UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data

Matthaios, Vasileios N.; Knibbs, Luke D.; Kramer, Louisa J.; Crilley, Leigh R.; Bloss, William J.

DOI:
10.1016/j.atmosenv.2023.120233

License:
Creative Commons: Attribution (CC BY)

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available

Citation for published version (Harvard):
Matthaios, VN, Knibbs, LD, Kramer, LJ, Crilley, LR & Bloss, WJ 2023, 'Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data', *Atmospheric Environment*. https://doi.org/10.1016/j.atmosenv.2023.120233

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

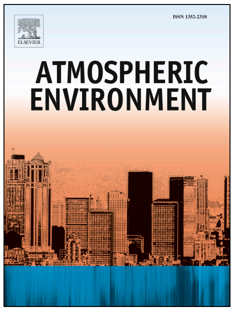
Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 30. Apr. 2024



Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data

Vasileios N. Matthaios, Luke D. Knibbs, Louisa J. Kramer, Leigh R. Crilley, William J. Bloss

PII: S1352-2310(23)00659-3

DOI: https://doi.org/10.1016/j.atmosenv.2023.120233

Reference: AEA 120233

To appear in: *Atmospheric Environment*

Received Date: 17 August 2023

Revised Date: 10 November 2023

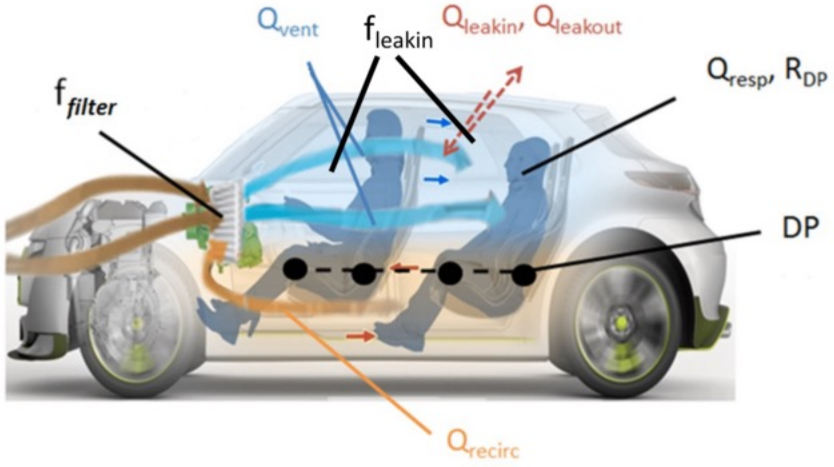
Accepted Date: 20 November 2023

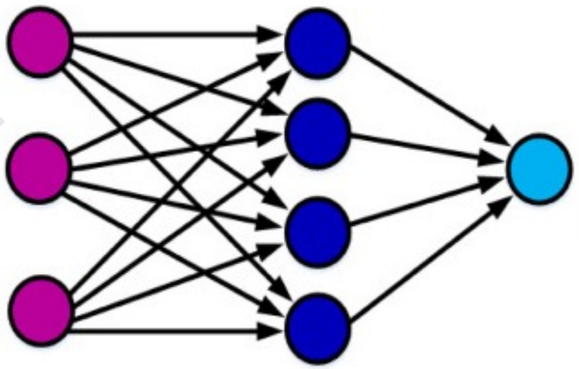
Please cite this article as: Matthaios, V.N., Knibbs, L.D., Kramer, L.J., Crilley, L.R., Bloss, W.J., Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data, *Atmospheric Environment* (2023), doi: https://doi.org/10.1016/j.atmosenv.2023.120233.

Author contributions

VNM conceived the idea, performed the analysis and wrote the first draft. LDK helped with the development of the idea. LJK and LRC helped with the experimental data collection. WJB supervised the project. VNM and WJB prepared the manuscript with contribution from all authors.

In-vehicle processes





On-road air pollution

Vehicle Characteristics

Air quality monitoring data

In-vehicle air pollution exposure

MACHINE LEARNING

Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data

Vasileios N. Matthaios[a*], Luke D. Knibbs[b,c], Louisa J. Kramer[d1], Leigh R. Crilley[d2] and William J. Bloss[d]

[a]Department of Public Health, Policy and Systems, University of Liverpool, Liverpool L69 3GB, United Kingdom

[b]School of Public Health, The University of Sydney, NSW 2006, Australia

[c]Public Health Research Analytics and Methods for Evidence, Public Health Unit, Sydney Local Health District, Camperdown, NSW 2050, Australia

[d]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[1]now at Ricardo, Harwell, Oxfordshire, OX11 0QR, United Kingdom

[2]now at WSP Australia, Brisbane, 4006 Australia

*Corresponding author: Department of Public Health, Policy and Systems, Institute of Population Health, University of Liverpool, Liverpool L69 3GB, UK

Email: V.Matthaios@liverpool.ac.uk

Modelling the air pollutant concentrations within-vehicles is an essential step to estimate our daily exposure to air pollution. This is a challenging issue however, since the processes that affect the exposures within-vehicles change with different driving patterns and ventilation settings. This study introduces an innovative approach that combines mass-balance principles and machine learning techniques, leveraging ambient air quality, on-road and within-vehicle measurements of particulate matter (PM_{10}, $PM_{2.5}$, PM_1), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), aerosol lung surface deposited area (LSDA) and ultrafine particles (UFP) under different ventilation settings to estimate air pollution exposure levels within vehicles. The first model (MB) includes basic physical and chemical processes and follows a mass-balance approach to estimate the within-vehicle concentrations. The second model (ML) applies data driven machine learning algorithms to a training set of observations to predict unseen within-vehicle concentrations. By using a number generator, the whole observational dataset was divided to 80:20 and 80% was used to build and train the ML model, while 20% was used for validation. Both models demonstrated good predictions of observations apart from an underestimation in UFP and LSDA. The ML model showed better predictive power than the MB model and had skill in predicting the unseen within-vehicle exposures. The ML model predictions were as good as the MB model for most of the species and improved for NO_2. The ML model demonstrated good index of agreement (IOA > 0.69) and Pearson correlation coefficient ($r > 0.80$) for all the species. The inclusion of air quality data from nearby monitoring stations instead of on-road (sampled while driving), in the ML model showed promising and new capabilities to within-vehicle exposure predictions. In an era where air pollution is a growing concern, understanding and predicting within-vehicle air pollution exposure is of great importance for public health and environmental research. This research not only advances the field of exposure assessment but (at no extra cost) also demonstrates practical implications for real-time exposure mapping and health impact assessment of vehicle occupants with existing infrastructure.

Keywords: within-vehicle cabin modelling, daily exposure, air pollution, machine learning, indoor air quality

Introduction

Road traffic is the dominant source of nitrogen dioxide (NO_2) and a significant contributor to particulate matter (PM_{10}, $PM_{2.5}$, PM_1 and ultrafine particles – UFP) in the atmospheres of urban environments. Numerous studies have highlighted the relationship between traffic related air pollution and adverse health effects such as cardiopulmonary disease, respiratory symptoms, reduced lung function changes in cardiac function and increased lung cancer risk (Adam et al., 2015; Hamra et al., 2015; IARC, 2014; Heal et al., 2012; Atkinson et al., 2010; De Hartog et al., 2010; Delfino et al., 2005). The road traffic dominance of many primary air pollutant emissions in urban areas leads to strong roadside concentration increments relative to urban background and rural areas (Harrison, 2018).

The interior of vehicles represents a further microenvironment where exposure to traffic related air pollution can occur, enhanced or reduced relative to the roadside environment, moderated through air exchange with the ambient environment, and within-vehicle sources, physical and chemical processing which can affect species concentrations. The significance of within-vehicle exposure varies with travel mode, environment, duration and personal commuting behaviour. In the

UK, there are approximately 32 million registered full driving license holders, of which 6% are professional drivers (DfT, 2017) who may be subject to particularly extended and elevated exposures of within-vehicle air pollution (Frederickson et al., 2020). Previous studies measuring exposure inside vehicles have found within-vehicle concentrations of $PM_{2.5}$ to be a factor of 2-3 larger than in other transport modes (e.g. De Nazelle et al., 2012; Zuurbier et al., 2010; Kumar et al., 2018), while BC and NO_2 levels inside cars can be 4.5 and 1.4 times greater than ambient concentrations (Delgado-Saborit, 2012). Other studies investigated the impact of ventilation settings on within-vehicle exposure and found that exposure was highly dependent on the air intake, vehicle age and air leaks (Kumar et al., 2021; Martin et al., 2016; Hudda et al., 2012; Knibbs et al., 2010). To inform policies, studies have also identified filtration media and usage as important factors that can help reduce within-vehicle exposures (Hachem et al., 2021; Lim et al., 2021; Matthaios et al., 2023a; Matthaios et al., 2023b). Limited studies have also directly compared pollutant levels within-vehicle with those immediately outside/adjacent to the vehicle, for both particulate and gaseous species highlighting the potentially greater health impact of NO_2 over PM exposure (Yamada et al., 2016). However, measuring within-vehicle exposure to air pollution with direct certified methods is very expensive and challenging and, given that it needs continuous monitoring, only offers a snapshot of the actual exposures. Therefore, alternative indirect approaches, such as the modelling that utilize already available air quality measurements from monitoring sites need to be explored.

Knowing that transport microenvironments represent on average 6% of our time, but account for 26% of daily total BC exposure (Dons et al., 2011); modelling the within-vehicle concentrations is an important step to assess and hence minimize personal air pollution exposure. Vehicle use changes not only from region to region but also due to meteorological conditions (e.g. more people may commute by car under cold weather). This increase in vehicle use results in more vehicle emissions not only due to the higher number of vehicles on road, but also due to the way their after-treatment abatement technologies work under cold weather (Matthaios et al., 2019). In turn these elevated vehicle emissions can result in greater exposure for vehicle occupants, depending upon ventilation and filtration media choices.

