Direct assessment of health impacts on hospital admission from traffic intensity in Madrid

Ricardo Navares, Julio Diaz, Jose L Aznarte , Cristina Linares.

Environ Res . 2020 May;184:109254.

Noname manuscript No.
(will be inserted by the editor)

Direct assessment of health impacts on hospital admission from traffic intensity in Madrid

Ricardo Navares, Julio Díaz, José L. Aznarte,

Cristina Linares

Received: date / Accepted: date

Abstract In this paper we establish the attributable risk on respiratory and cardiovascular disorders related to traffic intensity in Madrid. In contrast to previous related studies, the proposed approach directly associates road traffic counts to patient emergency admission rates instead of using primary air pollutants. By applying Shapley values over gradient boosting machines, a first selection step is performed among all traffic observation points based on their influence on patient emergency admissions at Gregorio Marañon hospital. A subsequent quantification of the relative risk associated to traffic intensity of the selected point is calculated via ARIMA and log-linear Poisson regression models. The results obtained show that 13% of respiratory cases are related to traffic intensity while, in the case of cardiovascular disorders, the percentage increases to 39%.

Keywords traffic intensity; hospital admissions; attributable risk; ARIMA; Poisson regression.

Ricardo Navares and José Luis Aznarte
Department of Artificial Intelligence, UNED
Juan del Rosal, 16, 28040 – Madrid, Spain
e-mail: jlaznarte@dia.uned.es

Julio Díaz and Cristina Linares
Department of Epidemiology and Biostatistic.
ENS, Carlos III Institute of Health.
Avda. Monforte de Lemos, 5, 28029 – Madrid, Spain

José Luis Aznarte, Julio Díaz and Cristina Linares
Instituto Mixto de Investigación ENS-UNED (IMIENS)

1 Introduction

The impact of atmospheric pollution on hospital admissions in Madrid has been extensively researched over the last decade, both related to chemical air pollutants (Linares and Díaz 2010a,b; de Miguel-Díez et al 2019; Marques-Mejías et al 2018) and to environmental noise levels (Tobías et al 2001; Linares and Díaz 2010a; Díaz et al 2020; Carmona et al 2018). All these research studies establish the relations between levels of pollutants measured at different observation stations in Madrid, mainly NO_2, PM_{10}, $PM_{2.5}$ and noise levels, and the correspondent health indicators.

In urban areas, over 55% of particulate matter (PM) is directly related to road traffic as well as about 70% of NO_2 emissions (Quero et al 2012). With respect to environmental noise levels, the percentage associated with road traffic surpasses 70% (Recio et al 2016). Even though road traffic is the major source of pollution in big cities, there are no previous studies which relate the main cause (road traffic) with the effect (health indicators).

The main objective of this paper is to analyze the association between traffic intensity and hospital admissions. This differs from previous research proposals since it directly analyses the impact of the daily number of vehicles on hospital admissions due to respiratory and cardiovascular disorders. As a consequence of the results obtained, it is also intended to increase the comprehension of the effects of intense road traffic in major cities.

Exposure to transport-related air pollution increases the risk of premature death due to respiratory and cardiovascular causes (Krzyzanowski et al 2005; WHO Regional Office for Europe 2013; Burns et al 2020; EEA 2020; Mannucci et al 2019). Knowing the attributable risk associated with road traffic not only enables traffic control policies to local authorities, but also increases the awareness of the aftereffect to its exposure.

Among all traffic observation points surrounding hospital Gregorio Marañon in Madrid (Figure 2), the proposal selects the one which has the most impact

on the number of emergency admissions due to respiratory and cardiovascular cases recorded. Gradient boosting machines (Friedman 2001) were used to perform variable selection. Tree-based models are a popular and effective method for feature selection (Xu et al 2019), however its interpretation might vary depending on the assumptions taken over the metric used to calculate variable relative importance. Therefore the approach proposed by Lundberg and Lee (2017), which is based on Shapley values (Shapley 1953) was used to provide a more comprehensive analysis.

In order to estimate the impact of road traffic on hospital admissions two approaches were taken: autoregressive integrated moving average models (ARIMA) and log-linear Poisson regression models. Both models have been extensively used not only for forecasting the evolution of time series in a wide range of fields such as environmental atmospheric (Díaz et al 1999; Navares et al 2018) but also in studying the eventuality of epidemiological diseases (Tobías et al 2001; Linares and Díaz 2010a,b).

2 Materials and methods

2.1 Data description

2.1.1 Target Variables

Target variables consist of daily hospital admissions recorded at hospital Gregorio Marañon in Madrid due to respiratory (ICD-10: J00-J99) and cardiovascular diseases (ICD-10: I00-I99). Gregorio Marañon hospital covers the assistance of a population of 320.000 individuals which correspond to 12 primary care centers.

The study period lays between 01-01-2010 and 31-12-2012 (Figure 1) and, due to data confidentiality laws, the exact origin of the patients is not provided. Consequently, independent variables source data is collected from surrounding areas assuming emergency cases reported far away are diverted to other hospitals.

