } T_{i,j}^{(k)} > 0, \tag{4}$$

where, $\psi_0^{(k)}$ is the intercept and $\psi_1^{(k)}$ is the parameter of the log of the $k^{th}$ exposure measure. Note that when years are used as an exposure measure, $T_{i,j}^{(0)} = 1$, $\forall\ i, j$. Thus the log of the exposure becomes 0, which turns $\sigma_{i,j}^{(0)}$ into a constant.

The mean and the variance of our Zero-Inflated Poisson model are given by:

$$\begin{aligned}
\text{E}\left[N_{i,j}|\mathbf{X}_i, \mathbf{Z}_{i,j}, T_{i,j}^{(k)}\right] &= \left(1 - \sigma_{i,j}^{(k)}\right)\mu_{i,j}^{(k)}, \\
\text{Var}\left[N_{i,j}|\mathbf{X}_i, \mathbf{Z}_{i,j}, T_{i,j}^{(k)}\right] &= \left(1 - \sigma_{i,j}^{(k)}\right)\mu_{i,j}^{(k)}\left(1 + 2\mu_{i,j}^{(k)}\right).
\end{aligned}$$

# 4 Results

## 4.1 Data and Descriptive Statistics

In our study, we use a data set containing information about 17,405 drivers between 2017 and 2018, obtained from a Spanish insurance company. A notable advantage of this dataset is its considerable size, which includes over 1,700 claims declared during the two-year observation period.Table 2 contains the claim frequency per driver during these two years.

Table 2: Frequency of claims per driver

| | At-fault claim counts $n = 17405$ | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Absolute Frequency | 15835 | 1450 | 112 | 6 | 2 |

In addition to the recorded claim counts, we have access to general information about each driver and vehicle as static covariates. We also have telematics data collected from the cars through onboard devices. This data includes the total kilometers driven, categorized by three distinct road types: Urban, Rural, and Motorway. Table 3 presents the descriptive statistics for the yearly distance traveled in each road type. We observe that, on average, drivers in our sample tend to drive most frequently on motorways, followed by rural roads, and least often in urban areas.

Table 3: Descriptive statistics for the distance driven in each road type (total yearly kilometers traveled by driver)

| Descriptive | Road type | | | |
|---|---|---|---|---|
| Statistics | Urban | Rural | Motorway | All |
| Mean | 2349 | 2741 | 5142 | 10233 |
| 1st Quartile | 1240 | 889 | 1442 | 5105 |
| Median | 2043 | 1924 | 3657 | 8990 |
| 2nd Quartile | 3082 | 3640 | 7307 | 13947 |
| Standard Deviation | 1663 | 2760 | 5043 | 6942 |

We also have data on the number of times each driver exceeded the speed limit and about the number of kilometers driven at night. Table 4 presents the descriptive statistics for both telematics and static covariates.

Based on these preliminary results, we note differences between the sample of drivers with and without claims. For non-telematics covariates, the riskier cohort tends to be younger on average and includes a slightly higher proportion of male drivers. Additionally, their vehicles generally have more powerful engines. As for telematics variables, drivers with claims are more likely to drive more at night, exceed speed limits, and drive more in urban areas.

Table 4: Covariates descriptive statistics

| Covariate | All sample $n = 17405$ (100%) | | Drivers without claims $n = 15835$ (91%) | | Drivers with claims $n = 1570$ (9%) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Driver Age | 29.07 | 4.62 | 29.10 | 4.60 | 28.77 | 4.78 |
| Gender Male % | 0.43 | - | 0.43 | - | 0.45 | - |
| Vehicle Power | 102.71 | 30.37 | 102.58 | 30.19 | 104.06 | 32.10 |
| Km at Night % | 10.02 | 7.80 | 9.89 | 7.73 | 11.32 | 8.38 |
| Km overspeed % | 3.60 | 6.07 | 3.58 | 6.04 | 3.84 | 6.33 |
| Km urban % | 26.39 | 15.93 | 26.21 | 15.89 | 28.23 | 16.22 |

Our data set does not contain data regarding air pollution caused by drivers. Therefore, we estimate these particles through an emission forecasting model.

### 4.2 Emission Modeling

We employ a macroscopic model to estimate emissions, utilizing emission factors provided by the National Atmospheric Emissions Inventory (NAEI), the UK agency responsible for measuring annual emissions within the country. These factors enable us to broadly estimate drivers' emissions. As previously mentioned, our data offers essential insights into road accident risk as it contains telematics variables and claim counts in a relatively large data set. While a microscopic model could yield more precise results, we lack access to sufficiently detailed telematics information about drivers' trips necessary to effectively implement current state-of-the-art microscopic approaches.

The NAIE's macroscopic model uses emission factors to estimate average vehicle emissions in the United Kingdom, accounting for road conditions similar to those in our data set. We acknowledge that the fleet of vehicles from our Spanish data set is likely different from the NAIE's. However, a broad extrapolation is possible by matching the year the NAIE collected its data (i.e., 2017). For this analysis, we focus on two key pollutants: nitrous oxides (NOx) and carbon monoxide (CO), both of which significantly impact public health. (Notably, the NAIE did not report CO2 emissions from private cars in 2017.) The emission factors from the NAIE are summarized in Table 5. Additional descriptive statistics for our emission estimates are presented in the Appendix (see Table 9). According to the NAIE, in 2017, vehicles emitted the highest levels of pollutant particles in urban areas, followed by motorways and rural regions.