In light of the range of potential implications of improving the air quality in one of the most common microenvironments, and to provide new capabilities in real-time predicting and regulating the exposure of vehicle occupants, this study reports the development of two innovative and complementary approaches to simulate within-vehicle passenger exposure to air pollutants as a function of outside (ambient) levels and vehicle ventilation conditions. The first approach involves the development of a mass-balance (MB) model, which explicitly represents the aforementioned (predominant) physical and chemical processes which drive changes in within-vehicle air pollutant abundance. The second approach uses machine-learning algorithms (ML model), which seek to replicate the observed within-vehicle data based upon a training set of observations of internal and external (outside, ambient) pollutant concentrations, and which does not include any mechanistic representation. The results from the MB model are compared with time series measurements of within-vehicle concentrations, while the results from the ML model are compared with a subset of observations which were excluded from the training dataset. The performance of both models in estimating within-vehicle air pollution exposure is evaluated using two contrasting measures of outside (ambient) pollutant levels: (i) observations obtained directly outside the test vehicles and (ii) observations from roadside air quality monitoring stations within the same locality as the vehicle, but at some distance away from its immediate location. The objective of this study is not only to evaluate the effectiveness of this approach but to unveil its far-reaching implications for real-time exposure mapping, health impact assessment, and policy development.

2. Methods

2.1 Measurements, tested vehicles and ventilation conditions

Model development and validation was supported by measurements of NO, NO_2, O_3, PM_{10}, $PM_{2.5}$, PM_1, ultrafine particle number (UFP) and aerosol lung surface deposited area (LSDA), which were performed concurrently within vehicle cabins (in the breathing zone of the driver) and directly outside (at the side window of) the tested vehicle. CO_2 measurements were performed with two LICOR LI-820 infra-red analysers, NO_x (NO + NO_2) with chemiluminescent 42i and 42C thermo-scientific analysers, O_3 with 49i thermo-scientific analysers, PM with alphasense OPC-N2, and UFP/LSDA with DiSCmini. Temperature and relative humidity were also measured inside the vehicle cabin using HOBO sensors. Measurements were performed during two periods in 2017 in four study vehicles (see Table 1) in Birmingham (UK). Five core ventilation settings were investigated and a sixth setting was applied in two out of four vehicles: (a) front windows (of driver and co-driver) fully open, fans and AC off, (b) all windows closed; ventilation fans on (c) all windows closed; ventilation fans on with air-conditioning (AC) (d) all windows closed, ventilations fans on, recirculation mode (no AC) (e) all windows closed, ventilation fans on, recirculation mode, AC on (in two vehicles) and (f) all windows closed, ventilation system off. Fan power (air flow setting) was varied in some vehicles as outlined later. Details of the sampling campaign and quality assurance of the measurements are discussed elsewhere (Matthaios et al., 2020).

Vehicle characteristics	Ford Focus	Vauxhall Insignia	Hyundai i800	Ford Transit
Vehicle type	Estate	Estate	9 seater van	Closed cabin van
Model year	2013	2016	2017	2009
AC	Yes	Yes	Yes	No
Estimated cabin volume (m^3)	11.66	13.27	19.03	2.813
Estimated cabin geometric surface area (m^2)	34.04	37.92	47.02	14.59
Internal cabin surface:volume ratio	2.92	2.86	2.47	5.19
Air filter (as supplied)	Pollen	Pollen	Pollen	None

Table 1. Vehicles and their characteristics used in this study.

2.2 Description of within-vehicle processes and modelling

Physical air exchange processes are represented schematically in Figure 1. These give rise to an overall cabin air exchange rate from a combination of active ventilation options, passive in-built ventilation and/or leaks. The introduction of ambient pollutants may be further modified by filtering (in the case of the ventilation system). These physical processes may be described by the parameters summarised in Table 2. Considering mechanical flow alone, under recirculatory ventilation conditions, $Q_{leakin} = Q_{leakout}$ and $Q_{vent} = 0$, while under non-recirculatory ventilation settings, $Q_{vent} + Q_{leakin} = Q_{leakout}$ and $Q_{recirc} = 0$. The penetration (or removal) of air pollutants through each cabin entry mechanism can be represented by a dimensionless filtration efficiency, f, which represents the fraction of a given pollutant removed by each entry process. Deposition characterises the rate at which pollutants have losses to surfaces.

Table 2. Parameters describing the physical processes inside the vehicle cabin. Note that windows open is considered as a ventilation setting with associated values for Q_{vent} and Q_{leakin}.

Process	**Parameter**	**Nature**	**Units**	**Value Used**
Ambient air entering through ventilation system	Q_{vent}	Flow rate	$m^3\ h^{-1}$	Vehicle & ventilation setting specific
Recirculation flow through the ventilations system	Q_{recirc}	Flow rate	$m^3\ h^{-1}$	Vehicle & ventilation setting specific
Leakage: Ambient air into cabin	Q_{leakin}	Flow rate	$m^3\ h^{-1}$	Vehicle specific
Leakage: Ambient air in and out of cabin	$Q_{leakout}$	Flow rate	$m^3\ h^{-1}$	Vehicle specific
Occupant Respiration	Q_{resp}	Flow rate	$m^3\ h^{-1}$	Fixed value used for all simulations (2 occupants assumed)
Fraction of air pollutant species removed from ventilation system inflow (non-recirculatory)	f_{vent}		Dimensionless	Species specific – flow rate dependent
Fractions of air pollutants species removed during recirculation	f_{recirc}		Dimensionless	Species-specific values used, recirculation flow rate dependent
Fraction of air pollutant species removed during leak in (penetration)	f_{leakin}		Dimensionless	Species-specific values used
Fraction of pollutants lost through respiration	RD_p	Fraction of air pollutants removed during inhalation/exhalation	Dimensionless	Species-specific values used (two occupants assumed)

Losses through surface deposition	Dp_{O3} Dp_{NO} Dp_{NO2}	Species deposition rate coefficient	h^{-1}	Species-specific values used
Vehicle volume	V	volume	m^3	Vehicle specific

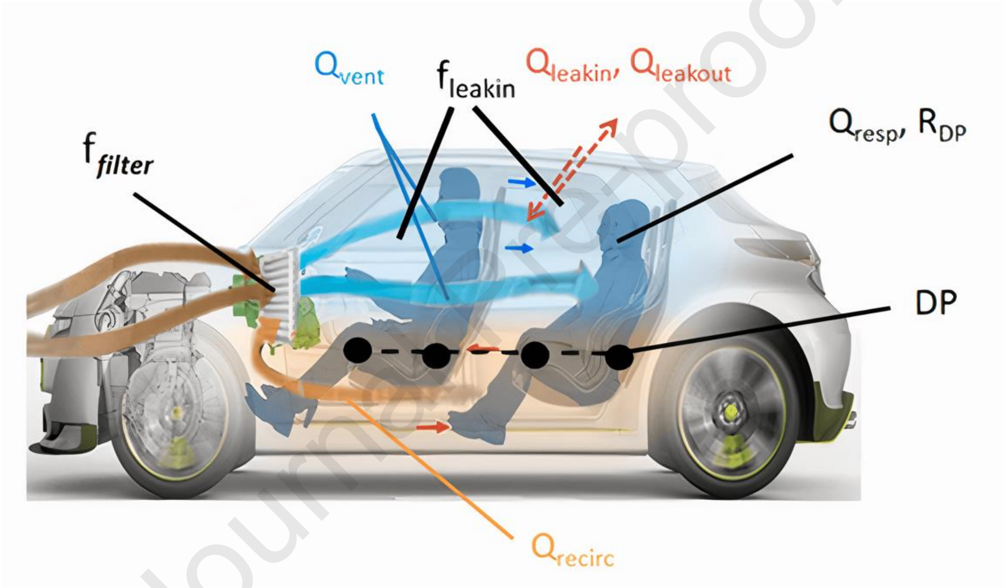


Figure 1. Schematic representation of the principal physical air exchange processes inside a typical vehicle cabin with windows closed. f_{filter}: filtration of air supply via cabin air filter; Q_{vent}: ventilation supplied flow (blue arrow). Q_{recirc}: recirculated supplied flow (orange arrow). Q_{resp}: occupant breathing rate; Dp: deposition; f_{leakin}: penetration/leaks of outside pollutants inside and vehicle leaks $Q_{Lin} and\ Q_{Lout}$: vehicle leaked flows in and out of the cabin.

2.3. Mass balance modelling approach (MB)

2.3.1 Mechanism

The mass balance (MB) model developed in this study predicts air pollutant concentrations within vehicles taking into account the physical processes illustrated in Figure 1 and a representation of the gas-phase NO_x-O_3 photostationary steady state chemistry; no other physical or chemical processes are considered here. For a given time interval, the MB model defines the rate of change of the within-vehicle air pollution concentration (following Xu and Zhu, 2009; Knibbs et al., 2010) as arising from the sum of pollutant inflow from outside (ambient) air, adjusted for filtration factors, pollutant outflow from the vehicle (both ventilation dependent), cabin surface and occupant inhalation deposition, and photochemical formation and removal (for NO_x - O_3). Air is assumed to be instantaneously homogeneously mixed within the vehicle cabin. No chemical processing of PM is considered. The mathematical equation for the MB model is given in Eq (1):

$$\frac{d(C_{inj}V)}{dt} = C_{out\,j}\left[\,Q_{vent}(1-f_{vent}) + \ Q_{Lin}f_{leakin_j}\ \ \right] - \ C_{inj}[Q_{resp}RD_p + \ Dp_jV + (Q_{vent} + Q_{Lout}) + \textstyle\sum_{j=1}^{n} R_{ij}] \qquad [1],$$

where C_{inj} is the *j* concentration inside the vehicle, $C_{out\,j}$ is the *j* concentration outside the vehicle, Q_{vent} is the mechanical supply flow, Q_L is the leakage flow (in and out as indicated by the subscript), Q_{resp} is the respiratory breathing rate of the vehicle occupants, V is the volume of the vehicle, f_{vent} is the filtration efficiency, f_{leakin_j} is the leak of pollutants that enter the cabin through cracks (penetration factor), RD_p is the deposition rate coefficient of the respiratory system of vehicle occupants, Dp is the deposition rate coefficient inside the vehicle, and R_{ij} represents the chemical / photochemical reactions (consumption and production) of species *i* and *j*. Equation (1) can be integrated numerically using a time-step approach, initial conditions and knowledge of the (time varying) outside concentrations. The different ventilation options are described in Table S1. For gases, only the NO_x-O_3 photostationary steady state reactions were included.