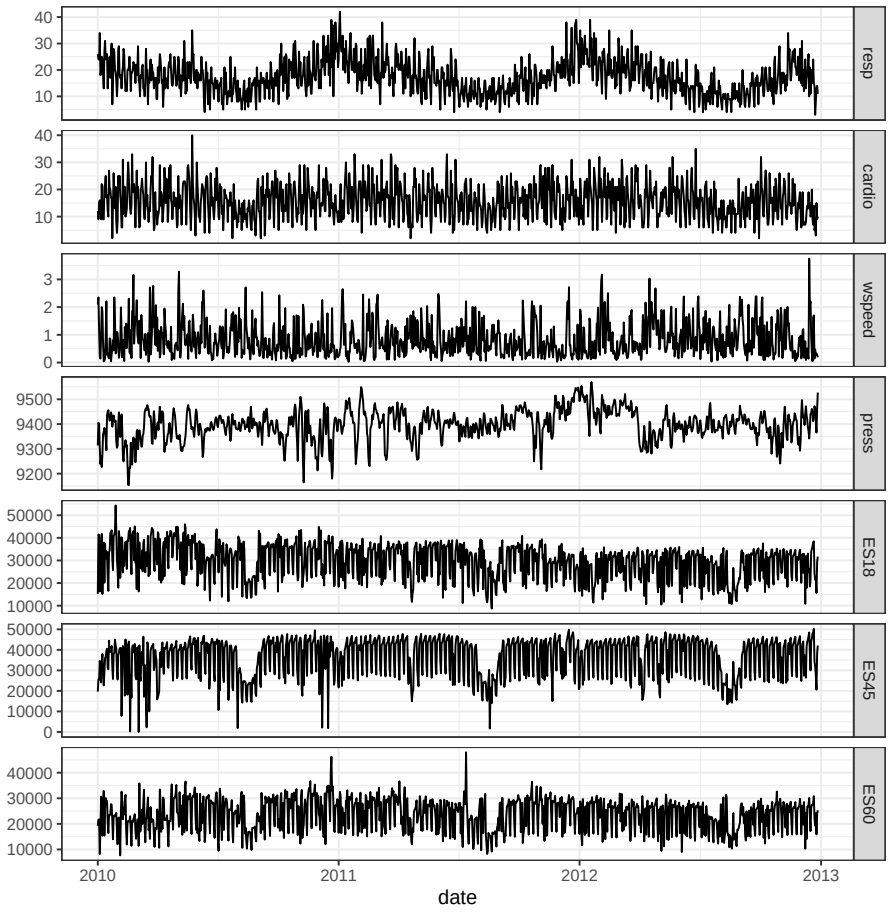

Fig. 1: Hospital admissions time series along with a sample of independent variables in this study.

2.1.2 Independent Variables.

Atmospheric conditions play an important role in the convection-diffusion process of pollutants (Li et al 2017). Consequently, pressure and wind were included to represent the convection and advection processes respectively.

Wind data. Observations consist of hourly wind speed measures in m/s and wind direction in degrees. The data is provided by the Autonomous Community of

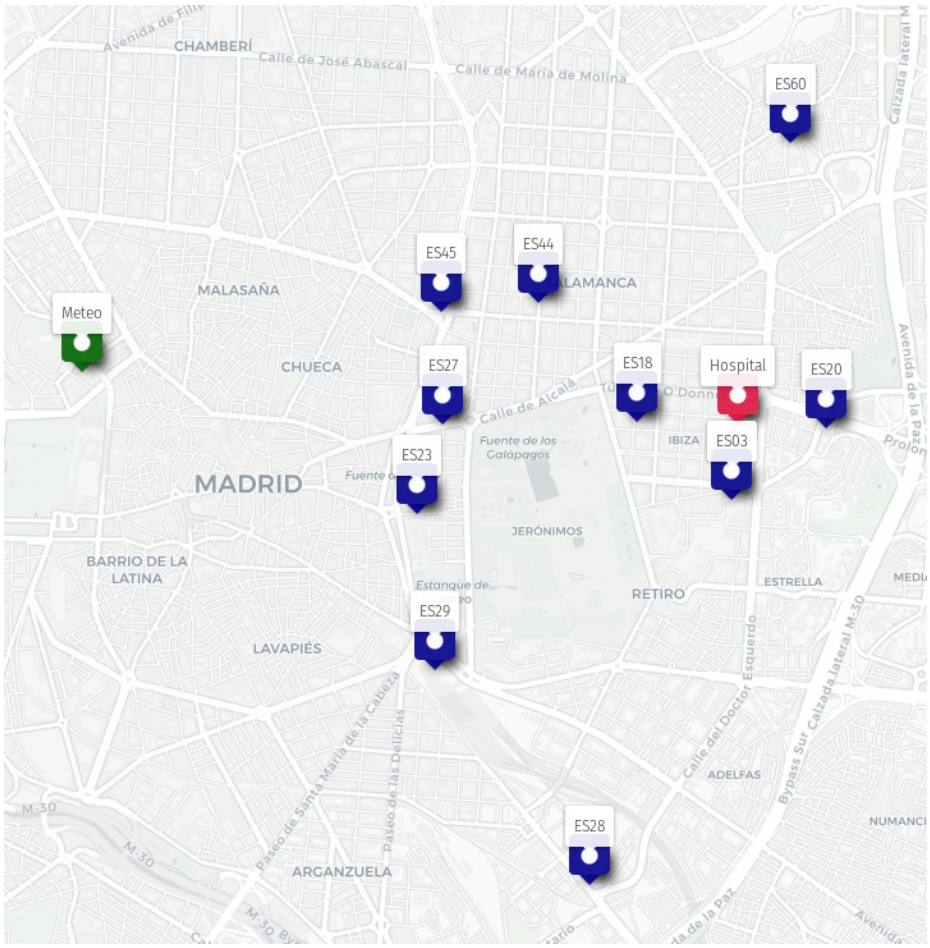

Fig. 2: Locations of the Hospital, the traffic stations and the weather station.

Madrid at Plaza de España which is the closest observation station to Gregorio Marañon hospital separated by 3 km.

Levels of immission are determined by pollutant dispersion processes in the atmosphere. Generally, these processes are driven by convection, which is pollutant dispersion to higher layers of the atmosphere, or advection which is the horizontal movement of the pollutants. Advection processes are mainly driven by wind while convection processes are driven by cyclonic (low pressure) or anticyclonic structures. Cyclonic structures are distinguished by the presence of upward currents

which ease pollutant dispersion. Conversely, downward currents are characteristic of anticyclonic structures which hinder pollutant dispersion.