Table 5: Combined hot exhaust and cold start emission factors for cars by road type (NAEI 2017)

| Pollutants | Road type | | |
|---|---|---|---|
| (in grams per km) | Urban | Rural | Motor Way |
| NOx | 0.3728 | 0.2797 | 0.3537 |
| CO | 0.7956 | 0.2606 | 0.3077 |

We run a bivariate kernel density estimation for the variable pairs (km,CO) and (km,NOx) to visualize the relationships between particle emissions and kilometers driven. The kernel density estimation for each pair is constructed using a multivariate normal distribution and a normal scale bandwidth selector, following established methods in the multivariate kernel density literature (see, for example, Chapters 2 and 3 of Chacón & Duong (2018)). Figure 2 provides a graphical representation of the results through heatmaps. We observe that the variance depends on the pollutant selected in each pair. Specifically, the pair (km,NOx) shows a strong positive correlation, with increases in one variable closely mirroring increases in the other. In contrast, while (km,CO) also exhibits a correlation, it is notably weaker, particularly for longer distances and higher emissions levels.

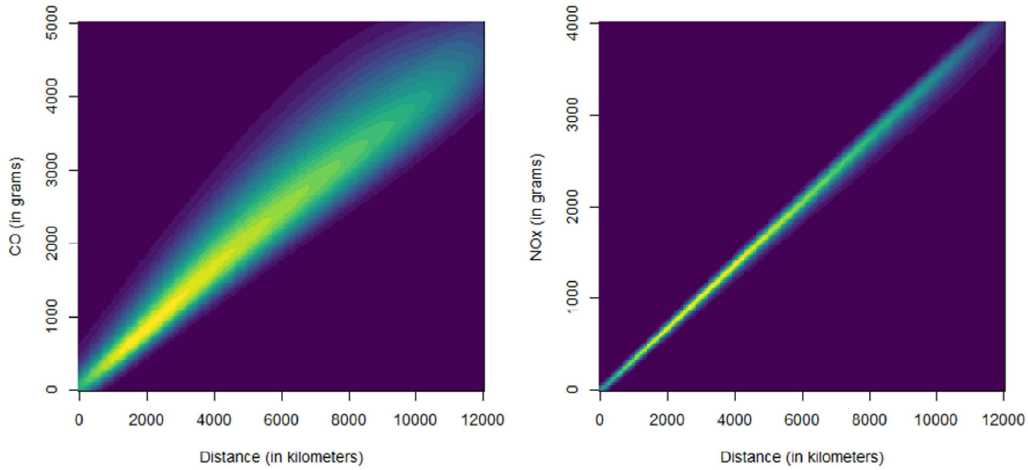


Figure 2: Bivariate kernel density distribution for kilometers and pollutant particles. On the left is the density with CO, and on the right is the density with NOx.

The macroscopic model applied in this paper yields a linear relationship between distance and particle emissions. We also observe that the choice of the particle can affect the correlation between the two. Next, we incorporate these predictions into a road accident model in Subsection 4.3.

### 4.3 Road Accident Modeling

To model road accident risk using the estimates of particle emissions, we first provide a comparative analysis of the impact of exposure measures on road accidents. Then, we introduce models used for predictive analysis, run them, and compare their predictive power and goodness-of-fit. Finally, we provide an out-of-sample simulation analysis to compare emission-based and distance-based approaches.

#### 4.3.1 *Exposure Measure Comparison*

We run a comparative analysis of our exposure measures, focusing on their impact on road accident risk. Figures 3 and 4 visually illustrate this relationship. For the analysis, we split data into two subsamples: one containing driver-year observations with no at-fault reported claims and another including driver-year observations with at least one reported claim. Using the bivariate kernel approach outlined in Section 4.2, we estimate the density distributions for each subsample.

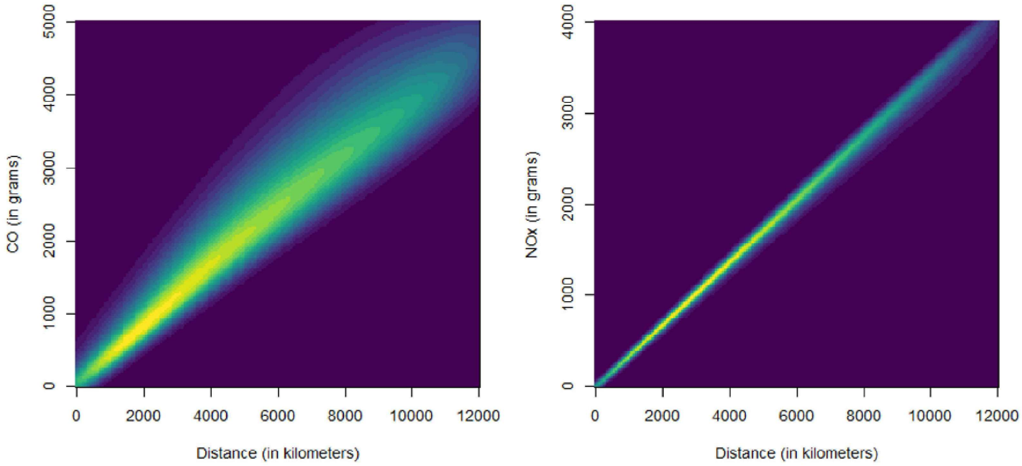


Figure 3: Bivariate kernel density distribution for kilometers and pollutant particles based on the sample without claims. On the left is the density with CO, and on the right is the density with NOx.