2.3.2 Parameters and initial conditions for Mass Balance (MB) model

Ventilation supply flow (Q_{vent}): The supply flow is calculated by multiplying the number of vents that were used with the surface area of the air vent and the air flow speed. Within the model, 4 vents with a constant size of 40 cm^2 were assumed for all vehicles. For full fan power an air flow speed of 6 $m\ s^{-1}$ was selected, while for intermediate fan power levels a value of 2.5 $m\ s^{-1}$ was applied from

Xu and Zhu, (2009). The calculated mechanical flows were 346 $m^3\ h^{-1}$, and 173 $m^3\ h^{-1}$ for full and intermediate fan power levels respectively, while for the two front fully open windows a flow of 692 $m^3\ h^{-1}$ was used (assuming two-fold amplification of the fan full power; Ott et al.,2008; Knibbs et al., 2009; Mathai et al., 2021).

Leakage flow (Q_{Lin}; Q_{Lout}): Leakage flow in and out of the vehicles is driven by the pressure difference between the interior and outdoor environment. The leakage flow depends on the ventilation settings, the vehicle characteristics, and the driving speed of the vehicle. Here leakage Q_L was based upon experiments measuring CO_2 equilibrium inside 50 vehicle cabins as reported by Hudda et al., (2012), assuming a speed of 30 *km/h* as per Eq (2):

$$\ln(Q_L) = 2.79 + (0.019\ \times S) + (0.015\ \times v.age + 3.3 \times\ 10^{-3} v.age^2) + (-0.023\ \times V +\ 6.6\ \times\ 10^{-5} V^2) + m, \qquad [2]$$

where, *S* is the vehicle speed, *V* is the volume of the cabin, *v.age* is the vehicle's age and *m* is the manufacturer adjustment (Hudda et al., 2012).

Human Respiratory inhalation flow (Q_{resp}): The Inhalation flow represents the breathing rate of the vehicle occupants. A breathing rate of 1.38 $m^3\ h^{-1}$ for males and 1.16 $m^3\ h^{-1}$ for female according to the study of Adams, (1993) was used (to match vehicle occupation during the measurements). Exhalation is a very small source for the (non-VOC) species considered here and may be neglected for most air pollutants (Knibbs et al., 2011).

Respiratory deposition coefficient (RD_p): Respiratory deposition is the net loss of particles in the human respiratory system. Here, the respiratory deposition coefficient can be considered analogous to filtration efficiency, where it represents the fractional loss of pollutant species during breathing. For UFP and LSDA (median measured value of 50 *nm*) we adopted the RD_p from Hinds, (1999) for light exercise (0.55): For PM_{10} and $PM_{2.5}$ the equivalent RD_p is 0.65 while for PM_1 it is 0.55. For NO and NO_2 a respiratory deposition coefficient of 0.67 as reported in Postlethwait and Bidani, (1990) was used.

Deposition rate coefficient (D_p): Dry deposition is a surface loss mechanism inside vehicles (Thutcher et al., 2002). Deposition rate coefficients differ between within-cabin and indoor microenvironments, as air exchange rates are much greater inside vehicles (Ott et al., 2007; Knibbs et al., 2010; Hudda et al., 2012) comparing to buildings (Yamamoto et al., 2010) if there is no indoor particle source. For UFP and LSDA (50nm size) we used the fixed deposition rate coefficient of 10 h^{-1}, as in Gong et al., (2009). This value was applied for two reasons: 1) the mean size of our UFP and LSDA for particles was 50 *nm* which is possibly due to the nature of the particles (i.e. coming from diesel exhaust) and 2) the deposition rate for particles in the range of 100 – 30 *nm* in the observational study of Gong et al., 2009 showed little variation from deposition rate spanning from 9.5 – 11.5 h^{-1}. For PM deposition values in Table 3 for different ventilation options we used the values provided by Ott et al., (2007). For NO and NO_2 we used values from Nazaroff and Cass, (1987) for indoor NO_2 decay rates in a house. Values were applied to all study vehicles.

Ventilation filtration efficiency (f_{vent}): The filtration efficiency is how well the vehicle's air filtration system removes pollutants in the incoming airflow. This filtration efficiency varies for PM_{10} and $PM_{2.5}$ depending on the experimental conditions and filter characteristics. However, since the filtration efficiency was not tested in this study, values from Qi et al., (2008), who tested vehicle particle filter efficiency in two different velocities representing low and full power fan settings, were adopted (see Table 3). Pollen filter efficacy of 0.10 based on Matthaios et al., (2023a) was applied for gases, as none of the test vehicles had activated charcoal filtration for NO_2 removal (three of the vehicles were equipped with pollen filters and one had no filter).

Fraction of species removed during leak in (penetration) f_{leakin}: f_{leak} determined the transmission efficiency for pollutants during leak entry to the vehicle. The values of f_{leak} for each particle size used in this study are summarized in Table 3 and were adjusted from indoor air quality in buildings (Chen and Zhao, 2011). It has to be noted here that no factors could be found gaseous species therefore we assumed an equivalent behaviour to fine particles ($PM_{2.5}$).

Reaction and Photolysis rates: The only reactions considered here are the (overall) photostationary steady state reactions of $NO_2 + hv \rightarrow NO + O$, $NO + O_3 \rightarrow NO_2 + O_2$ and $O + O_2 + M \rightarrow O_3 + M$. The NO + O_3 reaction rate constant was calculated using the Arrhenius expression with the measured temperatures within-cabin, and the O + O_2 recombination reaction was assumed to be instantaneous. The photolysis frequency varies based on the window design, vehicle orientation and incident sunlight (time, location). These variations can result in differences in the experienced actinic flux (Carslaw, 2007). A ratio of photolysis frequencies of 1:10 for $j(NO_2)_{indoor}$:$j(NO_2)_{outdoor}$ values reported in Carslaw, (2007) for buildings was used. The corresponding outdoor photolysis rates $j(NO_2)$ outdoor were taken from the TUV model (Madronich, 1993) for each measurement time / location, assuming clear-sky conditions.

The model was used to simulate the time-varying within-cabin pollutant concentrations for each vehicle, and each ventilation setting. This typically corresponded to a total run-time of 35 minutes, using a model timestep of 1 second. The timescale for PSS reactions is approximately 50s (under typical continental boundary layer daytime conditions), while the typical air residence time inside the vehicle can be as little as 16*s* (for an inflow of 0.192 m^3/s under windows open) and 31*s* and 63*s* (for an inflow of 0.096 m^3/s and 0.048 m^3/s under full and intermediate fan power ventilation settings respectively). In each case, outside pollutant concentrations were set to their actual (measured, time-varying) levels. The model was initiated with actual measured within-cabin pollutant concentrations.

Table 3: Parameters used for the Eq (1), (2) and (3); a) from Ott et al., 2008, b) Calculated in the study, c) Values from Gong et al., (2009) for the median UFP (50nm) size in this study d) Values from Nazaroff and Cass, (1987) for indoor NO_2 decay rates in a house e) Values from Thatcher et al., (2003), f) Values from Williams et al., (2003), g) average value from the studies reported in Chen and Zhao, (2011), h) According to light exercise and sitting from Hinds (1999) for UFP size 50nm, i) Postlethwait and Bidani, (1990) j) Values from Qi et al., (2008); +: Values used for Windows open, ++: Values used for Fan on, AC on, +++: Values used for All closed, Recirculation on; *: Full fan power, **: Low fan power; ‡: No filter efficiency was applied none of the cars was equipped with charcoal filter.

Species	Deposition rate coefficient (*Dp*)	Penetration factor (*P*)	Respiratory deposition coefficient (*RDp*)	Filter efficiency (*f.ef*)
PM_{10}	123.76^{b+}, 27.03^{b++}, 13.26^{b+++}	0.6^{e}	0.65^{h}	0.8^{j*}, 0.6^{j**}
$PM_{2.5}$	72.8^{a+},15.9^{a++},7.8^{a+++}	0.72^{f}	0.65^{h}	0.65j*, 0.45j**
PM_{1}	54.82^{b+},11.93^{b++}, 5.85^{b+++}	0.8^{g}	0.55^{h}	0.4^{j}
UFP	10^{c}	0.8^{g}	0.55^{h}	0.25^{j}
LSDA	10^{c}	0.8^{g}	0.55^{h}	0.25^{j}
NO_2	39.6^{d}	0.7	0.67^{i}	$0.1^{‡}$
NO	39.6^{d}	0.7	0.67^{i}	$0.1^{‡}$

Table 4: Parameters changed during the modelling between different vehicles. Q_s: Mechanical supplied air, Q_L: vehicle leakage, **: full fan strength; *: intermediate fan strength; $^{+}$: front windows fully open; $^{++}$ leakage at 30 kmh.