Pressure. Daily average pressure was provided by the Agencia Estatal de Meteorología (AEMET) at the observation station located in Retiro which is 700 m from Gregorio Maranõn Hospital. In order to consider a synoptic scale of meteorological changes, a variable $deltaPress = P_t - P_{t-1}$ is defined being P_t the average daily pressure at time t. This variable serves to represent the trends which are related to more or less intense wind presence in the case of cyclonic ($deltaPress < 0$) or anticyclonic ($deltaPress > 0$) atmosphere respectively.

Traffic intensity. Hourly traffic intensity was provided by the Madrid Municipal Traffic Grid which consists of 4079 electromagnetic sensors placed under the pavement to detect vehicles mass that pass over the system. Instead of using the full grid, the 10 observations composing a 20 km radius which surround the hospital were selected. Records are provided as the number of cars per hour which are aggregated at each location (Figure 2) to obtain the total number of cars per day.

2.2 Methodology

The proposal consists in a first part to discriminate those variables which are not relevant, in terms of predictive capabilities, for both dependent variables estimation. With this first selection we simplify subsequent models and with their consequent gain in interpretability. As a second part, once the variables are filtered by importance, linear models are applied to analyze the independent selected variables attributable risk to the number of admissions. Firstly, a data preprocessing (Appendix A) was performed to eliminate collinearity and excessive correlations to prevent complications when applying variable selection and attribution models.

2.2.1 Variable selection

Linear models describe the relation between variables and the predictions since they have a single vector of coefficients. This interpretability eases diagnosis and it is clearly and advantage when compared to more complex computational intelligence models even though these last ones might better extract relevant information for prediction.

Nonetheless, tree-based computational intelligence models success in providing good predictive performance and interpretability. Among them, gradient boosting machines (GBM) Friedman (2001) is an ensemble technique which combines multiple weak learners to form a strong learner by additive training. At each iteration, a new weak learner (tree) is added to optimize the error function obtained by previous iteration fitted model. Even though tree based algorithms easily provide a way to extract variable importances, it is not always straightforward which assumption needs to be taken in order to obtain the importances. In tree-based models, variable importance can be interpreted as either the number of times a variable is used to split data across trees (weight) or the reduction gained in the loss function when that variable is used for splitting (gain). This decision might lead to different interpretations.

Lundberg and Lee (2017) introduce a unified approach based on Shapley values (Shapley 1953) which describes the effect of each variable on the prediction of each data point by approximating the effect of eliminating a variable from the model. The Shapley (ϕ) value of a feature x_i is defined by

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \big(f_{S \cup \{i\}} - f_S\big), \tag{1}$$

where F denotes the set of all feature space, S a subset of F and f_S is the evaluation of the algorithm given a subset S of input variables (Ichiishi 1983). Shapley values compare a prediction to a subset, which can be also composed by a single variable, instead of comparing it with the average prediction for the whole dataset allowing

more contrastive explanation when compared to local surrogate models such as the local interpretable model-agnostic explanations LIME (Ribeiro et al 2016). Lundberg and Lee (2017) provides proof of the consistency and accuracy of using Shapley values in contrast to gain and weight approaches mentioned before.

2.3 Linear Models

In order to estimate the impact of the total number of vehicles per day on hospital admissions due to respiratory and cardiovascular causes, both ARIMA models (Appendix B.1) and Log-linear Poisson regression models were used in this study. These two methodologies are comparable from the point of view of the quantification of the impact on health in normal distributions (Tobías et al 2001).

2.3.1 Poisson regression models

Poisson distributions are a particularly useful theoretical model to study the contingency of epidemiological diseases. A random variable X representing the number of occurrences of an event happens in a period of time t, follows a probability Poisson distribution if complies with the following hypothesis with respect to the cumulative incidence of the disease: proportionality, stationarity and independence. Under these assumptions, the probability of an event k during a time period t for a random variable Y that follows a Poisson distribution is defined by

$$P(Y = k) = \mathrm{e}^{-\mu} \cdot \frac{\mu^k}{k!}, \tag{2}$$

where μ represents the expected number of events during a period t (Pastor-Barriuso 2012). One of the advantages of this type of models is that they allow to determine an estimation of the effect of certain event on health, taking into account ecological studies which use aggregated data of the population. This effect is known as relative risk (RR) which is represented in this study as the number of emergency admissions due to respiratory and cardiovascular disorders.

RR represents the difference in the risk of suffering the health event between exposed and unexposed individuals due to an increase in the unit of the corresponding independent variable. The linear regression model is constructed in which the probability of a count is determined by a Poisson distribution, where the average of the distribution is a function of the external independent variables (explanatory) as follows:

$$\ln(\hat{\mu}) = \beta_0 + \beta_1 x, \tag{3}$$

where x is the explanatory variable, β_0 is the intersection and β_1 the trend. Taking exponentials of both sides, it can be derived:

$$\hat{\mu} = \mathrm{e}^{\beta_0+\beta_1 x} = \mathrm{e}^{\beta_0} \cdot \mathrm{e}^{\beta_1 x}, \tag{4}$$

where e^{β} represents the RR for each correspondent variable.

The following covariables were included in the analysis, in order to control for the trend and seasonalities of the series, as well as the lags in the Theta mentioned before:

- Sine and Cosine functions of 365, 180, 120, 90 and 60 days to account for annual, six, four, three and two month periodicities.
- The trend of the series, using a counter (n1), which is 1 for the first day of the series, 2 for the second day, and so on, successively.
- Days of the week, using dummy variable.