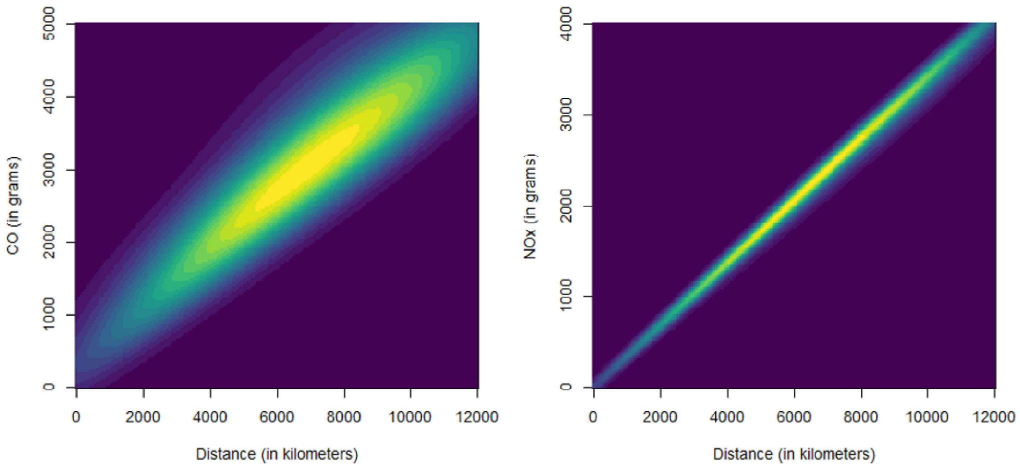


Figure 4: Bivariate kernel density distribution for kilometers and pollutant particles based on the sample with claims. On the left is the density with CO, and on the right is the density with NOx.

Figures 3 and 4 present heatmaps for the subsamples without and with claims, respectively. A key observation is that driver-year observations with at least one claim tend to exhibit higher values for kilometers driven and emissions as the mode of the distributions shifts proportionally. Additionally, consistent with the findings in Section 4.2, the choice of pollutant affects the variance of the distributions, as the (km,CO) pair's distribution showcases higher dispersion than the (km,NOx) pair. Visual inspection of the figures shows that all three exposure measures—kilometers driven (Km), CO emissions, and NOx emissions—are significant for explaining road accident frequency.

### 4.3.2 *Model Description*

After presenting a comparative analysis of the exposure measure using kernel distributions, we proceed to train models and evaluate the significance of covariates based on three statistical distributions introduced in Section 3: the Poisson distribution (Equation (1)), the Negative Binomial distribution (Equation (2)), and the Zero-Inflated Poisson distribution (Equation (3)). The corresponding mean parameters for these distributions are summarized in Table 1. For model development, we use data from 2017 as the training set and 2018 as the test set to validate the models.

From now on, we discuss models of type (**III**), and we analyze the significance of covariates using *z-tests*. Table 6 presents the results for the Negative Binomial models. Comparable results for the other two distributions, in terms of significance, are provided in the Appendix (see Tables 10- 11).

Table 6: Negative Binomial model with static covariates, telematics covariates, and exposure measures as covariates

| | Risk exposure measures | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Years | | Km | | NOx | | CO | |
| | Coef. | p-value | Coef. | p-value | Coef. | p-value | Coef. | p-value |
| ($\mu$ intercept) | -3.317 | $< 0.001$ | -5.391 | $< 0.001$ | -9.673 | $< 0.001$ | -9.844 | $< 0.001$ |
| Driver Age | -0.017 | 0.036 | -0.014 | 0.094 | -0.014 | 0.092 | -0.014 | 0.091 |
| Gender Male | -0.015 | 0.838 | -0.060 | 0.414 | -0.061 | 0.409 | -0.064 | 0.388 |
| Vehicle Power | 0.003 | 0.002 | 0.003 | 0.019 | 0.003 | 0.019 | 0.003 | 0.022 |
| Km Night | 0.018 | $< 0.001$ | 0.013 | 0.001 | 0.013 | 0.001 | 0.013 | 0.002 |
| Km overspeed | -0.002 | 0.666 | -0.008 | 0.147 | -0.009 | 0.124 | -0.008 | 0.152 |
| Km Urban | 0.010 | $< 0.001$ | 0.026 | $< 0.001$ | 0.025 | $< 0.001$ | 0.018 | $< 0.001$ |
| $\log\left(T_{i,j}^{(k)}\right)$ | - | - | 0.752 | $< 0.001$ | 0.743 | $< 0.001$ | 0.766 | $< 0.001$ |
| $\sigma$ | 1.883 | 0.021 | 3.227 | 0.080 | 3.227 | 0.080 | 3.358 | 0.086 |

Regarding the significance of static and telematics covariates, we observe a consistent pattern across all models, regardless of the exposure measure (or distribution applied). When applying a 5% p-value threshold, the driver's age and vehicle power stand out as the significant static covariates. As for the telematics covariates, the percentage of kilometers driven in urban areas and at night consistently falls below the 5% significance threshold. Two covariates—gender and the percentage of kilometers driven over the speed limit—consistently underperform in terms of their *z-test* results. The insignificance of gender is unsurprising, aligning with prior findings that it becomes non-significant when telematics variables are considered (see, for example, Ayuso *et al.* (2014)). However, the non-significance of overspeeding contradicts the consensus in the literature (see, for example, Steg & van Brussel (2009); Pawar *et al.* (2020)). We hypothesize that this effect is likely captured by other correlated components of the model, such as vehicle power and distance driven in urban areas.