	Ford Focus	Vauxhall Insignia	Hyundai i800	Ford Transit
Q_{vent} (m^3 h^{-1})	692^{+}/346**	692^{+}/346** / 173*	692^{+}/346**	692^{+}/346** / 173*

$Q_{Lin}; Q_{Lout}$ (m^3 h^{-1})	28^{++}	27^{++}	25^{++}	39^{++}

2.4 Machine learning model (ML) and cross validation.

Machine learning (ML) algorithms learn directly from the data and can be broadly categorised into supervised or unsupervised approaches. In the former case, a known dataset is used to combine input variables in such a way as to predict the outcome using classification or regression methods. In unsupervised learning, methods such as clustering are used to recognise patterns in the data without reference to the outputs. The majority of practical machine learning uses supervised learning.

There are several supervised ML algorithms that can be used for model training and prediction. As a rule, no single learning algorithm can uniformly outperform other algorithms over all datasets. However, they can be evaluated for their (1) accuracy, (2) speed of learning, (3) speed of classification, (4) ability to deal with discrete/binary and continuous data, (5) danger of overfitting, (6) attempts required for incremental learning, (7) ability to handle model parameters and explain classifications, (8) tolerance to missing values and noise. In this study the k-Nearest Neighbour (kNN) algorithm was used. kNN is a statistical instance-based learning method used for regressions and classifications that matches which already stored instance is mostly similar to the new instance (Cover and Hart, 1975; Weinberger et al., 2006). When a new instance is inputted, the algorithm searches similar instances from memory using the distance metric (Euclidean, Manhattan, Minkowski, etc.) and then matches the new record by identifying the single most frequent label. This method is robust to noisy and large training datasets (Wettschereck et al., 1997) since it considers the query instance when deciding how to generalize beyond the training data, whereas a different machine learning method may have chosen the time where the query instance was observed (Aquilina et al., 2018). However, kNN algorithms require large storage for the model training, are sensitive to the choice of the similarity function (function which is used to compare instances) and lack of universal way to choose the best k (number of nearest neighbour) except through cross-validation (Kotsiantis, 2007).

The machine learning applied in this study used the original 80% of the within-vehicle observations of the complete dataset, selected using a random number generator. The remaining 20% was reserved to validate and test the model’s predictability and response (after the ML training) to fresh unseen data. In detail, the ML training dataset used within-vehicle concentrations as the response variable, and the training was built upon the variables of on-road concentrations, time of day, day of week ventilation power (expressed as 0, 50 and 100), ventilation type (expressed from 1 to 6), and cabin surface area and cabin volume of the vehicles. The kNN ML training and hyperparameter tuning (number of neighbours (k); distance metric; weighing of neighbours) followed the repeated grid search and k-fold cross validation approach. Mathematical description of the kNN algorithm used here can be found in supplementary information. In this method, after randomly splitting the training data into k-folds (10 in this case), a ML model was trained for k-1 folds (training fold) of the dataset and tested on the k^{th} (testing fold). For each fold/subset that was held out, the model was trained on all other subsets. This training process was repeated 1000 times and the final model accuracy was taken as the average of those repeats. More repetitions provide better accuracy for each instance in the dataset, however it should be mentioned that this requires more computational power. This process maximizes the training and the testing of the ML algorithm and has the advantage that for a single dataset all the available values are used for training and testing. This method is robust for estimating the accuracy of the model and the size of k and tunes the amount of bias in the predictions; Principles which are critical when using a kNN approach (Kotsiantis, 2007). Finally, the ML model (built from k-1 folds and tested on the k^{th} fold with 1000 repeats) was evaluated against the 20% of the initially randomly excluded data to assess its performance. A comparison of the three machine learning algorithms tested are listed in Table S2.

2.5 Model evaluation and real-world application scenarios

To evaluate / validate the MB and ML models we used the statistical indices of: 1) Root mean square error (RMSE) between the predicted and observed pollutant concentrations, where the closer the RMSE is to 0 the better the model prediction (Aidaoui et al., 2015; Matthaios et al., 2017); 2) fraction of predictions within a factor of 2 of observations (FAC2), where the predictions vary between $0.5 \leq FAC2 \leq 2$ and $FAC2 = 1$ is the perfect prediction; 3) mean bias (MB), which is the relative mean over or under estimation of the model predictions; 4) Mean Gross Error (MGE), which provides an indication of the mean error of the model regardless of whether it is an over or under estimate; 5) Pearson correlation coefficient (r), which represents the strength of the linear relationship between two variables; 6) Index of agreement (IOA) which is a measure of how well the predicted variations are represented around the mean observations and ranges from 0 to 1, and 7) comparison of means (for observed and predicted values). The model evaluation statistics were performed with openair package in R (Carslaw, 2019; Carslaw and Ropkins, 2012)

To examine the predictability of the MB model and the applicability of the ML model, we tested two further cases: (i) in the MB model we replaced the initial within-vehicle concentrations with the median observed within-vehicle concentration (for each ventilation setting in each car) and we re-ran the MB model to ensure that there was minimal dependence upon the model initial conditions. In case (ii), the ML model was retrained with initial concentrations set to the median within-vehicle levels, and with the outside levels taken from the closest roadside air quality station, rather than using the actual on-road measurements measured adjacent to the vehicle. This case was built to reflect a potential real-world situation *i.e.* where only monitoring station data is likely to be available. Again, the ML model followed the 80:20 approach with 1000 iterations. Table 5 summarises the constructed cases.

Table 5. Modelling cases constructed to test the application of the model. C'_{inmj}: denotes predicted median concentration; C'_{inmj}: denotes within-vehicle median levels. All the remaining parameters in the model are taken from the values in Table 3.

Case	Equation
Initial model	$(C'_{inj} - C_{inj})V = \left[C_{outj}\left(Q_{vent}(1 - f_{vent}) + Q_{Leakin}f_{leakin_j}\right) - C_{inj}\left(Q_{resp}RDp_j + DP_jV + (Q_{vent} + Q_{Leakout})\right) + \sum_{j=1}^{n} R_{ij}\right]\Delta t$
Case (i)	$(C'_{inmj} - C_{inmj})V = \left[C_{outj}\left(Q_{vent}(1 - f_{vent}) + Q_{Leakin}f_{leakin_j}\right) - C_{inmj}\left(Q_{resp}RDp_j + DP_jV + (Q_{vent} + Q_{Leakout})\right) + \sum_{j=1}^{n} R_{ij}\right]\Delta t$

3. Results

3.1 Measured concentrations. The measurements of ventilation-setting-dependent within-vehicle concentrations is discussed briefly in section 2.1 and in detail in Matthaios et al., (2020). Here, Table 6 presents the median of the concentrations measured. As anticipated, the highest exposure to exhaust-related gaseous (NO_2 and NO_x) and particulate (UFP and LSDA) pollutants was measured with open windows (ventilation option a). Under closed windows, the highest median exposure to particulate pollution (PM_{10}, $PM_{2.5}$, PM_1) was measured when the fan was on bringing air from outside inside (ventilation option b). The lowest mean exposure for PM_{10}, $PM_{2.5}$, PM_1, UFP and LSDA occurs when ventilation recirculation option is selected (ventilation options d and e). The within-vehicle measurements show a strong dependence upon ventilation setting, highlighting the importance of ventilation representation for accurate within-vehicle pollutant prediction.

Table 6. Median within-vehicle concentrations of PM_{10}, $PM_{2.5}$, PM_1, LSDA, NO_2, NO_x, UFP and CO_2 under ventilation settings: (a) windows open, fans and AC off, (b) Fans on - AC & recirculation off, windows closed, (c) Fan plus AC on, recirculation off, windows closed (d), Fan plus recirculation on, AC off, windows closed, (e) Fan plus AC and recirculation on, windows closed and (f) windows closed, AC, fans and recirculation off.

Species	**Ventilation (a)**	**Ventilation (b)**	**Ventilation (c)**	**Ventilation (d)**	**Ventilation (e)**	**Ventilation (f)**
PM_{10} ($\mu g/m^3$)	15	24	6	8	3	13
$PM_{2.5}$ ($\mu g/m^3$)	8	15	4	4	3	5
PM_1 ($\mu g/m^3$)	5	13	3	3	2	3
LSDA ($\mu m^2/cm^3$)	52	39	38	12	6	26
NO_2 (*ppb*)	53	48	40	48	32	31
NO (*ppb*)	232	210	209	227	245	125
UFP (*pt/cm³*)	44816	31960	27265	5466	400	19110
O_3 (*ppb*)	8.6	4.1	4.4	2	2.2	5

3.2 Modelling results – Comparison with observations

3.2.1 Mass-Balance model simulations

Figure 3 compares the timeseries of mass-balance (MB) model predictions and measured levels of (within-vehicle) UFP and NO_2 from one of the test vehicles. For UFP, the model performs well under windows-open, fan-on and AC-on modes, but overpredicts the observed levels under the no-ventilation and recirculation modes. For NO_2, the MB model performs well under no-ventilation and recirculation conditions but underestimates the observations for windows-open and AC-on, and overestimates for fan-on and AC-with-recirculation.

To examine the performance of the MB model across all the measurements, the data are aggregated in Figure 4, which shows the measured vs. MB model values for all measurements. Individual ventilation setting predictions can be found in supplementary information Figures S2 – S7. PM_{10}, $PM_{2.5}$ and PM_1 species are predicted well by the model and are within the ±10% of the 1:1 line, however, a clear under estimation is evident for UFP and LSDA. This is possibly because the model parameter values for filtration efficiency, deposition rate coefficient and penetration factors were taken from the literature, rather than reflecting the specific vehicle under evaluation. Furthermore, internal sources of particle generation were not considered, which could contribute to the under-prediction in those species. For NO we see some overpredictions at mid to high mixing ratios (>250 *ppb*), however in general the majority of the predictions are well within ±10% of the measured data. For NO_2 the predictions vs observations are clearly more scattered than for the other pollutants, and the model predicts well the low levels <60 *ppb* clearly underpredicts levels from 75 – 150 *ppb*.