The p-value was determined using the step-back procedure, in which the complete model that included all the analyzed explanatory variables , those concluded relevant out of the computational intelligence method applied in the first step of the analysis, was initially implemented, with those variables that individually showed less statistical significance gradually eliminated until concluding with a model that included just the statistically significant variables ($p < 0.05$). The percentage of population attributable risk (PAR) is calculated, based on RR, as follows: $\%PAR = 100 \cdot [(RR - 1)/RR]$ (Coste and Spira 1991) representing the

percentage of increment in emergency hospital admissions associated under the hypothesis of full population exposure. We have also to assume that all the other factors that might potentially influence should remain stable. All analyses were performed using the software IBM SPSS Statistics 22 and STATA v14.1.

3 Results

3.1 Variable selection

As we have seen in Appendix A the data obtained at ES03 was highly correlated (>70%) with the majority of the other observation points. A further collinearity analysis shows that removing ES03 drives not only the VIF with respect other observation points below the threshold of 5 as suggested by O' Brien (2007) (Table A.2), but also its impact on the linear regression over the target variables is residual. Consequently, ES03 was removed from the study.

Figure 3 (a) shows the Shapley values for respiratory cases. Among all traffic observation points, ES45 shows the higher overall influence where positive Shapley values (~ 2) correspond to high influx of patients due to this disorder. Conversely, Shpley values around -2 are related to low levels of admissions. Consequently, Table A.1 shows the highest correlation for ES45. Anticyclonic atmosphere conditions (increase in pressure) show as an important factor clearly influencing high number of patient admissions due to respiratory cases.

On the other hand, atmospheric conditions seem to have less impact on the cardiovascular cases (Figure 3 (b)), staying ES45 as the most influential among traffic observation points while being also the most correlated as low / high Shapley values correspond to low / high patient influx respectively.

3.2 Impact on hospital admissions

The results obtained for emergency hospital admissions due to respiratory causes through ARIMA and the calculation of risk using linear Poisson regression models

(a) Respiratory cases , (b) Cardiovascular cases

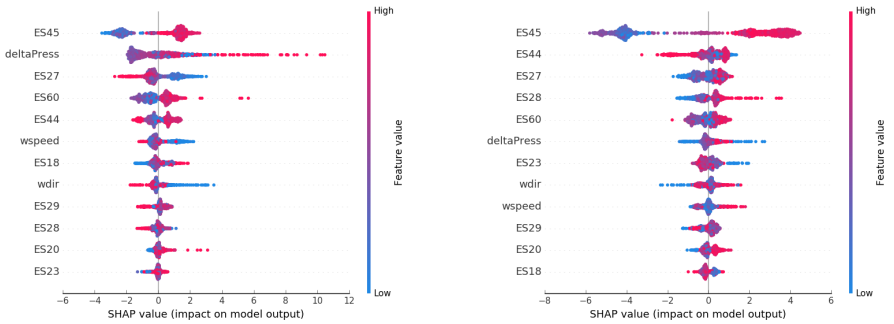

Fig. 3: GBM variable importance based on Shapley values for respiratory (a) and cardiovascular (b) cases.

Table 1: Explanatory variables obtained by the ARIMA model for the respiratory-related admissions.

		Estimations	Std. Error	t	Sig.
Non Seasonal	AR1	0.959	0.018	52493	0
	MA1	0.864	0.03	28499	0
Reg. Coefficients	Sine 365 days	2151	0.61	3528	0
	Cosine 365 days	4420	0.597	7399	0
	Sine 180 days	-1057	0.525	-2014	0.044
	Monday	4421	0.575	7684	0
	Tuesday	2841	0.592	4802	0
	Wednesday	3029	0.608	4981	0
	Thursday	1503	0.606	2481	0.013
	Friday	-1612	0.61	-2644	0.008
	Saturday	-3319	0.502	-6608	0
	LAGS(wspeed.2)	-0.428	0.228	-1875	0.061
	LAGS(deltaPress.3)	-0.007	0.004	-1854	0.064
	ES45 (*)	0.056	0.022	2474	0.014
Constant		14566	0.818	17806	0

(*) in thousands

are shown in Tables 1 and 2 respectively. Both models are consistent with respect to the resulting independent and control variables. Also, variables that control the annual and semiannual seasonality of the series, as well as the days of the week from Monday to Friday appear as significant. Regarding the independent variables, wind speed in two-days lag and the difference in atmospheric pressure in lagged three days are significant and with a negative coefficient, that is, at lower wind speed (less dispersion) and anticyclonic situations (greater atmospheric stability) they

Table 2: Poisson regression for respiratory admissions.

Poisson regression					Num. obs.: 1088	
					LR χ^2_{13}: 1511.27	
					Prob. > χ^2: 0	
Log. Lik. -3173.861					Pseudo R2: 0.1923	
	IRR	Std. Error	z	P> $\|z\|$	[95% Conf.	Interval]
LAGS(resp, 1)	1.0088	0.00141	6.25	0	1.0060	1.0116
N1 (trend)	0.9999	0.00002	-3.72	0	0.9999	1.0000
Sine 365 days	1.1174	0.01265	9.81	0	1.0929	1.1425
Sine 180 days	0.9331	0.01015	-6.36	0	0.9134	0.9532
Cosine 365 days	1.2633	0.01581	18.68	0	1.2327	1.2947
ES45 (*)	1.0045	0.00096	4.66	0	1.0026	1.0064
LAGS(wspeed.2)	0.9673	0.01153	-2.79	0.005	0.9449	0.9901
LAGS(deltaPress.3)	0.9996	0.00020	-2.05	0.041	0.9992	1.0000
Monday	1.2135	0.02779	8.45	0	1.1602	1.2692
Tuesday	1.0661	0.02570	2.65	0.008	1.0169	1.1177
Wednesday	1.0885	0.02605	3.54	0	1.0386	1.1408
Friday	0.8426	0.02186	-6.6	0	0.8008	0.8865
Saturday	0.7685	0.02063	-9.81	0	0.7291	0.8100
Constant	13.2985	0.53997	63.73	0	12.2812	14.4001

(*) in thousands

Table 3: Explanatory variables obtained by the ARIMA model for the cardiovascular admissions.