Finally, the exposure measures demonstrate substantial predictive power across all models. Specifically, in the Zero-Inflated Poisson model, we observe that the exposure measures are significant for both parameters: the probability of selecting a Poisson distribution ($\sigma$) and the mean parameter of the Poisson distribution ($\mu$).

Having studied the covariates and exposure measures integrated into the models of our case study, we can now compare them using a range of statistical tools.

#### 4.3.3 *Goodness-of-fit Comparison*

We now compare the performance of the aforementioned models using the Akaike Information Criterion (AIC) and the in-sample and out-of-sample Mean Squared Error (MSE), as summarized in Table 7. The first finding is that incorporating any of the three exposure measures, either as an offset or as a covariate (models (**I**) and (**II**), respectively), significantly enhances the baseline models (**0**) across all distributions and evaluation criteria. Notably, treating the exposure measures as covariates rather than offsets reduces both AIC and MSE across all models. Adding telematics covariates leads to additional model improvements, as evidenced by the general performance gains observed when comparing models (**III**) to (**II**). These improvements are consistent across all distributions and exposure measures, reinforcing the value of incorporating telematics covariates.

Table 7: Akaike Information Criterion (AIC) and Mean Squared Error (MSE) (in-sample and out-of-sample) for selected models with various exposure measures

| | | Poisson | | | Negative Binomial | | | Zero-Inflated Poisson | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Criteria | Model | Km | NOx | CO | Km | NOx | CO | Km | NOx | CO |
| AIC | (**0**) | 6915.9 | 6915.9 | 6915.9 | 6910.8 | 6910.8 | 6910.8 | 6911.9 | 6911.9 | 6911.9 |
| | (**I**) | 6944.8 | 6935.3 | **6819.6** | 6936.6 | 6927.5 | **6816.1** | 6840.2 | 6835.5 | **6786.6** |
| | (**II**) | 6836.4 | 6832.6 | **6785.4** | 6833.0 | 6829.3 | **6783.1** | 6821.9 | 6818.6 | **6779.8** |
| | (**III**) | 6713.9 | 6714.6 | **6707.6** | 6713.4 | 6714.1 | **6707.3** | 6707.2 | 6708.1 | **<u>6704.8</u>** |
| In-S. MSE* | (**0**) | 5.0885 | 5.0885 | 5.0885 | 5.0886 | 5.0886 | 5.0886 | 5.1587 | 5.1587 | 5.1587 |
| | (**I**) | 5.1161 | 5.1137 | **5.0722** | 5.1174 | 5.1149 | **5.0724** | 5.4187 | 5.3942 | **5.1768** |
| | (**II**) | 5.0711 | 5.0702 | **5.0570** | 5.0712 | 5.0703 | **5.0571** | 5.1944 | 5.1873 | **5.1077** |
| | (**III**) | 5.0286 | 5.0286 | **<u>5.0255</u>** | 5.0286 | 5.0286 | **5.0255** | 5.1084 | 5.1053 | **5.0626** |
| Out-of-S. MSE* | (**0**) | 5.0814 | 5.0814 | 5.0814 | 5.0813 | 5.0813 | 5.0813 | 5.1599 | 5.1599 | 5.1599 |
| | (**I**) | 5.0776 | 5.0760 | **5.0479** | 5.0782 | 5.0765 | **5.0480** | 5.4282 | 5.4044 | **5.1988** |
| | (**II**) | 5.0511 | 5.0501 | **5.0389** | 5.0510 | 5.0500 | **5.0389** | 5.1905 | 5.1841 | **5.1128** |
| | (**III**) | 5.0277 | 5.0277 | **5.0263** | 5.0277 | 5.0277 | **<u>5.0263</u>** | 5.1343 | 5.1308 | **5.0888** |

*multiplied by 100.

Results in Table 7 show that when comparing exposure measures, all three indicators yield similar results. However, CO is generally the most effective indicator (results highlighted in bold font). Additionally, NOx exposure measure outperforms kilometers in models (**I**) and (**II**). Meanwhile, when all telematics covariates are included (model (**III**)), kilometers become a better yet comparable predictor than NOx. Specifically, in terms of the in-sample and out-of-sample MSE, models (**III**) with Km and NOx exposures have similar results.

The choice of distribution produces mixed results regarding the models' performance, depending on the evaluation criteria. Specifically, the Zero-Inflated Poisson distribution is preferred based on the AIC, while the Poisson distribution performs best when assessed using the in-sample MSE. Conversely, the Negative Binomial distribution performs better when evaluated with the out-of-sample MSE. These findings suggest that none of these models wins the horserace of the overall performance because the evaluation of their predictive power depends on the criterion used.

We conclude that employing specific pollutants, such as CO, as exposure measures can improve the quality of the models that estimate road accident risk, compared to using traditional exposure measures, such as the distance driven. In contrast, other pollutants, such as NOx, yield results comparable to those of traditional exposure measures. Therefore, our findings suggest that pollutant emissions can serve as effective alternatives to traditional exposure measures, such as distance-driven, in modeling road accident risk.

#### 4.3.4 *Out-of-sample Simulation Study*

In the final step, we perform an out-of-sample simulation study to evaluate the performance of different exposure measures in predicting road accidents. To do so, we run 100,000 simulations of the total claim counts in 2018, using models trained on 2017 data. Specifically, we consider two distinct frameworks: (Model (**I**)), which includes only the exposure measures (used as an offset) and excludes other telematics covariates, and (Model (**III**)), which incorporates all telematics variables, including the exposure measures, as covariates. For this analysis, we select the Negative Binomial distribution, as it yielded the best out-of-sample MSE results. Table 8 summarizes the key results, which are further illustrated for models (**I**) and (**III**) in Figures 5 and 6, respectively.

Table 8: Out-of-sample distribution of total claim counts in 2018 for Negative Binomial models

| Model Tag | Dist | $k$ | Mean | SD | Quantile 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (**I**) | NB | KM | 779.86 | 29.05 | 713 | 733 | 743 | 760 | 780 | 799 | 817 | 828 | 848 |
| | | NOx | 779.77 | 29.13 | 713 | 732 | 743 | 760 | 780 | 799 | 817 | 828 | 849 |
| | | CO | 776.46 | 29.57 | 709 | 728 | 739 | 756 | 776 | 796 | 815 | 825 | 846 |
| (**III**) | NB | KM | 791.50 | 30.62 | 721 | 741 | 752 | 771 | 791 | 812 | 831 | 842 | 863 |
| | | NOx | 791.70 | 30.74 | 721 | 741 | 752 | 771 | 792 | 812 | 831 | 843 | 863 |
| | | CO | 789.26 | 30.77 | 719 | 739 | 750 | 768 | 789 | 810 | 829 | 840 | 862 |
| Observed | | | 843 | | | | | | | | | | |

Results presented in Table 8 suggest that all models underpredict the total claim counts. At a 95% Value-at-Risk (VaR), only 1 out of 6 models yield equal or higher VaRs than the observed claim count. When comparing models of the same type (i.e., Model (**I**) or (**III**)), the choice of exposure measure has minimal impact on the total claim count distribution. The mean, standard deviation, and quantiles are nearly identical (see Table 8), resulting in similar distributions regardless of the exposure measure used (see Figures 5 and 6).

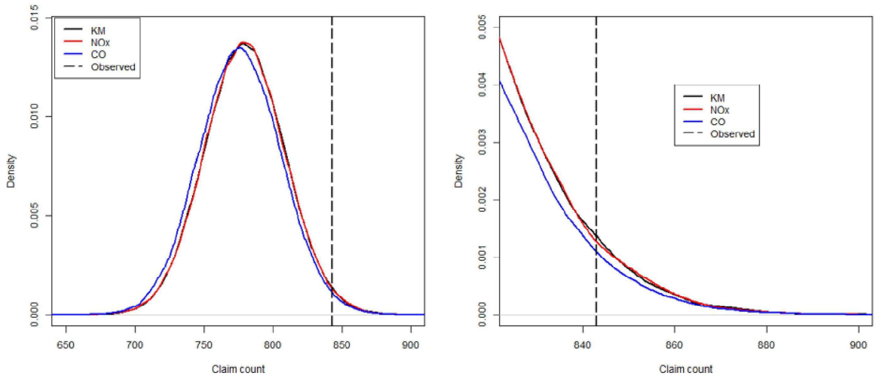


Figure 5: Negative Binomial model (**I**) distribution of the total claim count in 2018

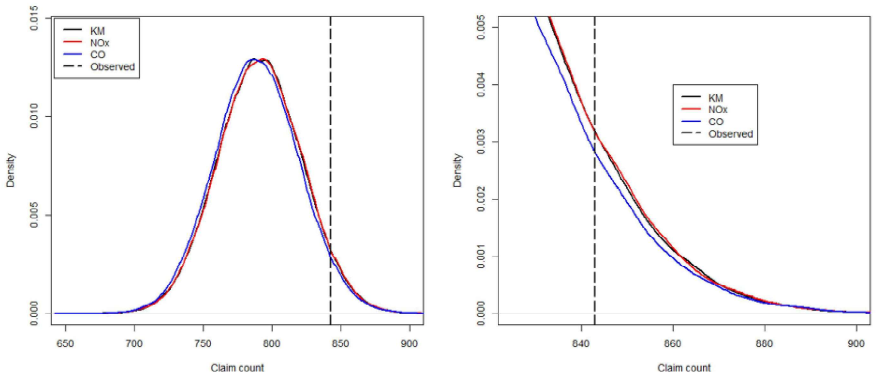


Figure 6: Negative Binomial model (**III**) distribution of the total claim count in 2018

Based on this analysis, we conclude that for large samples of driver and claim data, road accident risk can be estimated using models based on either distance-driven exposure measures or pollutant particle measures, without compromising prediction quality. Additionally, models incorporating pollution data offer a unique opportunity to increase road safety by using environmental protection arguments to influence driver behavior.

## 5 Conclusion

In this paper, we demonstrated the feasibility of understanding and evaluating road crash risk by incorporating pollutant emissions as exposure measures, replacing traditional metrics such as observation periods and distance driven. We explored various statistical frameworks and three distributions: the Poisson, Negative Binomial, and Zero-Inflated Poisson, to assess the impact of using CO and NOx emissions as alternative exposure measures. We showed that CO emissions consistently outperformed traditional exposure measures in predictive accuracy, as indicated by statistical tools such as AIC and MSE, while NOx provided comparable results to mileage.

To further illustrate our approach, we simulated the annual crash counts for 17,405 drivers, testing models with distinct covariate structures. By comparing the effect of different exposure measures on the total claim count distribution, we found that for a large number of vehicles, as in our case study, the distribution of road crash counts remains virtually indistinguishable between traditional mileage metrics and pollutant-based exposure measures.

While our macroscopic approach is well-suited for large groups of vehicles, its reliance on aggregated telematics data limits its level of detail. We propose that future research could employ a more detailed microscopic approach that uses data about specific trip characteristics or individual driver behaviors. Combining extensive datasets, as in our study, with more granular telematics data could significantly enhance the precision and applicability of this modeling framework. Such extension would allow for the development of risk and emission measures specific to individual drivers and trips, offering new insights into both accident and environmental risk.

Our study introduces a novel modeling approach to road crash risk assessment that integrates environmental considerations. By providing evidence that reducing pollutant emissions can also mitigate accident risk, we highlight a double benefit for public health and safety. This approach not only can be used to help drivers minimize their environmental footprint but also contributes to accident prevention, benefiting society as a whole.

## References


Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record, 1897*(1), 88-95.

Ahn, K., Rakha, H., Trani, A., & Van Aerde, M. (2002). Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *Journal of Transportation Engineering, 128*(2), 182-190.

Anenberg, S.C., Belova, A., Brandt, J., Fann, N., Greco, S., Guttikunda, S., Heroux, M.-E., Hurley, F., Krzyzanowski, M., Medina, S., Miller, B., Pandey, K., Roos, J. & Van Dingenen, R. (2016). Survey of ambient air pollution health risk assessment tools. *Risk analysis, 36*(9), 1718-1736.

André, M. (2004). The ARTEMIS European driving cycles for measuring car pollutant emissions. *Science of the Total Environment, 334*, 73-84.

Ayuso, M., Guillen, M. & Pérez-Marín, A. M. (2016). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies, 68*, 160-167.

Ayuso, M., Guillen, M. & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention, 73*, 125-131.

Becker, N., Rust, H. W., & Ulbrich, U. (2022). Weather impacts on various types of road crashes: a quantitative analysis using generalized additive models. *European Transport Research Review, 14*(1), 37.

Boucher, J. P., Côté, S. & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks, 5*(4), 54.

Chacón, J. E., & Duong, T. (2018). *Multivariate kernel smoothing and its applications.* Chapman and Hall/CRC.

Chipman, M. L., MacGregor, C. G., Smiley, A. M., & Lee-Gosselin, M. (1992). Time vs. distance as measures of exposure in driving surveys. *Accident Analysis & Prevention, 24*(6), 679-684.

Davison, J., Rose, R. A., Farren, N. J., Wagner, R. L., Murrells, T. P., & Carslaw, D. C. (2021). Verification of a national emission inventory and influence of on-road vehicle manufacturer-level emissions. *Environmental Science & Technology, 55*(8), 4452-4461.

De Haan, P., & Keller, M. (2004). Modelling fuel consumption and pollutant emissions based on real-world driving patterns: the HBEFA approach. *International Journal of Environment and Pollution, 22*(3), 240-258.

Duarte, G. O., Gonçalves, G. A., & Farias, T. L. (2016). Analysis of fuel consumption and pollutant emissions of regulated and alternative driving cycles based on real-world measurements. *Transportation Research Part D: Transport and Environment, 44*, 43-54.

El-Basyouny, K., Barua, S., & Islam, M. T. (2014). Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. *Accident Analysis & Prevention, 73*, 91-99.

European Commission (2024). *Annual statistical report on road safety in the EU 2024*. European Commission. https://road-safety.transport.ec.europa.eu/european-road-safety-observatory/data-and-analysis/annual-statistical-report_en

European Environmental Agency (2024). *New registrations of electric vehicles in Europe*. European Environmental Agency. https://www.eea.europa.eu/en/analysis/indicators/new-registrations-of-electric-vehicles/new-registration-of-electric-cars-eu-27

European Environmental Agency (2024). *EEA greenhouse gases — data viewer*. European Environmental Agency. https://www.eea.europa.eu/en/analysis/maps-and-charts/greenhouse-gases-viewer-data-viewers

European Environmental Agency (2021). *Air quality in Europe 2021 Europe's air quality status 2021- update*. European Environmental Agency. https://www.eea.europa.eu/publications/air-quality-in-europe-2021/air-quality-status-briefing-2021

Fafoutellis, P., Mantouka, E. G., & Vlahogianni, E. I. (2020). Eco-driving and its impacts on fuel efficiency: An overview of technologies and data-driven methods. *Sustainability, 13*(1), 226.

Fomunung, I., Washington, S., Guensler, R., & Bachman, W. (2000). Validation of the MEASURE automobile emissions model: a statistical analysis. *Journal of Transportation and Statistics, 3*, 65-84.

Gao, G. & Wüthrich, M. V. (2019). Convolutional neural network classification of telematics car driving data. *Risks, 7*(1), 6.

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk analysis, 39*(3), 662-672.

Guo, M., Zhao, X., Yao, Y., Yan, P., Su, Y., Bi, C., & Wu, D. (2021). A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. *Accident Analysis & Prevention, 160*, 106328.

Hausberger, S., Rexeis, M., & Luz, R. (2013). PHEM Passenger car and Heavy duty Emissions Model User guide for Version 11, Graz University of Technology. *Institute for Internal Combustion Engines and Thermodynamic.*

Huang, Y. & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems, 127*, 113156.

Huboyo, H. S., Handayani, W., & Samadikun, B. P. (2017, June). Potential air pollutant emission from private vehicles based on vehicle route. In *IOP Conference Series: Earth and Environmental Science* (Vol. 70, No. 1, p. 012013). IOP Publishing.

Liu, H., Chen, X., Wang, Y., & Han, S. (2013). Vehicle emission and near-road air quality modeling for shanghai, china: Based on global positioning system data from taxis and revised moves emission inventory. *Transportation Research Record, 2340*(1), 38-48.

Ma, Y. L., Zhu, X., Hu, X. & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance ratemaking. *Transportation Research Part A: Policy and Practice, 113*, 243-258.

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention, 122*, 181-188.

McDonnell, K., Sheehan, B., Murphy, F., & Guillen, M. (2024). Are electric vehicles riskier? A comparative study of driving behaviour and insurance claims for internal combustion engine, hybrid and electric vehicles. *Accident Analysis & Prevention, 207*, 107761.

Meseguer, J. E., Toh, C. K., Calafate, C. T., Cano, J. C., & Manzoni, P. (2017). Drivingstyles: A mobile platform for driving styles and fuel consumption characterization. *Journal of Communications and Networks, 19*(2), 162-168.

Naik, B., Tung, L. W., Zhao, S., & Khattak, A. J. (2016). Weather impacts on single-vehicle truck crash injury severity. *Journal of Safety Research, 58*, 57-65.

Negrenti, E. (1996). TEE: The ENEA traffic emissions and energetics model micro-scale applications. *Science of the Total Environment, 189*, 167-174.

Nejadkoorki, F., Nicholson, K., Lake, I., & Davies, T. (2008). An approach for modelling CO2 emissions from road traffic in urban areas. *Science of the Total Environment, 406*(1-2), 269-278.

Nocera, S., Basso, M., & Cavallaro, F. (2017). Micro and Macro modelling approaches for the evaluation of the carbon impacts of transportation. *Transportation Research Procedia, 24*, 146-154.

Ntziachristos, L., Gkatzoflias, D., Kouridis, C., & Samaras, Z. (2009). COPERT: a European road transport emission inventory model. In *Information Technologies in Environmental Engineering: Proceedings of the 4th International ICSC Symposium Thessaloniki, Greece, May 28-29, 2009* (pp. 491-504). Springer Berlin Heidelberg.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 174). Berlin: Springer.

Pawar, N. M., Khanuja, R. K., Choudhary, P.,& Velaga, N. R. (2020). Modelling braking behaviour and accident probability of drivers under increasing time pressure conditions. *Accident Analysis & Prevention, 136*, 105401.

Pei, X., Wong, S. C., & Sze, N. N. (2012). The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention, 48*, 464-471.

Pirayre, A., Michel, P., Rodriguez, S. S., & Chasse, A. (2022). Driving behavior identification and real-world fuel consumption estimation with crowdsensing data. *IEEE Transactions on Intelligent Transportation Systems, 23*(10), 18378-18391.

Quddus, M. A., Noland, R. B. & Chin, H. C. (2002). An analysis of motorcycle injury and vehicle damage severity using ordered probit models. *Journal of Safety Research, 33*(4), 445-462.

Rakha, H., Ahn, K., & Trani, A. (2004). Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions. *Transportation Research Part D: Transport and Environment, 9*(1), 49-74.

Retallack, A. E., & Ostendorf, B. (2019). Current understanding of the effects of congestion on traffic accidents. *International Journal of Environmental Research and Public Health, 16*(18), 3400.

Rodriguez, S. S., Alix, G., Cheimariotis, I., & Dégeilh, P. (2023). Monitoring the real-driving emissions considering the fleet and the driving behavior by geographical area. *Transportation Research Procedia, 72*, 2692-2699.

Scora, G., & Barth, M. (2006). Comprehensive modal emissions model (cmem), version 3.01. User guide. *Centre for Environmental Research and Technology. University of California, Riverside, 1070*, 79.

Sentoff, K. M., Aultman-Hall, L., & Holmén, B. A. (2015). Implications of driving style and road grade for accurate vehicle activity data and emissions estimates. *Transportation Research Part D: Transport and Environment, 35*, 175-188.

Segui-Gomez, M., Lopez-Valdes, F. J., Guillen-Grima, F., Smyth, E., Llorca, J., & de Irala, J. (2011). Exposure to traffic and risk of hospitalization due to injuries. *Risk Analysis: An International Journal, 31*(3), 466-474.

Shen, S., Benedetti, M. H., Zhao, S., Wei, L., & Zhu, M. (2020). Comparing distance and time as driving exposure measures to evaluate fatal crash risk ratios. *Accident Analysis & Prevention, 142*, 105576.

Smit, R., Ntziachristos, L., & Boulter, P. (2010). Validation of road vehicle and traffic emission models–A review and meta-analysis. *Atmospheric Environment, 44*(25), 2943-2953.

Smit, R., Brown, A. L., & Chan, Y. C. (2008a). Do air pollution emissions and fuel consumption models for roadways include the effects of congestion in the roadway traffic flow?. *Environmental Modelling & Software, 23*(10-11), 1262-1270.

Smit, R., Poelman, M., & Schrijver, J. (2008). Improved road traffic emission inventories by adding mean speed distributions. *Atmospheric Environment, 42*(5), 916-926.

Smit, R., Smokers, R., & Rabé, E. (2007). A new modelling approach for road traffic emissions: VERSIT+. *Transportation Research Part D: Transport and Environment, 12*(6), 414-422.

Steg, L., & van Brussel, A. (2009). Accidents, aberrant behaviours, and speeding of young moped riders. *Transportation Research Part F: Traffic Psychology and Behaviour, 12*(6), 503-511.

Stipancic, J., Miranda-Moreno, L. & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. *Accident Analysis & Prevention, 115*, 160-169.

Tang, Y., Xiang, L., & Zhu, Z. (2014). Risk factor selection in rate making: EM adaptive LASSO for zero-inflated Poisson regression models. *Risk Analysis, 34*(6), 1112-1127.

Thibault, L., Degeilh, P., Lepreux, O., Voise, L., Alix, G., & Corde, G. (2016). A new GPS-based method to estimate real driving emissions. *In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1628-1633). IEEE.

Wang, C., Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety Science, 57*, 264-275.

World Health Organization (2023). *Global status report on road safety 2023.* World Health Organization. `https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023`

World Health Organization (2021). *What are the WHO Air quality guidelines?*. World Health Organization. `https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines`

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal, 7*(1), 89–108.

Yao, Y., Zhao, X., Liu, C., Rong, J., Zhang, Y., Dong, Z., & Su, Y. (2020). Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones. *Journal of Advanced Transportation, 2020*(1), 9263605.

Yu, R., & Abdel-Aty, M. (2013). Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. *Accident Analysis & Prevention, 58*, 97-105.

## Appendix: Additional Tables

Table 9: Descriptive statistics for the risk exposure measures (total yearly exposure by driver)

| Descriptive | Risk exposure measure | | |
|---|---|---|---|
| Statistics | NOx* | CO* | Kilometres |
| Mean | 3462 | 4166 | 10233 |
| 1st Quartile | 1732 | 2300 | 5105 |
| Median | 3043 | 3816 | 8990 |
| 2nd Quartile | 4700 | 5606 | 13947 |
| Standard Deviation | 2350 | 2561 | 6942 |

* in grams

Table 10: Poisson model with static covariates, telematics covariates, and exposure measures as covariates

| | Risk exposure measures | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Years | | Km | | NOx | | CO | |
| | Coef. | p-value | Coef. | p-value | Coef. | p-value | Coef. | p-value |
| ($\mu$ intercept) | -3.312 | $< 0.001$ | -5.388 | $< 0.001$ | -9.666 | $< 0.001$ | -9.839 | $< 0.001$ |
| Driver Age | -0.018 | 0.031 | -0.014 | 0.089 | -0.014 | 0.087 | -0.014 | 0.086 |
| Gender Male | -0.015 | 0.839 | -0.059 | 0.415 | -0.060 | 0.411 | -0.063 | 0.389 |
| Vehicle Power | 0.003 | 0.002 | 0.003 | 0.017 | 0.003 | 0.017 | 0.003 | 0.019 |
| Km Night | 0.018 | $< 0.001$ | 0.013 | 0.001 | 0.013 | 0.001 | 0.013 | 0.002 |
| Km Overspeed | -0.002 | 0.665 | -0.008 | 0.145 | -0.009 | 0.122 | -0.008 | 0.149 |
| Km Urban | 0.010 | $< 0.001$ | 0.026 | $< 0.001$ | 0.025 | $< 0.001$ | 0.018 | $< 0.001$ |
| $\log\left(T_{i,j}^{(k)}\right)$ | - | - | 0.751 | $< 0.001$ | 0.743 | $< 0.001$ | 0.766 | $< 0.001$ |

Table 11: Zero-Inflated Poisson model with static covariates, telematics covariates, and exposure measures as covariates

| | Risk exposure measures | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Years | | Km | | NOx | | CO | |
| | Coef. | p-value | Coef. | p-value | Coef. | p-value | Coef. | p-value |
| ($\mu$ intercept) | -2.938 | $< 0.001$ | -4.159 | $< 0.001$ | -6.251 | $< 0.001$ | -6.873 | $< 0.001$ |
| Driver Age | -0.017 | 0.035 | -0.014 | 0.094 | -0.014 | 0.092 | -0.014 | 0.088 |
| Gender Male | -0.015 | 0.842 | -0.053 | 0.475 | -0.053 | 0.472 | -0.056 | 0.445 |
| Vehicle Power | 0.003 | 0.002 | 0.003 | 0.016 | 0.003 | 0.016 | 0.003 | 0.018 |
| Km Night | 0.018 | $< 0.001$ | 0.013 | 0.002 | 0.013 | 0.002 | 0.013 | 0.002 |
| Km Overspeed | -0.002 | 0.672 | -0.008 | 0.187 | -0.008 | 0.163 | -0.008 | 0.195 |
| Km Urban | 0.010 | $< 0.001$ | 0.025 | $< 0.001$ | 0.024 | $< 0.001$ | 0.018 | $< 0.001$ |
| $\log\left(T_{i,j}^{(k)}\right)$ | - | - | 0.372 | 0.010 | 0.366 | 0.011 | 0.447 | 0.005 |
| ($\sigma$ intercept) | 0.156 | $< 0.001$ | 1.676 | 0.012 | 8.528 | $< 0.001$ | 8.325 | 0.002 |
| $\log\left(T_{i,j}^{(k)}\right)$ | - | - | -1.182 | $< 0.001$ | -1.177 | $< 0.001$ | -1.139 | 0.001 |

UBIREA
Institut de Recerca en Economia Aplicada Regional i Pública
*Research Institute of Applied Economics*
**Universitat de Barcelona**
Av. Diagonal, 690 • 08034 Barcelona

**WEBSITE:** www.ub.edu/irea/ • **CONTACT:** irea@ub.edu