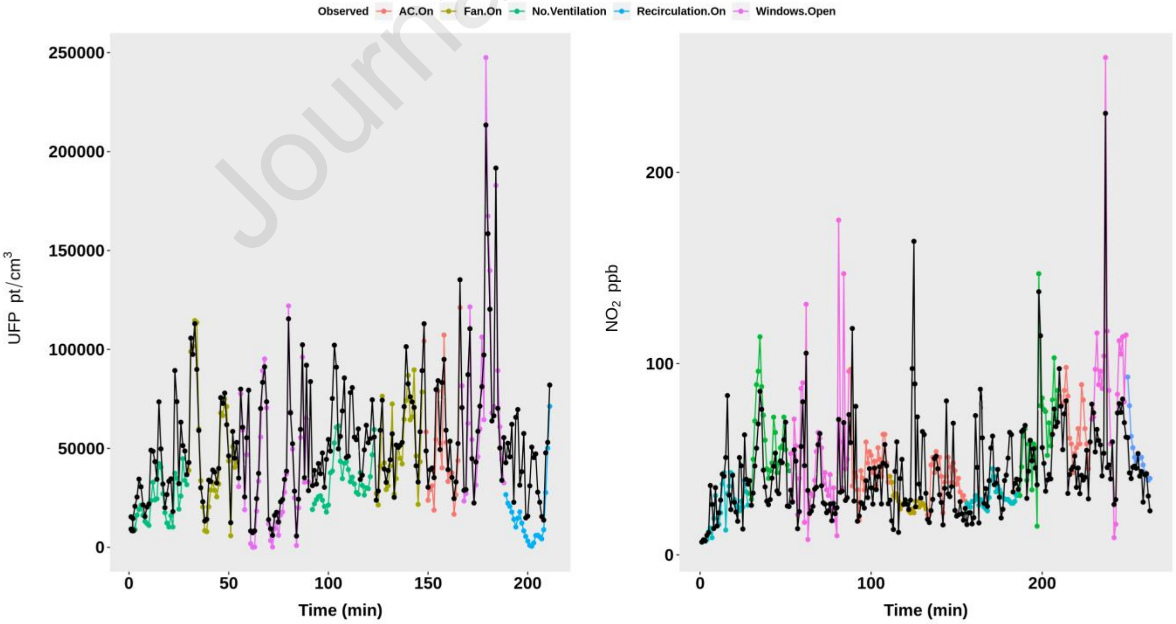


Figure 3. Time series modelled and observed values for UFP and NO_2 in Vauxhall Insignia. Different colours indicate the different ventilations, while the solid black line shows the modelled data.

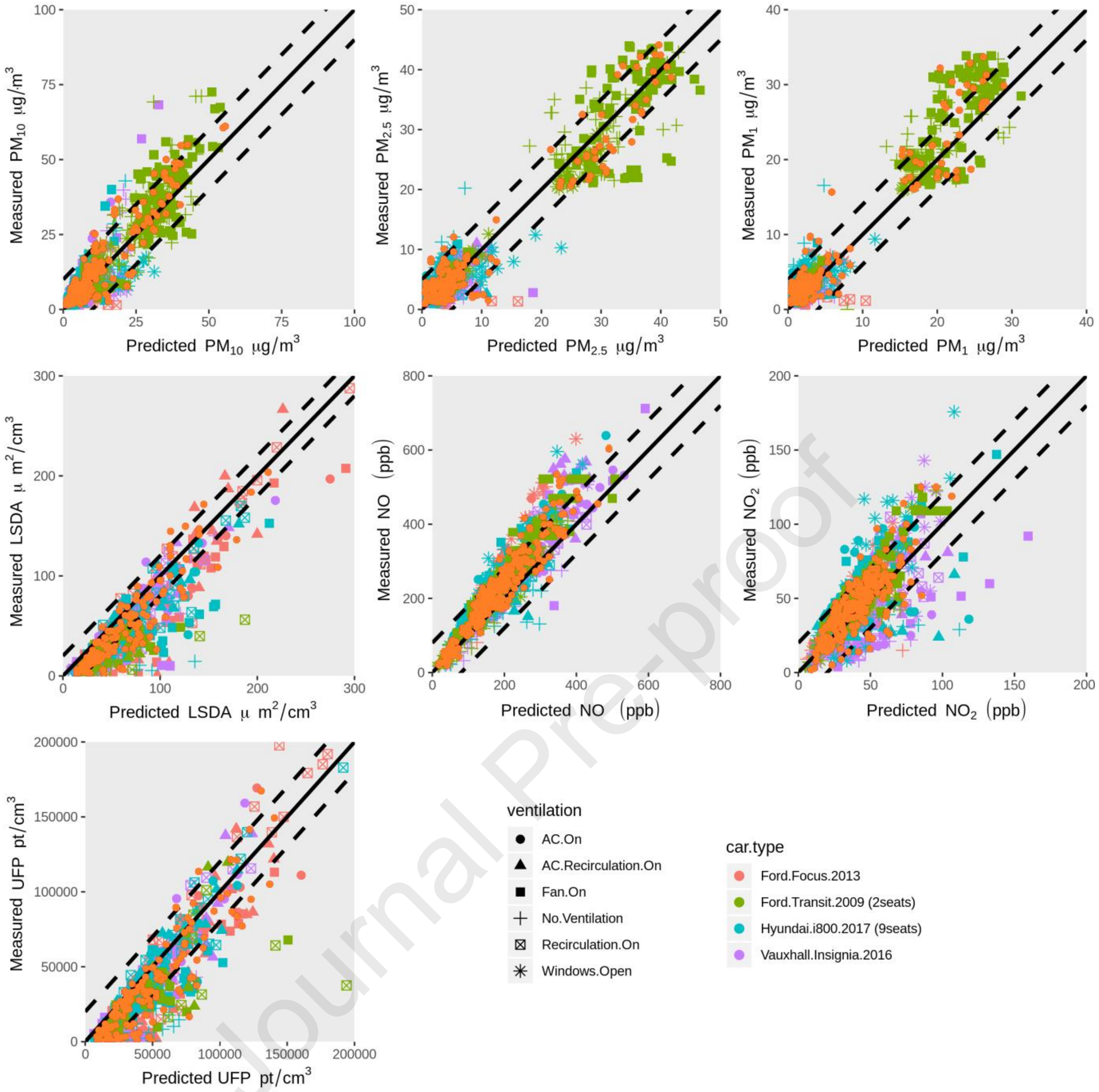


Figure 4. Measured vs MB and ML model within-vehicle concentrations of PM_{10}, $PM_{2.5}$, PM_1, LSDA, NO, NO_2 and UFP. The orange dots indicate the ML predictions for 20% of randomly excluded data. The solid line denotes the perfect model 1:1. The dashed lines indicate the ±10% of the perfect model.

3.2.2 Machine Learning (ML) model predictions

The machine learning (ML) model training method (80:20) is by definition expected to yield generally good predictions. In Figure 4 the orange dots also show the comparison between the observed and the ML modelled values for the 20% of measurements excluded from the training dataset. The ML model shows similar performance to the MB model and in some cases, such as for NO_2, it improves upon the MB model predictions. Most of the ML model predictions in almost all the species are equally spread around the 1:1 line, however, an under-prediction still occurs in the LSDA and UFP species.

Table 8 summarises the ML and MB model performance statistics against the observations (20% of withheld data in the ML case) respectively. It can be seen both models show good skill in predicting within-vehicle concentrations for all species. Pearson correlation coefficients for the ML model between ML predicted and observed values are higher than 0.80, while an IOA (index of agreement) is greater than 0.69 for all the species. For the MB model, the two indices between MB predicted and observed concentrations were slightly worse, varying between 0.45 – 0.82 and 0.48 - 0.83 for Pearson correlation coefficient and IOA respectively. However, values of IOA greater than 0.5 in general indicate good model predictions (Hurley et al., 2005; Matthaios et al., 2017). The mean gross error (MGE) of the ML and MB model's performance was less than 2.4 and 3.4 $\mu g\ m^{-3}$ respectively for all the particle classes (PM_{10}, $PM_{2.5}$ and PM_1) and 10.4 and 14.1 *ppb* for NO_2. The biggest error is evidenced in NO and UFP, which is almost the same as the mean bias. The model's fraction of predictions within a factor of two of observations (FAC2) is also in good agreement with observations for the ML model (higher than 0.66 for all the species), while noteworthy is the fact the ML model's FAC2 score is very high (0.89) for NO_2. For the MB model the FAC2 factor shows low prediction values for LSDA and UFP. NO had FAC2 greater than 1 values which indicates overprediction. The mean bias indicates that the ML model under-predicts the particulate species by less than < 1 $\mu g\ m^{-3}$ and the NO_2 by less than <5 *ppb*, while slightly greater mean bias for these species is observed for the MB model. The biggest under-prediction occurs for UFP and NO. For NO the ML has a mean underprediction of 26 *ppb* while the MB model has a mean overprediction of 35.4 *ppb*. Events such as overtaking or congestion that can result in greater NO outside and consequently inside, and particle leaks from the engine or generation of already deposited particles (in the seats or fabrics) due to vibration or movement cannot be captured in the MB model and can generate tails and cause skewness in the data. kNN algorithms are known to suffer from skewed distributions if those observations are very frequent in the data (Aha et al., 1991). Overall it can be stated that both MB and ML models showed good skill in predicting the measurement data however better predictions were observed in the ML model most likely due to the way the algorithm incorporates the data. The fact that ML improves the model's performance was also found in other studies (Ozcift and Gulten, 2011; Aquilina et al., 2018).