		Estimations	Std. Error	t	Sig.
Non Seasonal	AR1	0.09	0.292	0.307	0.759
	MA1	-0.015	0.293	-0.052	0.958
Reg. Coefficients	N1 (trend)	-0.001	0	-2083	0.038
	Sine 365 days	0.714	0.217	3293	0.001
	Cosine 365 days	0.59	0.209	2814	0.005
	Cosine 180 days	-0.469	0.207	-2265	0.024
	Sine 120 days	0.792	0.216	3663	0
	Cosine 120 days	-0.584	0.21	-2778	0.006
	Sine 60 days	0.441	0.209	2107	0.035
	Monday	6714	0.536	12519	0
	Tuesday	4397	0.571	7703	0
	Wednesday	3705	0.586	6322	0
	Thursday	2208	0.584	3780	0
	Friday	-1451	0.586	-2475	0.013
	Saturday	-5245	0.472	-11102	0
	ES45 (*)	0.17	0.02	8390	0
Constante		8561	0.644	13299	0

(*) in thousands

are related to an increase in the number of hospital admissions due to respiratory causes in the analyzed period.

Figure 3 (a) shows that ES45 station was the most influential among all traffic observation points for respiratory cases. As can be seen in Table 1, ARIMA mod-

Table 4: Poisson regression for cardiovascular admissions.

Poisson regression					Num. obs.: 1090	
					LR χ^2_{15}: 1746.19	
					Prob. > χ^2: 0	
Log. Lik.-3083.0027					Pseudo R2: 0.2207	
	IRR	Std. Error	z	P> $\|z\|$	[95% Conf.	Interval]
LAGS(cardio, 1)	1.0052	0.00173	3.02	0.003	1.0018	1.0086
N1 (trend)	0.9999	0.00003	-2.50	0.012	0.9999	1.0000
Sine 365 days	1.0420	0.01187	3.61	0.000	1.0190	1.0655
Cosine 365 days	1.0388	0.01149	3.45	0.001	1.0166	1.0616
Cosine 180 days	0.9721	0.01054	-2.61	0.009	0.9516	0.9930
Sine 120 days	1.0466	0.01198	3.98	0.000	1.0234	1.0704
Cosine 120 days	0.9677	0.01077	-2.95	0.003	0.9468	0.9890
Sine 60 days	1.0282	0.01125	2.54	0.011	1.0064	1.0505
Monday	1.4432	0.04798	11.03	0.000	1.3521	1.5403
Tuesday	1.2204	0.04843	5.02	0.000	1.1291	1.3191
Wednesday	1.1873	0.04573	4.46	0.000	1.1010	1.2804
Thursday	1.0921	0.04226	2.28	0.023	1.0123	1.1782
Friday	0.8662	0.03404	-3.66	0.000	0.8020	0.9355
Saturday	0.5787	0.02217	-14.27	0.000	0.5369	0.6239
ES45 (*)	1.0118	0.00128	9.27	0.000	1.0093	1.0143
Constant	9.0576	0.36176	55.17	0.000	8.3757	9.7952

(*) in thousands

eling for respiratory causes shows ES45 is significant with a positive coefficient of 0.056, meaning that for each thousand vehicles per day registered at this observation point, there is an absolute increment of 0.056 in admissions. Being the average daily number of admissions of 17.30 patients, the percentage of admissions would be 0.33% over 100 patients. Since the average daily traffic intensity at ES45 is 36.3 thousand vehicles, 12% of daily admissions due to respiratory cases are associated to the number of vehicles registered at this observation point.

Regarding the results of the Poisson modeling for the calculation of the risk, it is obtained that the ES45 station presents an increase in the relative risk, IRR = 1.004 [1.003, 1.006]; If we apply the equation to calculate the population attributable risk (PAR), we obtain that PAR = 0.40%. This result represents and increase of 0.40% patients per thousand vehicles per day. With the average of 36.3 thousand vehicles per day, it can be said that 14.5% of emergency cases due to respiratory cases are attributable to traffic intensity at ES45.

Tables 3 and 4 show the results of the ARIMA and the poisson regression models for cardiovascular cases respectively. Annual, semiannual, quarterly and

bi-monthly seasonality patterns have statistical significance as well as weekdays and Saturday. Figure 3 (b) again shows ES45 as the most influential traffic intensity observation point among all considered. ES45 station is also positive associated with cardiovascular cases according to the ARIMA (Table 3), although the coefficient is higher (0.17) when compared to the number of respiratory cases. This coefficient of 0.17 corresponds to 39.3% of daily admissions due to cardiovascular disorders which are attributable to the number of vehicles registered in ES45.

The results of the estimation of attributable impact using log-linear Poisson regression (Table 4) obtain an IRR = 1.012 [1.009, 1.014] with a PAR or 1.19%. Every thousand vehicles per day registered at ES45 increases hospital admissions due to cardiovascular cases by 1.19%. Being the daily average at ES45 of 36.3 thousand vehicles a day, the percentage of cardiovascular causes at Gregorio Marañon hospital attributable to traffic intensity is 43%.

4 Discussion

Even though the effects of pollution concentrations in the air and its influence on human health has been thoroughly studied, the relation between main pollution source in cities and related health disorders was not previously established. As we have seen in Section 3 road traffic relates to 13% of respiratory cases and 39% of cardiovascular disorders.

Variable selection through permutation over feature importance in tree-based models creates an interpretable output without the need to apply any transformation to the variable involved. However, it was mentioned the instability of the results since permutation adds randomness to the measurement, especially in the presence of highly correlated variables. In order to control these situations, a previous collinearity analysis was included along with tree-based variant of Shapley values computation proposed by Lundberg et al (2018).

Among the limitations of this proposal, there are those inherent in every longitudinal ecological study that prevent extrapolating the results at individual levels.