Table 8. Model evaluation statistics against 20% random observation data after the machine learning approach. n: indicates the number of compared values. FAC2: fraction of predictions within a factor of two of observations –perfect model FAC2 = 1. MB: Mean bias – indication of the mean over or underestimate of predictions. MGE: Mean gross error – indication of the mean error regardless of whether it is an over or underestimate. RMSE: Root mean squared error – a measure of how close predicted values are to observed values. r: Pearson correlation coefficient – values from -1 to 1 while values of 0 no prediction. IOA: Index of Agreement – values from -1 to 1. $\overline{m_O}$, $\overline{m_p}$: Mean values of observations and predictions respectively. SD: Standard Deviation.

Species	n_{ML}	n_{MB}	$FAC2_{ML}$	$FAC2_{MB}$	MB_{ML}	MB_{MB}	MGE_{ML}	MGE_{MB}	$RMSE_{ML}$	$RMSE_{MB}$	r_{ML}	r_{MB}	IOA_{ML}	IOA_{MB}	$\overline{m_O}$	$\overline{m_{ML}}$	$\overline{m_{MB}}$	SD_o	SD_{ML}	SD_{MB}
PM_{10}	196	1176	0.76	0.69	-1.06	-1.18	2.4	3.4	6.8	7.5	0.89	0.69	0.80	0.76	15	13.9	12.3	14.6	15.5	11.4
$PM_{2.5}$	196	1176	0.78	0.71	0.14	-0.25	2.3	2.8	3.4	4.2	0.94	0.80	0.87	0.83	9.9	10.2	9.4	11.8	13.4	7.8
PM_1	196	1176	0.81	0.74	-0.8	-0.9	1.6	2.1	2.3	2.8	0.96	0.82	0.89	0.83	7.6	6.8	6.7	9.02	11.1	8.2
LSDA	140	840	0.69	0.38	20.9	-18.8	22.3	29.2	28.8	32.6	0.92	0.48	0.69	0.51	48.5	69.5	26.7	50.9	52.1	82.7
NO_2	256	1536	0.89	0.55	-5.0	-8.8	10.4	14.1	15.4	22.4	0.89	0.52	0.79	0.58	45.5	40.5	36.2	24.27	33.2	49.4
NO	256	1536	0.83	1.22	-25.9	35.4	23.9	31.5	76.9	89.2	0.84	0.58	0.75	0.63	246.8	197	255.4	144.7	124	145.2
UFP	140	840	0.66	0.45	13405	18754	16518	21540	13209	26430	0.90	0.45	0.73	0.48	29841	38793	45759	43031	19870	54655

3.2.3 Extended application of ML model using data from monitoring stations

In the predictions discussed above, each model utilized external concentrations of air pollutants measured directly outside the study vehicle, to either to drive the calculated pollutant exchange (MB model), or as input for the ML model. However, in order to explore the ML model's potential wider application under real world circumstances we explored case (ii) where both within and directly-outside vehicle pollutant concentrations are unknown and the only data available is from nearby air quality monitoring stations (see 2.5). In this case, the ML model used a median within-vehicle level from all vehicles and hourly outdoor air quality measurements. The air quality levels from the monitoring sites were taken from urban-traffic locations representing different locations of the testing route. Figure 6 shows the case (ii) comparison of the ML model predicted within-vehicle pollutant concentrations, vs those measured. The results generally show some notable discrepancies for NO greater than 260 *ppb* and NO_2 greater than 60 *ppb*, of the within-vehicle air quality for a given air quality value, however the applicability of the method provides an indication of within-vehicle exposure without the need for directly-outside measurement. The ML predictions would have been more representative of the actual exposures in case where more information of the accurate representation of the ventilation system, filtration and air exchange, vehicle number and fleet composition were available.

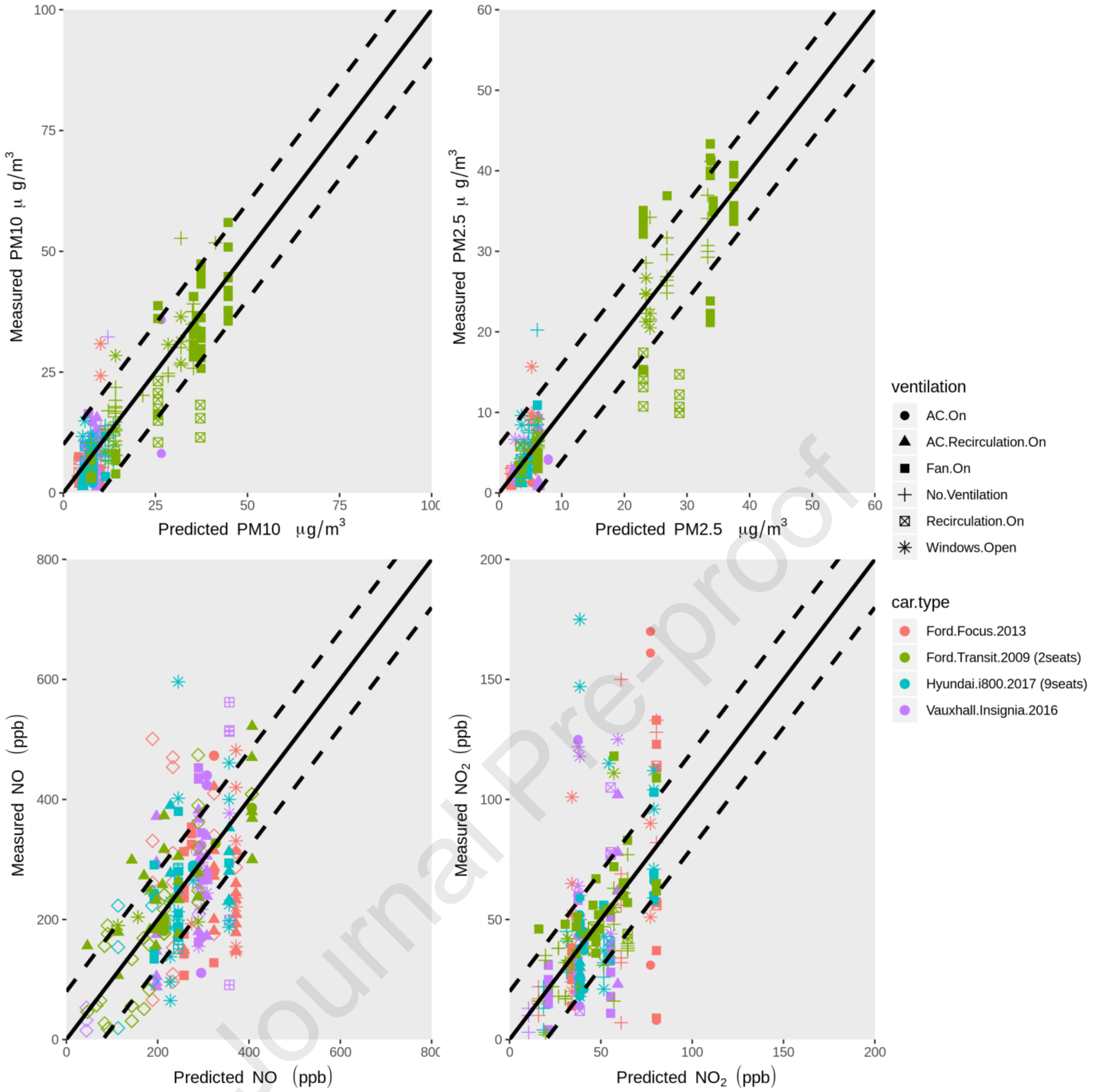


Figure 5. Comparison of within-vehicle ML modelled and measured species. For the learning of the ML model, a median within-vehicle level from all vehicles and hourly outdoor air quality measurements were used.

4 Comparison with other studies and limitations

The study investigated in-vehicle air pollution exposure with novel complementary modelling techniques using mass balance and machine learning approaches. Studies that used ML algorithms to predict in-vehicle air quality typically used low-cost sensors to calculate an air quality index that involved CO_2 and $PM_{2.5}$ and tested the performance of supervised ML algorithms against traditional regression techniques and deep-learning techniques (Sukor et al., 2022; Goh et al., 2021). Similarly, Lohani et al., 2022 compared traditional auto-regressive integrated moving average (ARIMA) and ML support vector regression (SVR) to investigate their performance against in-vehicle CO_2 levels. Chung and Kim, (2020), developed an anomaly detection system inside cars based on ML algorithms to prevent fatigue and drowsiness due to CO_2 and reduction in $PM_{2.5}$ exposures. Baldi et al., (2022), measured the performance of several ML algorithms against observations of PM_{10}, $PM_{2.5}$, PM_1, CO_2 and formaldehyde and found good results. Our study, apart from the application of ML to predict in-vehicle exposures, it offered novel expansion upon real-world applications with the implementation of air quality data from nearby monitoring sites. Several MB models have been reported for the prediction of within-vehicle concentrations of air pollutants, albeit focusing on different aspects of the problem, for example the models of Hudda et al., (2012); Knibbs et al., (2010) or Xu and Zhu, (2009). The model developed by Hudda et al., 2012, used measured data from a large number of vehicles and multi linear regression approaches and generalized estimating equations to estimate within-vehicle concentrations of UFP, while the models of Knibbs et al., (2010) and Xu and Zhu, (2009) are mass-balance based models. The differential equations applied in this work build on the mass balance studies of Knibbs et al., 2010 and Xu and Zhu, (2009), with some modifications in the equations, including incorporation of key aspects of chemical processing. The reason for the difference in some modelled vs observed levels is likely due to values such as deposition coefficients, filtration efficiency and penetration factors were taken from literature and often from experiments conducted in houses which are larger volumes than vehicle cabins and do not reflect actual within-vehicle values. Another reason might be due to our simplified approach of not having a speed dependent pressure difference penetration factor. As highlighted in Lee et al., (2015a), those factors depend on the combined effects of the ventilation conditions (i.e., ventilation mode and fan settings) and the aerodynamic changes on the vehicle envelope (i.e., driving speed and vehicle shapes) which have not yet been incorporated in this model. It should be further noted that the importance of physical air exchange processes of the outside measurements often dominate comparing to the other indoor sinks and when a rapid change of the outdoor concentrations (i.e. vehicle overtaking, high emitters etc) occurs it has implications for the modelling of within-vehicle NO and NO_2. This is likely the reason that the model underpredicts the high levels of within-vehicle NO_2.

The current MB model and methodology likely has limitations in the prediction of other more reactive species within-vehicles, where chemical processing is more important (relatively) to ingress and deposition and needs to be considered for those species; this also implies a more sophisticated treatment of physical conditions (including photolysis frequencies). The MB model assumes a well-mixed (within-vehicle) microenvironment, which may not reflect reality. Furthermore, the MB and ML models are dependent upon the initial parameters (e.g. vehicle characteristics, fan power and other within-vehicle parameters to build the model) and therefore they might be case-dependent and their applicability needs to be tested in other cases. In the model the leakage rate/passive ventilation was calculated using the equations of Hudda et al., (2012). However, since that method uses generalized regression models based on vehicle age, driving speed, and fan strength, the method may impose uncertainty across different vehicle models and other approaches to calculate the leakage flow/passive ventilation, for example based on the pressure difference (Lee et al.,2015b), or using an explicit CO_2 tracer, may be tested for suitability. Engine/fuel leaks can generate gaseous and particulate pollution and other organic gas compounds such as, benzene, toluene, xylene, and methyl-tertiary butyl ether (Faber et al., 2013; Fedoruk and Kerger, 2003; Jo and Park, 1998; Duffy and Nelson, 1997) that can enter the interior of the vehicles via the ventilation system. This source is not currently included in the model of this study. Finally, carcinogenic/toxic species such as volatile organic compounds which are released from plastics and fabrics on exposure to sunlight and heat (Yoshida and Matsunaga, 2006, You et al., 2007) and heterogeneous surface reactions or reactions of peroxy radicals with NO, can play a role in the within-vehicle chemistry and improve NO_2 predictions. The model currently is limited in omitting representation of such detailed chemistry, secondary aerosol formation and other particle physics processes.

5 Implications

The modelling methodology presented here can be developed into a useful tool that can be used by policymakers in order to estimate the air pollutant concentration levels inside vehicles. The approach presented here for the use of machine learning algorithms to predict within-vehicle exposure, showed promising applicability elsewhere and for different species.

The use of ambient monitoring data (rather than adjacent-to-vehicle measurement) to predict within-vehicle concentrations gave promising results highlighting that within-vehicle exposure can be estimated from existing air quality "infrastructure", and modelling techniques such as those presented here can be applied to estimate the associated health risks.

Future work should focus on developing more comprehensive exposure predictive models for car passengers. These models will need to account for various driving conditions (e.g., urban and motorway driving), driving durations, passenger characteristics (e.g., differing breathing rates, metabolism, sex, weight), and pathways for pollutant infiltration and penetration, including the assessment of potential in-cabin sources like engine leaks. Such information will be critical for the application of air quality management policies and new technologies such as within-vehicle air purifiers or high selectivity air cabin filters to reduce air pollution exposure. In conclusion, our study presents a novel method to predict within-vehicle air pollution exposure, which has far-reaching implications for public health and environmental research. The study has successfully demonstrated the effectiveness of the approach in providing real-time exposure estimates and mapping. We believe that this work serves as a foundational contribution to the field of real-time air pollution exposure assessment, offering a path towards cleaner and healthier urban environments. While our study is a significant step forward, we acknowledge that further research is essential to refine our approach and enhance its accuracy.

Data availability

The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments

The project was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie Grant Agreement No 895851. The measurements were supported by the UK NERC projects (SNAABL, NE/M013405/1) and WM-Air (NE/S003487/1). VNM also gratefully acknowledges University of Birmingham U21 funding and Royal Society of Chemistry Research Mobility grant that supported his travel to Australia for this study.

References

Adams W. C., 1993. Measurement of breathing rate and volume in routinely performed daily activities. Final Report Contract No. A033- 205, Air Resources Board, California Environmental Protection Agency, Sacramento, CA

Adam, M., Schikowski, T., Carsin, A. E., Cai, Y., Jacquemin, B., Sanchez, Met al., 2015. Adult lung function and long-term air pollution exposure. ESCAPE: a multicentre cohort study and meta-analysis. Eur Respir J, 45, 38–50. https://doi.org/10.1183/09031936.00130014

Aha, D. W., Kibler, D., Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37–66. doi:10.1007/bf00153759

Aidaoui, L., Triantafyllou, A.G., Azzi, A., Garas, S.K. and Matthaios, V.N., 2015. Elevated stacks' pollutants' dispersion and its contributions to photochemical smog formation in a heavily industrialized area. Air Quality, Atmosphere & Health, 8, pp.213-227.

Atkinson RW, Fuller GW, Anderson HR, Harrison RM, Armstrong B, 2010. Urban ambient particle metrics and health: a time-series analysis. Epidemiology, 21:501–11.

Aquilina N., J., Delgado-Saborit M., J., Bugelli S., Ginies J., P.,Harrison R., M., 2018. Comparison of Machine Learning Approaches with a General Linear Model To Predict Personal Exposure to

Benzene.Environmental Science & Technology 2018 52 (19), 11215-11222. DOI: 10.1021/acs.est.8b03328

Baldi, T., Delnevo, G., Girau, R. and Mirri, S., 2022, August. On the prediction of air quality within vehicles using outdoor air pollution: sensors and machine learning algorithms. In Proceedings of the ACM SIGCOMM Workshop on Networked Sensing Systems for a Sustainable Society (pp. 14-19).

Carslaw N., 2007. A new detailed chemical model for indoor air pollution. Atmos. Environ, 41 (6), 1164–1179.

Carslaw, D.C. and K. Ropkins, (2012). openair — an R package for air qualitydata analysis. EnvironmentalModelling&Software. Volume27-28,pp. 52-61.

Carslaw, D.C. (2019). The openair manual — open-source tools for analysing air pollution data. Manual for version 2.6-6, University of York.

Chen, C., Zhao, B., 2011. Review of relationship between indoor and outdoor particles: I/Oratio, infiltration factor and penetration factor. Atmos. Environ. 45, 275–288. https://doi.org/10.1016/j.atmosenv.2010.09.048.

Chung, J.J. and Kim, H.J., 2020. An automobile environment detection system based on deep neural network and its implementation using IoT-enabled in-vehicle air quality sensors. Sustainability, 12(6), p.2475.

Cover T., Hart P., 1967. Nearest neighbor pattern classification. In IEEE Transactions in Information Theory, IT-13, pages 21–27.

Delgado-Saborit, J.M., 2012. Use of real-time sensors to characterise human exposures to combustion related pollutants. J. Environ. Monit. 14, 1824–1837.

Delfino, R.J., Malik, S., Sioutas, C., 2005. Potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health. Environmental Health Perspectives 113 (8), 934-946.

De Hartog, J.J., Ayres, J., Karakatsani, A., Analitis, A., ten Brink, H., Hameri, K., Harrison, R., Katsouyanni, K., Kotronarou, A., Kavouras, I., Meddings, C., Pekkanen, J., Hoek, G., 2010. Lung function and indicators of exposure to indoor and outdoor particulate matter among asthma and COPD patients. Occupational and Environmental Medicine 67, 2-10.

DfT, Department for Transport, National Travel Survey: England 2016, 2017, July 2017https://www.gov.uk/government/statistics/national-travel-survey-2016. (accessed July 2019)

Dons, E., Int Panis, L., Van Poppel, M., Theunis, J., Willems, H., Torfs, R., Wets, G., 2011. Impact of time–activity patterns on personal exposure to black carbon. Atmos. Environ.45, 3594–3602. https://doi.org/10.1016/j.atmosenv.2011.03.064.

Frederickson LB, Lim S, Russell HS, Kwiatkowski S, Bonomaully J, Schmidt JA, Hertel O, Mudway I, Barratt B, Johnson MS. Monitoring Excess Exposure to Air Pollution for Professional Drivers in London Using Low-Cost Sensors. Atmosphere. 2020; 11(7):749. https://doi.org/10.3390/atmos11070749

Fruin S. A., Hudda N., Sioutas C., Delfino R. J., 2011. Predictive Model for Vehicle Air Exchange Rates Based on a Large, Representative Sample. Environmental Science and Technology, Vol. 45, pp. 3,569-3,575.

Goh, C.C., Kamarudin, L.M., Zakaria, A., Nishizaki, H., Ramli, N., Mao, X., Syed Zakaria, S.M.M., Kanagaraj, E., Abdull Sukor, A.S. and Elham, M.F., 2021. Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. Sensors, 21(15), p.4956.

Gong L., Xu B., Zhu Y., 2009. Ultrafine particles deposition inside passenger vehicles. Aerosol Sci. Technol., 43, 544–553

Hachem, M., Saleh, N., Bensefa-Colas, L. and Momas, I. (2021), Determinants of ultrafine particles, black carbon, nitrogen dioxide, and carbon monoxide concentrations inside vehicles in the Paris area: PUF-TAXI study. Indoor Air, 31: 848- 859. https://doi.org/10.1111/ina.12779

Hamra, G. B., Laden, F., Cohen, A. J., Raaschou-Nielsen, O., Brauer, M., Loomis, D. 2015. Lung cancer and exposure to nitrogen dioxide and traffic: A systematic review and meta-analysis. Environmental Health Perspectives, 123 (11), 1107–1112. https://doi.org/10.1289/ehp.1408882

Harrison, R.M., 2018. Urban atmospheric chemistry: a very special case for study. npjClim. Atmos. Sci. 1, 5. https://doi.org/10.1038/s41612-017-0010-8.

Heal M. R., Kumar P., Harrison R. M, 2012. Particles, air quality, policy and health. Chem Soc Rev 41:6606–30.