On the other hand, averaged data from several vehicle counting points have been used, therefore, these measures do not represent an individual exposure. However, the methodology applied is common to studies in which the impact on health is analyzed through data from air pollution measurement stations (Samet et al 2000). These biases are minimized by including in the control variable models such as trend, seasonality and autoregressive factor of the series. Finally, as in all studies that analyze the effect of pollution on health variables there is a misalignment problem (Ingebrigtsen 2015).

Even though previously cited studies show a clear relation between primary pollutant levels immission and road traffic in large cities, atmosphere also plays an important role. In this study, the role of atmosphere was included through convection-diffusion process, via the variable deltaPress, and advection via wind speed (wspeed).

The results obtained in this study are inline with respect to the role atmospheric conditions play in pollutant diffusion (Li et al 2017). Specifically, negative trends in pressure, as those obtained in the proposed models, represent a cyclonic tendency which is distinguished by the altitude of the mixing layer and, consequently, with a better dispersion of pollutants and lower levels of immission (Li et al 2017). As a result, there is a decrease in hospital admissions (Díaz et al 1999; Linares and Díaz 2010a,b).

On the other hand, the negative sign obtained in wind influence on hospital admissions, represents lower levels of immission, therefore, lower number of hospital admissions (Díaz et al 1999; Linares and Díaz 2010a,b).

From a quantitative point of view, the influence of traffic is 13% per thousand vehicles in the case of respiratory disorders, and 39.3% in the case of circulatory. These results are inline when compared to the influence of chemical air pollutants and noise levels on hospital admissions where, the impact is higher on circulatory cases than in respiratory cases (Díaz et al 1999, 2001; Tobías et al 2001; Linares and Díaz 2010a,b; Recio et al 2016). It should also be emphasized that the per-

centages obtained in this study refer to traffic intensity in contrast to pollution levels, either chemical or noise. Therefore, the results are not directly comparable to the aforementioned researches or other study focused on pollution levels. Notwithstanding, the World Health Organization (WHO, 2014) [1] establishes that air pollution-causes deaths are 40% due to ischaemic heart disease, 40% due to stroke and 11% due to chronic obstructive pulmonary disease (COPD). Similar proportions were found in this study.

5 Conclusions

In this paper, a novel approach to quantify the effect of urban traffic on respiratory and cardiovascular diseases is presented. Although pollution-related effects on health have been extensively studied, the direct influence of what is considered a main driver of pollution (traffic) was not previously established.

Indeed, our results show that traffic is responsible for emergency hospital admissions: 13% of respiratory cases are related to traffic intensity while, in the case of cardiovascular disorders, the percentage increases to 39%.

These results represent a step forward in the understanding of how human health in contemporary cities is threatened by the way in which they are organized: concretely, it becomes clear that traffic is a major cause of respiratory and cardiovascular diseases. As a consequence, raising awareness about the risks of high traffic levels should arguably be a priority for urban institutions, which should put public health in the center of how the cities are understood and managed.

Appendix A Data preprocessing

Wind is usually reported as two quantities, speed in m/s and direction in degrees 0-359 where 0 represents wind blowing from the North. Since data is provided at hourly level, in order to aggregate (via average) at daily granularity its vector components $\mathbf{u}$ and $\mathbf{v}$ which represent

[1] `https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/`

Table A.1: Correlation matrix of model variables

		correlations														
	Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1.	ES03	1														
2.	ES18	0.85	1													
3.	ES20	0.65	0.64	1												
4.	ES23	0.38	0.18	0.27	1											
5.	ES27	0.04	-0.02	0.12	0.53	1										
6.	ES28	0.76	0.84	0.59	0.12	-0.03	1									
7.	ES29	0.71	0.72	0.6	0.44	0.25	0.68	1								
8.	ES44	0.38	0.28	0.35	0.4	0.41	0.25	0.38	1							
9.	ES45	0.52	0.48	0.42	0.33	0.31	0.42	0.49	0.72	1						
10.	ES60	0.84	0.76	0.64	0.38	0.11	0.74	0.72	0.43	0.52	1					
11.	wspeed	-0.02	-0.02	-0.01	0.04	-0.04	0	0.01	-0.05	0	-0.02	1				
12.	wdir	-0.02	0	-0.05	0	-0.05	-0.01	-0.03	-0.03	-0.07	-0.02	0.07	1			
13.	pressAvg	-0.05	-0.08	0.01	0.15	0.24	-0.06	0.05	0.15	0.09	0.01	-0.24	-0.26	1		
14.	resp	0.25	0.22	0.16	0.08	-0.07	0.22	0.16	0.2	0.3	0.25	-0.05	-0.09	0.07	1	
15.	cardio	0.34	0.33	0.26	0.14	0.12	0.35	0.28	0.34	0.5	0.33	0.03	-0.03	0.03	0.42	1

Table A.2: Stepwise VIF values at each iteration and impact on linear regression statistics.

	VIF	
Variable	Step 1	Step 2
ES03	6.22	-
ES18	6.09	4.79
ES20	1.98	1.96
ES23	2.10	1.86
ES27	1.74	1.66
ES28	3.98	3.96
ES29	3.10	3.10
ES44	2.42	2.41
ES45	2.56	2.55
ES60	4.18	3.53
wspeed	1.10	1.09
wdir	1.09	1.08
pressAvg	1.37	1.36
deltaPress	1.12	1.12
	Reg. cardiovascular	
Residual SE	5.527	5.526
R2	0.27	0.27
	Reg. respiratory	
Residual SE	6.254	6.254
R2	0.13	0.14

the east-west and the north-south components respectively (Glickman 2000) and are defined by

$$\mathbf{u_i} = -u_i \sin\left(2\pi \frac{\theta_i}{360}\right), \quad \mathbf{v_i} = -u_i \cos\left(2\pi \frac{\theta_i}{360}\right), \tag{A.1}$$

where u_i is the wind speed at time i and θ_i the angle in degrees. These components can be averaged to obtain their daily levels: $\mathbf{u_t} = \frac{1}{24}\sum_{i=0}^{23} \mathbf{u_i}$ and $\mathbf{v_t} = \frac{1}{24}\sum_{i=0}^{23} \mathbf{v_i}$. Daily vector average wind speed at time t becomes $\text{wspeed}_t = (\mathbf{u_t}^2 + \mathbf{v_t}^2)^{\frac{1}{2}}$ and average wind direction is

defined by

$$\text{wdir}_t = \arctan\left(\frac{\mathbf{u_t}}{\mathbf{v_t}}\right) + C, \tag{A.2}$$

where $C = 180$ if $\left(\frac{\mathbf{u_t}}{\mathbf{v_t}}\right) < 180$ and $C = -180$ otherwise.

Collinearity or excessive correlation among variables is required to be checked to prevent complications when identifying and optimal subset of explanatory variables. High correlation and severe multicollinearity among predictors might result in instability of the coefficient estimates since confidence intervals for coefficients tend to be very wide and, as a consequence, makes models difficult to interpret as they lose statistical significance. Table A.1 shows the correlation among variables. It can be clearly seen some pairs of highly correlated variables such as ES03 and ES18 which are correlated at 85% or ES03 and ES60 with a correlation of 84% which, on the other hand it might be caused by the imputation method.

In order to examine for multicollinearity, a widely-used diagnostic called variance inflation factor (VIF) (Graham 2003) is calculated for each predictor by performing a linear regression of the predictor on the remaining other to obtain the R^2, being the VIF defined by $VIF_i = \frac{1}{1-R_i^2}$, where R_i^2 is the R^2-value obtained by regressing the i^{th} predictor on the remaining predictors. As a rule of thumb, VIF values in excess of 5 or 10 are used as an indicator of multicollinearity (Mason and Gunst 2003) although some studies warn about using a cutoff value of 10 (O' Brien 2007). Consequently, in this study the threshold to consider severe multicollinearity will be set at 5.

The stepwise procedure consists of calculating the VIF values iteratively, at each step the predictor variable with highest VIF is removed to subsequently recalculate the VIF until all predictors show a value lower than 5. Table A.2 shows the VIF at each step along with the linear regression statistics on each target variable. It can be clearly seen a very limited impact in the residual standard error and the R^2 for each respiratory and cardiovascular regressions with a reduced set of 13 variables (Table A.2) compared to the initial set of 14. this is an appendix

Appendix B Linear models

B.1 Arima

The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is, predicted value of Y equals a constant (μ) and/or a weighted sum of one or

more recent values of Y $(Y_{t-1}, \cdots, Y_{t-p})$ and a weighted sum of recent values of the errors $(e_{t-1}, \cdots, e_{t-q})$.

Lags of the stationary series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series. A nonseasonal ARIMA model is noted as an "ARIMA(p,d,q)" model, where

- p is the number of autoregressive terms,
- d is the number of nonseasonal differences needed for stationarity, and
- q is the number of lagged forecast errors in the prediction equation.

In terms of y, the general forecasting equation is:

$$\hat{y_t} = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_p e_{t-p}, \tag{B.1}$$

where ϕ_i is the coefficient of the autoregressive (AR) term i and θ_i represents the coefficient of the moving average (MA) term i. ARIMA models with exogenous variables (Makridakis et al 1983) include the values of these variables $(X \dots Z)$ with their correspondent lag $(s \dots m)$ along with the dependent variable Y, its lags (Y_{t-p}), the errors (e) and its lags (e_{t-q}) resulting in the following equation:

$$\begin{aligned} \hat{y_t} = \mu &+ \phi_1 y_{t-1} + ... + \phi_p y_{t-p} \\ &- \theta_1 e_{t-1} - ... - \theta_p e_{t-p} \\ &+ \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_s X_{t-s} + \dots \\ &+ \gamma_0 Z_t + \gamma_1 Z_{t-1} + ... + \gamma_m Z_{t-m}. \end{aligned} \tag{B.2}$$

The value of the estimators $\beta_0, \dots, \beta_s$ and $\gamma_0, \dots, \gamma_m$ of the variables that are significant at $p < 0.05$ (p-value provided by SPSS v15) indicating increased Y to increment by one unit of each independent variable $(X, \dots, Z)$ respectively. The model's goodness-of-fit was obtained by analysis of residuals (AIC, BIC, ACF, Box-Ljung).

References

Burns J, Boogaard H, Polus S, Pfadenhauer L, Rohwer A, van Erp A, Turley R, Rehfuess E (2020) Interventions to reduce ambient air pollution and their effects on health: An abridged cochrane systematic review. Environment International 135:105,400, DOI https://doi.org/10.1016/j.envint.2019.105400, URL http://www.sciencedirect.com/science/article/pii/S0160412019322056

Carmona R, Linares C, Recio A, Ortiz C, Díaz J (2018) Emergency multiple sclerosis hospital admissions attributable to chemical and acoustic pollution: Madrid (spain), 20012009. Science of The Total Environment 612:111 – 118, DOI https://doi.org/10.1016/j.scitotenv.2017.08.243, URL http://www.sciencedirect.com/science/article/pii/S0048969717322519

Coste J, Spira A (1991) Le proportion de cas attributable en santé publique: definition(s), estimation(s) et interpretation. Rev Epidemiol Santé Publique 51:399–411

Díaz J, García R, Ribera P, Alberdi JC, Hernández E, Pajares MS (1999) Modeling of air pollution and its relationship with mortality and morbidity in madrid (spain). Int Arch Occup Environ Health 75:366–376

Díaz J, Alberdi JC, Pajares MS, López R, López C, Otero A (2001) A model for forecasting emergency hospital admissions: effect of environmental variables. Journal of Environmental Health 64:9–15

Díaz J, López-Bueno J, López-Ossorio J, González J, Sánchez F, Linares C (2020) Short-term effects of traffic noise on suicides and emergency hospital admissions due to anxiety and depression in madrid (spain). Science of The Total Environment 710:136,315, DOI https://doi.org/10.1016/j.scitotenv.2019.136315, URL http://www.sciencedirect.com/science/article/pii/S0048969719363119

EEA (2020) Health impacts of air pollution. https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution

WHO Regional Office for Europe R (2013) Review of evidence on health aspects of air pollution REVIHAAP project: Technical report. Tech. rep.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29:1189–1232

Glickman TS (2000) Glossary of Meteorology. American Meteorological Society (AMS), The Council of AMS

Graham M (2003) Confronting multicollinearity in ecological multiple regression. Ecology 84:2809–2815

Ichiishi T (1983) Game Theory for Economic Analysis. Economic Theory, Econometrics and Mathematical Economics Series, Academic Press

Ingebrigtsen R (2015) Bayesian spatial modelling of non-stationary processes and misaligned data utilising markov properties for computational efficiency

Krzyzanowski M, Kuna-Dibbert B, Schneider J (2005) Health effects of transport-related air pollution. WHO Regional Office Europe

Li Z, Guo J, Ding A, Liao H, Liu J, Sun Y, Wang T, Xue H, Zhang H, Zhu B (2017) Aerosol and boundary-layer interactions and impact on air quality. National Science Review 4(6):810–

833, DOI 10.1093/nsr/nwx117, URL https://doi.org/10.1093/nsr/nwx117, http://oup.prod.sis.lan/nsr/article-pdf/4/6/810/23827203/nwx117.pdf

Linares C, Díaz J (2010a) Short-term effect of concentrations of fine particulate matter on hospital admissions due to cardiovascular and respiratory causes among the over-75 age group in madrid, spain. Public Health 124(1):28 – 36, DOI https://doi.org/10.1016/j.puhe.2009.11.007, URL http://www.sciencedirect.com/science/article/pii/S0033350609003564

Linares C, Díaz J (2010b) Short-term effect of pm2.5 on daily hospital admissions in madrid (20032005). International Journal of Environmental Health Research 20(2):129–140, DOI 10.1080/09603120903456810, URL https://doi.org/10.1080/09603120903456810, pMID: 20169485, https://doi.org/10.1080/09603120903456810

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp 4765–4774

Lundberg SM, Erion GG, Lee S (2018) Consistent individualized feature attribution for tree ensembles. CoRR abs/1802.03888, URL http://arxiv.org/abs/1802.03888, 1802.03888

Makridakis S, Wheelwright S, Mcgee V (1983) Forecasting methods and applications. Wiley, San Francisco

Mannucci PM, Harari S, Franchini M (2019) Novel evidence for a greater burden of ambient air pollution on cardiovascular disease. Haematologica 104(12):2349–2357, DOI 10.3324/haematol.2019.225086, URL http://www.haematologica.org/content/104/12/2349, http://www.haematologica.org/content/104/12/2349.full.pdf

Marques-Mejías M, Tomás-Pérez M, Hernández I, López I, Quirce S (2018) Asthma exacerbations in the pediatric emergency department at a tertiary hospital: Relationship with environmental factors. Journal of investigational allergology & clinical immunology 29, DOI 10.18176/jiaci.0364

Mason R, Gunst R (2003) Statistical Design and Analysis of Experiments: With Applications to Engineering and Science. John Wiley & Sons

de Miguel-Díez J, Hernández-Vázquez J, López-de Andrés A, Alvaro-Meca A, Hernández-Barrera V, Jiménez-García R (2019) Analysis of environmental risk factors for chronic obstructive pulmonary disease exacerbation: A case-crossover study (2004-2013). PLOS ONE 14(5):1–11, DOI 10.1371/journal.pone.0217143, URL https://doi.org/10.1371/journal.pone.0217143

Navares R, Díaz J, Linares C, Aznarte J (2018) Comparing arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in madrid. Stoch Environ Res Risk Assess pp 1–11, DOI 10.1007/s00477-018-1519-z

O' Brien R (2007) A caution regarding rules of thumb for variance inflation factors. Quality & Quantity 41:673–690, DOI 10.1007/s11135-006-9018-6

Pastor-Barriuso R (2012) Bioestadstica. Centro Nacional de Epidemiologa, Instituto de Salud-Carlos III

Quero X, Viana M, Moreno T, Alastuey A (2012) Bases científico-técnicas para un plan nacional de mejora de la calidad del aire. Informes CSIC

Recio A, Linares C, Banegas J, Daz J (2016) The short-term association of road traffic noise with cardiovascular, respiratory, and diabetes-related mortality. Environmental research 150:383–390, DOI 10.1016/j.envres.2016.06.014

Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. CoRR abs/1602.04938, URL http://arxiv.org/abs/1602.04938, 1602.04938

Samet J, Dominici F, Zeger S, Schwartz J, Dockery D (2000) The national morbidity, mortality, and air pollution study. part i: Methods and methodologic issues. Research report (Health Effects Institute) (94 Pt 1):514; discussion 7584, URL http://europepmc.org/abstract/MED/11098531

Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games II, Princeton University Press, Princeton, pp 307–317

Tobías A, Díaz J, Saez M, Carlos Alberdi J (2001) Use of poisson regression and box–jenkins models to evaluate the short-term effects of environmental noise levels on daily emergency admissions in madrid, spain. European Journal of Epidemiology 17(8):765–771

Xu ZE, Huang G, Weinberger KQ, Zheng AX (2019) Gradient boosted feature selection. CoRR abs/1901.04055, URL http://arxiv.org/abs/1901.04055, 1901.04055