Hinds W. C., 1999. Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles. Wiley, New York.

Hudda, N., Eckel, S.P., Knibbs, L.D., Sioutas, C., Delfino, R.J., Fruin, S.A., 2012. Linking in-vehicle ultrafine particle exposures to on-road concentrations. Atmos. Environ. 59,578–586. https://doi.org/10.1016/j.atmosenv.2012.05.021.

IARC, (2014). Diesel and Gasoline engine exhausts and some nitroarenes. Volume 105 IARC monographs on the evaluation of carcinogenic risks to humans. https://monographs.iarc.fr/wp-content/uploads/2018/06/mono105.pdf (accessed June 2020)

Kotsiantis S. B., 2007. Supervised machine learning: A review of classification techniques. Informatica, 31, 249–268.

Knibbs, L.D., de Dear, R.J., Morawska, L., 2010. Effect of cabin ventilation rate on ultrafineparticle exposure inside automobiles. Environ. Sci. Technol. 44, 3546–3551. https://doi.org/10.1021/es9038209.

Knibbs, L. D.; de Dear, R. J.; Atkinson, S. E. Field study of air change and flow rate in six automobiles. Indoor Air 2009, 303–313.

Kumar, P., Hama, S., Nogueira, T., Abbass, R.A., Brand, V.S., de Fatima Andrade, M., Asfaw, A., Aziz, K.H., Cao, S.J., El-Gendy, A. and Islam, S., 2021. In-car particulate matter exposure across ten global cities. Science of the total environment, 750, p.141395.

Kumar, P., Rivas, I., Singh, A.P., Ganesh, V.J., Ananya, M. and Frey, H.C., 2018. Dynamics of coarse and fine particle exposure in transport microenvironments. NPJ climate and atmospheric science, 1(1), p.11.

Lawin H., Fanou L. A., Hinson A. V., Stolbrink M., Houngbegnon P., Kedote N. M., Fayomi B., Kagima J., Katoto P., Ouendo E. M. D., Mortimer K., 2018. Health Risks Associated with Occupational Exposure to Ambient Air Pollution in Commercial Drivers: A Systematic Review. Int. J. Environ. Res. Public Health, 15, 2039; doi:10.3390/ijerph15092039.

Lee E. S., Stenstrom M. K., Zhu Y.F., 2015a. Ultrafine particles infiltration into passenger vehicles Part I: experimental evidences. Transp. Res. Part D.: Transp. Environ. 38, 156–165.

Lee E. S., Stenstrom M. K., Zhu Y.F.,2015b. Ultrafine particle infiltration into passenger vehicles, Part II: model analysis. Transp Res D.; 38: 144–155.

Lohani, D., Barthwal, A. and Acharya, D., 2022. Modeling vehicle indoor air quality using sensor data analytics. Journal of Reliable Intelligent Environments, pp.1-11.

Shanon Lim, Benjamin Barratt, Lois Holliday, Chris J. Griffiths, Ian S. Mudway, Characterising professional drivers' exposure to traffic-related air pollution: Evidence for reduction strategies from in-vehicle personal exposure monitoring, Environment International, Volume 153, 2021, 106532, https://doi.org/10.1016/j.envint.2021.106532.

Madronich, S.: The atmosphere and UV-B radiation at ground level. Environmental UV Photobiology, Plenum Press, 1–39, 1993.

Martin, A.N., Boulter, P.G., Roddis, D., McDonough, L., Patterson, M., Rodriguez del Barco, M., Mattes, A., Knibbs, L.D., 2016. In-vehicle nitrogen dioxide concentrations in road tunnels. Atmos. Environ. 144:234–248. http://dx.doi.org/10.1016/j.atmosenv.2016.08.083

Mathai, V.; Das, A.; Bailey, J.A.; Breuer, K. Airflows inside passenger cars and implications for airborne disease transmission. Sci.Adv. 2021, 7, eabe0166.

Matthaios V. N., Triantafyllou A. G., Albanis T. A., Sakkas V., Garas S., 2017. Performance and evaluation of a coupled prognostic model TAPM over a mountainous complex terrain industrial area. Theor Appl Climatol 132:885–903. https://doi.org/10.1007/s00704-017-2122-9.

Matthaios N. V., Kramer J. L., Sommariva R., Pope D. F., Bloss J. W., 2019. Investigation of vehicle cold starts primary NO_2 emissions from ambient monitoring data in the UK and their implications for urban air quality. Atmospheric Environment 199, 402-414, DOI: 10.1016/j.atmosenv.2018.11.

Matthaios, V. N., Kramer, L. J., Crilley, L. R., Sommariva, R., Pope, F. D., Bloss, W. J., 2020. Quantification of within-vehicle exposure to NOx and particles: Variation with outside air quality, route choice and ventilation options. Atmospheric Environment, 117810. doi:10.1016/j.atmosenv.2020.117810

Matthaios, V.N., Rooney, D., Harrison, R.M., Koutrakis, P. and Bloss, W.J., (2023a). NO_2 levels inside vehicle cabins with pollen and activated carbon filters: A real world targeted intervention to estimate NO_2 exposure reduction potential. Science of the Total Environment, 860, p.160395.

V.N. Matthaios, R.M. Harrison, P. Koutrakis, W.J. Bloss, (2023b). In-vehicle exposure to NO_2 and $PM_{2.5}$: A comprehensive assessment of controlling parameters and reduction strategies to minimise personal exposure, Science of the Total Environment, https://doi.org/10.1016/j.scitotenv.2023.165537

Nazaroff, W.N., Cass, G.R., 1986. Mathematical modeling of chemically reactive pollutants in indoor air. Environmental Science and Technology 20 (9), 924–934.

Ott W., Klepeis N., Switzer P., 2008. Air change rates of motor vehicles and in-vehicle pollutant concentrations from secondhand smoke. J. Exposure Sci. Environ. Epidemiol., 18, 312–325

Ozcift, A., Gulten, A., 2011. Classifier Ensemble Construction with Rotation Forest to Improve Medical Diagnosis Performance of Machine Learning Algorithms. Comput. Methods Programs Biomed. 552, 104 (3), 443–451.

Postlethwait, E. M., and Bidani, A., 1990. Reactive uptake governs the pulmonary airspace removal of inhaled nitrogen dioxide. J. Appl. Physiol. 68:594-603.

Qi, C.; Stanley, N.; Pui, D. Y. H.; Kuehn, T. H. Laboratory and on-road evaluations of cabin air filters using number and surface area concentration monitors. Environ. Sci. Technol. 2008, 42, 4128–4132.

Song, Y., Liang, J., Lu, J., Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26–34. doi:10.1016/j.neucom.2017.04.01

Sukor, A.S.A.; Cheik, G.C.; Kamarudin, L.M.; Mao, X.; Nishizaki, H.; Zakaria, A.; Syed Zakaria, S.M.M.

Predictive Analysis of In-Vehicle Air Quality Monitoring System Using Deep Learning Technique.

Atmosphere 2022, 13, 1587. https://doi.org/10.3390/atmos13101587

Thatcher, T.L., Lunden, M.M., Revzan, K.L., Sextro, R.G., Brown, N.J., 2003. A concentration rebound method for measuring particle penetration and deposition in the indoor environment. Aerosol Sci. Technol. 37, 847-864.

TomTom, 2019. World traffic index, measuring congestion worldwide https://www.tomtom.com/en_gb/trafficindex/list?citySize=LARGE&continent=ALL&country=ALL

WeinbergerK. Q., BlitzerJ., SaulL. K., 2006. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press 2, 3

Wettschereck D., Aha D. W., Mohri T., 1997. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. Artificial Intelligence Review 10:1–37.

Williams, R., Suggs, J., Rea, A., Sheldon, L., Rodes, C., Thornburg, J., 2003. The Research Triangle Park particulate matter panel study: modeling ambient source contribution to personal and residential PM mass concentrations. Atmos. Environ. 37, 5365-5378.

Wilks D. S., 2005. Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition (International Geophysics). 2nd ed. Academic Press (cit. on p. 238).

Xu B., Zhu Y., 2009. Quantitative analysis of the parameters affecting in-cabin to on-roadway (I/O) ultrafine particle concentration ratios. Aerosol Sci. Technol., 43, 400–410.

Yamada, H., Hayashi, R., Tonokura, K., 2016. Simultaneous measurements of on road/in-vehicle nanoparticles and NOx while driving: actual situations, passenger exposure and secondary formations. Sci. Total Environ. 563, 944-955.

Yamamoto, N., Shendell, D.G., Winer, A.M., Zhang, J., 2010. Residential air exchange rates in three major US metropolitan areas: results from the relationship among indoor, outdoor, and personal air study 1999-2001. Indoor Air 20, 85-90.

Yoshida, T., Matsunaga, I., 2006. A case study on identification of airborne organic com-pounds and time courses of their concentrations in the cabin of a new car for privateuse. Environ. Int. 32:58–79. http://dx.doi.org/10.1016/j.envint.2005.04.009.

You, K., Ge, Y., Hu, B., Ning, Z., Zhao, S., Zhang, Y., Xie, P., 2007. Measurement of in-vehiclevolatile organic compounds under static conditions. J. Environ. Sci. 19:1208–1213.http://dx.doi.org/10.1016/S1001-0742(07)60197-1.

Zuurbier, M., Hoek, G., Oldenwening, M., Lenters, V., Meliefste, K., van den Hazel, P., Brunekreef, B., 2010. Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. Environ. Health Perspect. 118, 783–789.

- Development of a mass-balance and a machine learning model for within-vehicle exposures
- Both models demonstrated good predictions of observations apart from an underestimation in UFP and LSDA.
- The ML model predictions were as good as the MB model for most of the species and improved for NO_2.
- Use of air quality monitoring data provides new capabilities for within-vehicle exposure predictions

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: