

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](https://doi.org/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

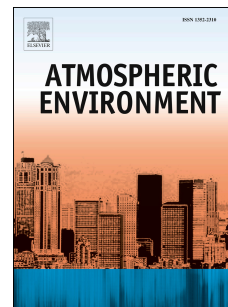
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

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.

Journal Pre-proof

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

Vasileios N. Matthaïos, 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: Matthaïos, 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>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

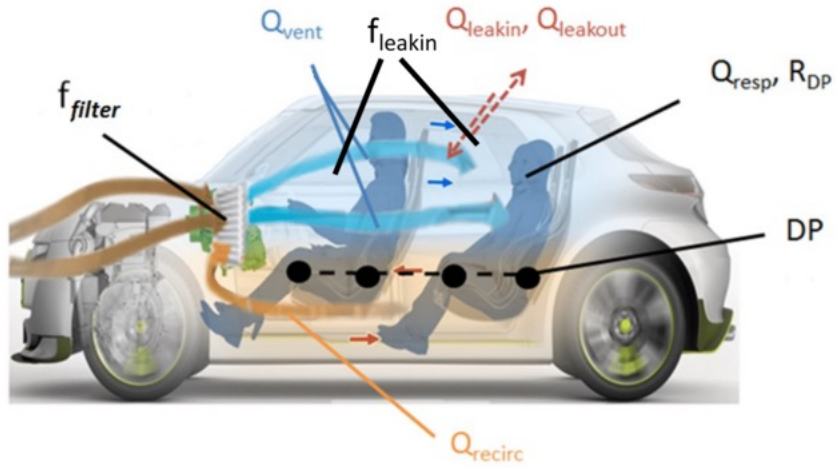
© 2023 Published by Elsevier Ltd.

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.

Journal Pre-proof

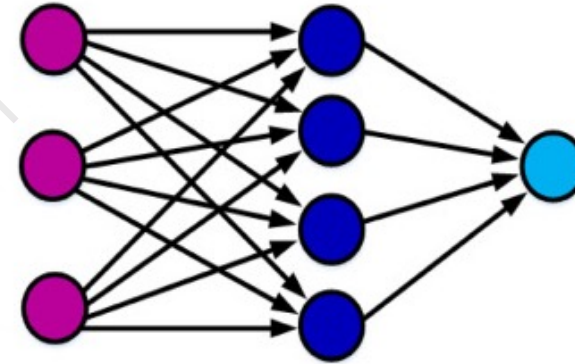
In-vehicle processes



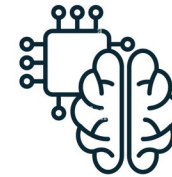
On-road air pollution

Vehicle Characteristics

Air quality monitoring data



In-vehicle air pollution exposure



MACHINE LEARNING

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

3
4 Vasileios N. Matthaios^{a*}, Luke D. Knibbs^{b,c}, Louisa J. Kramer^{d1}, Leigh R. Crilley^{d2} and William J. Bloss^d

5 ^aDepartment of Public Health, Policy and Systems, University of Liverpool, Liverpool L69 3GB, United
6 Kingdom

7 ^bSchool of Public Health, The University of Sydney, NSW 2006, Australia

8 ^cPublic Health Research Analytics and Methods for Evidence, Public Health Unit, Sydney Local Health
9 District, Camperdown, NSW 2050, Australia

10 ^dSchool of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston,
11 Birmingham, B15 2TT, United Kingdom

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

13 ²now at WSP Australia, Brisbane, 4006 Australia

14
15 *Corresponding author: Department of Public Health, Policy and Systems, Institute of Population
16 Health, University of Liverpool, Liverpool L69 3GB, UK

17 Email: V.Matthaios@liverpool.ac.uk

18
19 Modelling the air pollutant concentrations within-vehicles is an essential step to estimate our
20 daily exposure to air pollution. This is a challenging issue however, since the processes that affect the
21 exposures within-vehicles change with different driving patterns and ventilation settings. This study
22 introduces an innovative approach that combines mass-balance principles and machine learning
23 techniques, leveraging ambient air quality, on-road and within-vehicle measurements of particulate
24 matter (PM₁₀, PM_{2.5}, PM₁), nitrogen dioxide (NO₂), nitrogen oxides (NO_x), aerosol lung surface
25 deposited area (LSDA) and ultrafine particles (UFP) under different ventilation settings to estimate air
26 pollution exposure levels within vehicles. The first model (MB) includes basic physical and chemical
27 processes and follows a mass-balance approach to estimate the within-vehicle concentrations. The
28 second model (ML) applies data driven machine learning algorithms to a training set of observations
29 to predict unseen within-vehicle concentrations. By using a number generator, the whole

30 observational dataset was divided to 80:20 and 80% was used to build and train the ML model, while
31 20% was used for validation. Both models demonstrated good predictions of observations apart from
32 an underestimation in UFP and LSDA. The ML model showed better predictive power than the MB
33 model and had skill in predicting the unseen within-vehicle exposures. The ML model predictions were
34 as good as the MB model for most of the species and improved for NO₂. The ML model demonstrated
35 good index of agreement (IOA > 0.69) and Pearson correlation coefficient ($r > 0.80$) for all the species.
36 The inclusion of air quality data from nearby monitoring stations instead of on-road (sampled while
37 driving), in the ML model showed promising and new capabilities to within-vehicle exposure
38 predictions. In an era where air pollution is a growing concern, understanding and predicting within-
39 vehicle air pollution exposure is of great importance for public health and environmental research.
40 This research not only advances the field of exposure assessment but (at no extra cost) also
41 demonstrates practical implications for real-time exposure mapping and health impact assessment of
42 vehicle occupants with existing infrastructure.

43

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

46

47 Introduction

48 Road traffic is the dominant source of nitrogen dioxide (NO₂) and a significant contributor to
49 particulate matter (PM₁₀, PM_{2.5}, PM₁ and ultrafine particles – UFP) in the atmospheres of urban
50 environments. Numerous studies have highlighted the relationship between traffic related air
51 pollution and adverse health effects such as cardiopulmonary disease, respiratory symptoms, reduced
52 lung function changes in cardiac function and increased lung cancer risk (Adam et al., 2015; Hamra et
53 al., 2015; IARC, 2014; Heal et al., 2012; Atkinson et al., 2010; De Hartog et al., 2010; Delfino et al.,
54 2005). The road traffic dominance of many primary air pollutant emissions in urban areas leads to
55 strong roadside concentration increments relative to urban background and rural areas (Harrison,
56 2018).

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

62 UK, there are approximately 32 million registered full driving license holders, of which 6% are
63 professional drivers (DfT, 2017) who may be subject to particularly extended and elevated exposures
64 of within-vehicle air pollution (Frederickson et al., 2020). Previous studies measuring exposure inside
65 vehicles have found within-vehicle concentrations of $PM_{2.5}$ to be a factor of 2-3 larger than in other
66 transport modes (e.g. De Nazelle et al., 2012; Zuurbier et al., 2010; Kumar et al., 2018), while BC and
67 NO_2 levels inside cars can be 4.5 and 1.4 times greater than ambient concentrations (Delgado-Saborit,
68 2012). Other studies investigated the impact of ventilation settings on within-vehicle exposure and
69 found that exposure was highly dependent on the air intake, vehicle age and air leaks (Kumar et al.,
70 2021; Martin et al., 2016; Hudda et al., 2012; Knibbs et al., 2010). To inform policies, studies have also
71 identified filtration media and usage as important factors that can help reduce within-vehicle
72 exposures (Hachem et al., 2021; Lim et al., 2021; Matthaios et al., 2023a; Matthaios et al., 2023b).
73 Limited studies have also directly compared pollutant levels within-vehicle with those immediately
74 outside/adjacent to the vehicle, for both particulate and gaseous species highlighting the potentially
75 greater health impact of NO_2 over PM exposure (Yamada et al., 2016). However, measuring within-
76 vehicle exposure to air pollution with direct certified methods is very expensive and challenging and,
77 given that it needs continuous monitoring, only offers a snapshot of the actual exposures. Therefore,
78 alternative indirect approaches, such as the modelling that utilize already available air quality
79 measurements from monitoring sites need to be explored.

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

89 In light of the range of potential implications of improving the air quality in one of the most
90 common microenvironments, and to provide new capabilities in real-time predicting and regulating
91 the exposure of vehicle occupants, this study reports the development of two innovative and
92 complementary approaches to simulate within-vehicle passenger exposure to air pollutants as a
93 function of outside (ambient) levels and vehicle ventilation conditions. The first approach involves the
94 development of a mass-balance (MB) model, which explicitly represents the aforementioned
95 (predominant) physical and chemical processes which drive changes in within-vehicle air pollutant

96 abundance. The second approach uses machine-learning algorithms (ML model), which seek to
97 replicate the observed within-vehicle data based upon a training set of observations of internal and
98 external (outside, ambient) pollutant concentrations, and which does not include any mechanistic
99 representation. The results from the MB model are compared with time series measurements of
100 within-vehicle concentrations, while the results from the ML model are compared with a subset of
101 observations which were excluded from the training dataset. The performance of both models in
102 estimating within-vehicle air pollution exposure is evaluated using two contrasting measures of
103 outside (ambient) pollutant levels: (i) observations obtained directly outside the test vehicles and (ii)
104 observations from roadside air quality monitoring stations within the same locality as the vehicle, but
105 at some distance away from its immediate location. The objective of this study is not only to evaluate
106 the effectiveness of this approach but to unveil its far-reaching implications for real-time exposure
107 mapping, health impact assessment, and policy development.

109 2. Methods

111 2.1 Measurements, tested vehicles and ventilation conditions

112 Model development and validation was supported by measurements of NO, NO₂, O₃, PM₁₀,
113 PM_{2.5}, PM₁, ultrafine particle number (UFP) and aerosol lung surface deposited area (LSDA), which
114 were performed concurrently within vehicle cabins (in the breathing zone of the driver) and directly
115 outside (at the side window of) the tested vehicle. CO₂ measurements were performed with two LICOR
116 LI-820 infra-red analysers, NO_x (NO + NO₂) with chemiluminescent 42i and 42C thermo-scientific
117 analysers, O₃ with 49i thermo-scientific analysers, PM with alphasense OPC-N2, and UFP/LSDA with
118 DiSCmini. Temperature and relative humidity were also measured inside the vehicle cabin using HOBO
119 sensors. Measurements were performed during two periods in 2017 in four study vehicles (see Table
120 1) in Birmingham (UK). Five core ventilation settings were investigated and a sixth setting was applied
121 in two out of four vehicles: (a) front windows (of driver and co-driver) fully open, fans and AC off, (b)
122 all windows closed; ventilation fans on (c) all windows closed; ventilation fans on with air-conditioning
123 (AC) (d) all windows closed, ventilations fans on, recirculation mode (no AC) (e) all windows closed,
124 ventilation fans on, recirculation mode, AC on (in two vehicles) and (f) all windows closed, ventilation
125 system off. Fan power (air flow setting) was varied in some vehicles as outlined later. Details of the
126 sampling campaign and quality assurance of the measurements are discussed elsewhere (Matthaios
127 et al., 2020).

129

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

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

131

132 **2.2 Description of within-vehicle processes and modelling**

133 Physical air exchange processes are represented schematically in Figure 1. These give rise to
 134 an overall cabin air exchange rate from a combination of active ventilation options, passive in-built
 135 ventilation and/or leaks. The introduction of ambient pollutants may be further modified by filtering
 136 (in the case of the ventilation system). These physical processes may be described by the parameters
 137 summarised in Table 2. Considering mechanical flow alone, under recirculatory ventilation conditions,
 138 $Q_{leakin} = Q_{leakout}$ and $Q_{vent} = 0$, while under non-recirculatory ventilation settings, $Q_{vent} + Q_{leakin} = Q_{leakout}$
 139 and $Q_{recirc} = 0$. The penetration (or removal) of air pollutants through each cabin entry mechanism can
 140 be represented by a dimensionless filtration efficiency, f , which represents the fraction of a given
 141 pollutant removed by each entry process. Deposition characterises the rate at which pollutants have
 142 losses to surfaces.

143

144

145

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

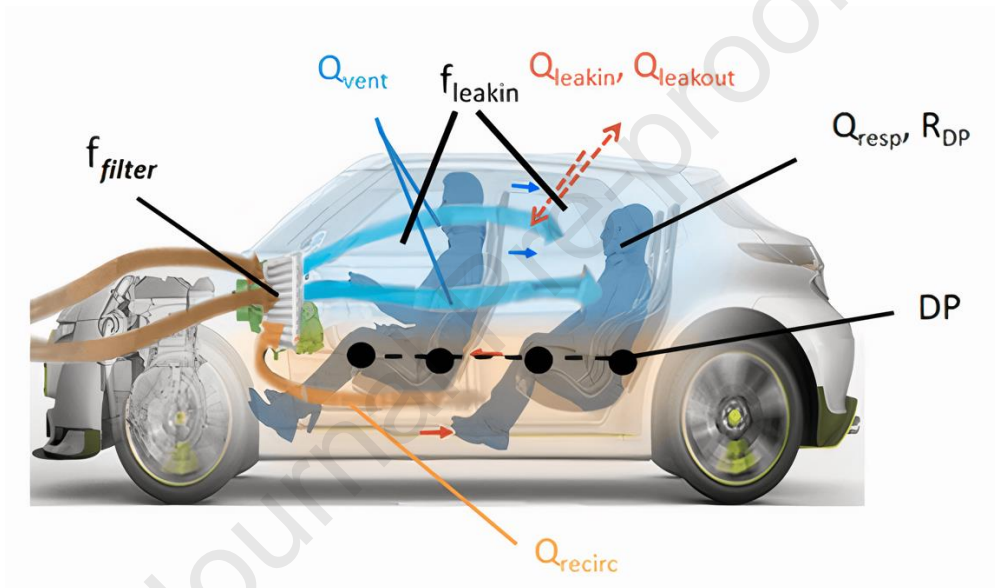
Process	Parameter	Nature	Units	Value Used
Ambient air entering through ventilation system	Q_{vent}	Flow rate	$m^3 h^{-1}$	Vehicle & ventilation setting specific
Recirculation flow through the ventilations system	Q_{recirc}	Flow rate	$m^3 h^{-1}$	Vehicle & ventilation setting specific
Leakage: Ambient 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_{O_3} Dp_{NO} Dp_{NO_2}	Species deposition rate coefficient	h^{-1}	Species-specific values used
Vehicle volume	V	volume	m^3	Vehicle specific

148

149

150



151

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

157

158 2.3. Mass balance modelling approach (MB)

159 2.3.1 Mechanism

160 The mass balance (MB) model developed in this study predicts air pollutant concentrations
 161 within vehicles taking into account the physical processes illustrated in Figure 1 and a representation
 162 of the gas-phase NO_x - O_3 photostationary steady state chemistry; no other physical or chemical

163 processes are considered here. For a given time interval, the MB model defines the rate of change of
 164 the within-vehicle air pollution concentration (following Xu and Zhu, 2009; Knibbs et al., 2010) as
 165 arising from the sum of pollutant inflow from outside (ambient) air, adjusted for filtration factors,
 166 pollutant outflow from the vehicle (both ventilation dependent), cabin surface and occupant
 167 inhalation deposition, and photochemical formation and removal (for NO_x - O_3). Air is assumed to be
 168 instantaneously homogeneously mixed within the vehicle cabin. No chemical processing of PM is
 169 considered. The mathematical equation for the MB model is given in Eq (1):

170

$$171 \frac{d(C_{inj}V)}{dt} = C_{outj} [Q_{vent}(1 - f_{vent}) + Q_{Lin}f_{leakinj}] - C_{inj}[Q_{resp}RD_p + Dp_jV + (Q_{vent} +$$

$$172 Q_{Lout}) + \sum_{j=1}^n R_{ij}] \quad [1],$$

173

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

184

185

186

187 2.3.2 Parameters and initial conditions for Mass Balance (MB) model

188

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

193 Xu and Zhu, (2009). The calculated mechanical flows were $346 \text{ m}^3 \text{ h}^{-1}$, and $173 \text{ m}^3 \text{ h}^{-1}$ for full and
 194 intermediate fan power levels respectively, while for the two front fully open windows a flow of 692
 195 $\text{m}^3 \text{ h}^{-1}$ was used (assuming two-fold amplification of the fan full power; Ott et al., 2008; Knibbs et al.,
 196 2009; Mathai et al., 2021).

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

$$202 \ln(Q_L) = 2.79 + (0.019 \times S) + (0.015 \times v.age + 3.3 \times 10^{-3} v.age^2) + (-0.023 \times V + 6.6 \times$$

$$203 10^{-5} V^2) + m, \quad [2]$$

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

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

212 **Respiratory deposition coefficient (RD_p):** Respiratory deposition is the net loss of particles in
 213 the human respiratory system. Here, the respiratory deposition coefficient can be considered
 214 analogous to filtration efficiency, where it represents the fractional loss of pollutant species during
 215 breathing. For UFP and LSDA (median measured value of 50 nm) we adopted the RD_p from Hinds,
 216 (1999) for light exercise (0.55): For PM_{10} and $\text{PM}_{2.5}$ the equivalent RD_p is 0.65 while for PM_1 it is 0.55.
 217 For NO and NO_2 a respiratory deposition coefficient of 0.67 as reported in Postlethwait and Bidani,
 218 (1990) was used.

219 **Deposition rate coefficient (D_p):** Dry deposition is a surface loss mechanism inside vehicles
 220 (Thutcher et al., 2002). Deposition rate coefficients differ between within-cabin and indoor
 221 microenvironments, as air exchange rates are much greater inside vehicles (Ott et al., 2007; Knibbs et
 222 al., 2010; Hudda et al., 2012) comparing to buildings (Yamamoto et al., 2010) if there is no indoor
 223 particle source. For UFP and LSDA (50 nm size) we used the fixed deposition rate coefficient of 10 h^{-1} ,
 224 as in Gong et al., (2009). This value was applied for two reasons: 1) the mean size of our UFP and LSDA
 225 for particles was 50 nm which is possibly due to the nature of the particles (i.e. coming from diesel
 226 exhaust) and 2) the deposition rate for particles in the range of $100 - 30 \text{ nm}$ in the observational study

227 of Gong et al., 2009 showed little variation from deposition rate spanning from 9.5 – 11.5 h^{-1} . For PM
 228 deposition values in Table 3 for different ventilation options we used the values provided by Ott et al.,
 229 (2007). For NO and NO₂ we used values from Nazaroff and Cass, (1987) for indoor NO₂ decay rates in
 230 a house. Values were applied to all study vehicles.

231 **Ventilation filtration efficiency (f_{vent}):** The filtration efficiency is how well the vehicle's air
 232 filtration system removes pollutants in the incoming airflow. This filtration efficiency varies for PM₁₀
 233 and PM_{2.5} depending on the experimental conditions and filter characteristics. However, since the
 234 filtration efficiency was not tested in this study, values from Qi et al., (2008), who tested vehicle
 235 particle filter efficiency in two different velocities representing low and full power fan settings, were
 236 adopted (see Table 3). Pollen filter efficacy of 0.10 based on Matthaïos et al., (2023a) was applied for
 237 gases, as none of the test vehicles had activated charcoal filtration for NO₂ removal (three of the
 238 vehicles were equipped with pollen filters and one had no filter).

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

244 **Reaction and Photolysis rates:** The only reactions considered here are the (overall)
 245 photostationary steady state reactions of $NO_2 + hv \rightarrow NO + O$, $NO + O_3 \rightarrow NO_2 + O_2$ and $O +$
 246 $O_2 + M \rightarrow O_3 + M$. The NO + O₃ reaction rate constant was calculated using the Arrhenius
 247 expression with the measured temperatures within-cabin, and the O + O₂ recombination reaction was
 248 assumed to be instantaneous. The photolysis frequency varies based on the window design, vehicle
 249 orientation and incident sunlight (time, location). These variations can result in differences in the
 250 experienced actinic flux (Carslaw, 2007). A ratio of photolysis frequencies of 1:10 for
 251 $j(NO_2)_{indoor}:j(NO_2)_{outdoor}$ values reported in Carslaw, (2007) for buildings was used. The corresponding
 252 outdoor photolysis rates $j(NO_2)$ outdoor were taken from the TUV model (Madronich, 1993) for each
 253 measurement time / location, assuming clear-sky conditions.

254 The model was used to simulate the time-varying within-cabin pollutant concentrations for
 255 each vehicle, and each ventilation setting. This typically corresponded to a total run-time of 35
 256 minutes, using a model timestep of 1 second. The timescale for PSS reactions is approximately 50s
 257 (under typical continental boundary layer daytime conditions), while the typical air residence time
 258 inside the vehicle can be as little as 16s (for an inflow of 0.192 m^3/s under windows open) and 31s and
 259 63s (for an inflow of 0.096 m^3/s and 0.048 m^3/s under full and intermediate fan power ventilation
 260 settings respectively). In each case, outside pollutant concentrations were set to their actual

261 (measured, time-varying) levels. The model was initiated with actual measured within-cabin pollutant
 262 concentrations.

263

264 Table 3: Parameters used for the Eq (1), (2) and (3); a) from Ott et al., 2008, b) Calculated in the study,
 265 c) Values from Gong et al., (2009) for the median UFP (50nm) size in this study d) Values from Nazaroff
 266 and Cass, (1987) for indoor NO₂ decay rates in a house e) Values from Thatcher et al., (2003), f) Values
 267 from Williams et al., (2003), g) average value from the studies reported in Chen and Zhao, (2011), h)
 268 According to light exercise and sitting from Hinds (1999) for UFP size 50nm, i) Postlethwait and Bidani,
 269 (1990) j) Values from Qi et al., (2008); +: Values used for Windows open, ++: Values used for Fan on,
 270 AC on, +++: Values used for All closed, Recirculation on; *: Full fan power, **: Low fan power; ‡: No
 271 filter efficiency was applied none of the cars was equipped with charcoal filter.

Species	Deposition rate coefficient (D_p)	Penetration factor (P)	Respiratory deposition coefficient (RD_p)	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	0.65 ^{j*} , 0.45 ^{j**}
PM ₁	54.82 ^{b+} , 11.93 ^{b++} , 5.85 ^{b+++}	0.8 ^g	0.55 ^h	0.4 ⁱ
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 [‡]

272

273

274 Table 4: Parameters changed during the modelling between different vehicles. Q_s : Mechanical
 275 supplied air, Q_l : vehicle leakage, **: full fan strength; *: intermediate fan strength; +: front windows
 276 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 ⁺⁺
----------------------------------	------------------	------------------	------------------	------------------

277

278

279 **2.4 Machine learning model (ML) and cross validation.**

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

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

302 The machine learning applied in this study used the original 80% of the within-vehicle
 303 observations of the complete dataset, selected using a random number generator. The remaining 20%
 304 was reserved to validate and test the model's predictability and response (after the ML training) to
 305 fresh unseen data. In detail, the ML training dataset used within-vehicle concentrations as the
 306 response variable, and the training was built upon the variables of on-road concentrations, time of

307 day, day of week ventilation power (expressed as 0, 50 and 100), ventilation type (expressed from 1
308 to 6), and cabin surface area and cabin volume of the vehicles. The kNN ML training and
309 hyperparameter tuning (number of neighbours (k); distance metric; weighing of neighbours) followed
310 the repeated grid search and k -fold cross validation approach. Mathematical description of the kNN
311 algorithm used here can be found in supplementary information. In this method, after randomly
312 splitting the training data into k -folds (10 in this case), a ML model was trained for $k-1$ folds (training
313 fold) of the dataset and tested on the k^{th} (testing fold). For each fold/subset that was held out, the
314 model was trained on all other subsets. This training process was repeated 1000 times and the final
315 model accuracy was taken as the average of those repeats. More repetitions provide better accuracy
316 for each instance in the dataset, however it should be mentioned that this requires more
317 computational power. This process maximizes the training and the testing of the ML algorithm and
318 has the advantage that for a single dataset all the available values are used for training and testing.
319 This method is robust for estimating the accuracy of the model and the size of k and tunes the amount
320 of bias in the predictions; Principles which are critical when using a kNN approach (Kotsiantis, 2007).
321 Finally, the ML model (built from $k-1$ folds and tested on the k^{th} fold with 1000 repeats) was evaluated
322 against the 20% of the initially randomly excluded data to assess its performance. A comparison of the
323 three machine learning algorithms tested are listed in Table S2.

324

325 **2.5 Model evaluation and real-world application scenarios**

326 To evaluate / validate the MB and ML models we used the statistical indices of: 1) Root mean
327 square error (RMSE) between the predicted and observed pollutant concentrations, where the closer
328 the RMSE is to 0 the better the model prediction (Aidaoui et al., 2015; Matthaios et al., 2017); 2)
329 fraction of predictions within a factor of 2 of observations (FAC2), where the predictions vary between
330 $0.5 \leq \text{FAC2} \leq 2$ and $\text{FAC2} = 1$ is the perfect prediction; 3) mean bias (MB), which is the relative mean
331 over or under estimation of the model predictions; 4) Mean Gross Error (MGE), which provides an
332 indication of the mean error of the model regardless of whether it is an over or under estimate; 5)
333 Pearson correlation coefficient (r), which represents the strength of the linear relationship between
334 two variables; 6) Index of agreement (IOA) which is a measure of how well the predicted variations
335 are represented around the mean observations and ranges from 0 to 1, and 7) comparison of means
336 (for observed and predicted values). The model evaluation statistics were performed with openair
337 package in R (Carslaw, 2019; Carslaw and Ropkins, 2012)

338 To examine the predictability of the MB model and the applicability of the ML model, we
339 tested two further cases: (i) in the MB model we replaced the initial within-vehicle concentrations

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

348
 349

350

351

352

353

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

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

357

358 **3. Results**

359

360 **3.1 Measured concentrations.** The measurements of ventilation-setting-dependent within-vehicle
 361 concentrations is discussed briefly in section 2.1 and in detail in Matthaios et al., (2020). Here, Table
 362 6 presents the median of the concentrations measured. As anticipated, the highest exposure to
 363 exhaust-related gaseous (NO_2 and NO_x) and particulate (UFP and LSDA) pollutants was measured with
 364 open windows (ventilation option a). Under closed windows, the highest median exposure to
 365 particulate pollution (PM_{10} , $\text{PM}_{2.5}$, PM_1) was measured when the fan was on bringing air from outside
 366 inside (ventilation option b). The lowest mean exposure for PM_{10} , $\text{PM}_{2.5}$, PM_1 , UFP and LSDA occurs
 367 when ventilation recirculation option is selected (ventilation options d and e). The within-vehicle
 368 measurements show a strong dependence upon ventilation setting, highlighting the importance of
 369 ventilation representation for accurate within-vehicle pollutant prediction.

370

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

Species	Ventilation (a)	Ventilation (b)	Ventilation (c)	Ventilation (d)	Ventilation (e)	Ventilation (f)
PM_{10} ($\mu\text{g}/\text{m}^3$)	15	24	6	8	3	13
$\text{PM}_{2.5}$ ($\mu\text{g}/\text{m}^3$)	8	15	4	4	3	5
PM_1 ($\mu\text{g}/\text{m}^3$)	5	13	3	3	2	3
LSDA ($\mu\text{m}^2/\text{cm}^3$)	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/cm ³)	44816	31960	27265	5466	400	19110
O_3 (ppb)	8.6	4.1	4.4	2	2.2	5

376

377

378 3.2 Modelling results – Comparison with observations

379

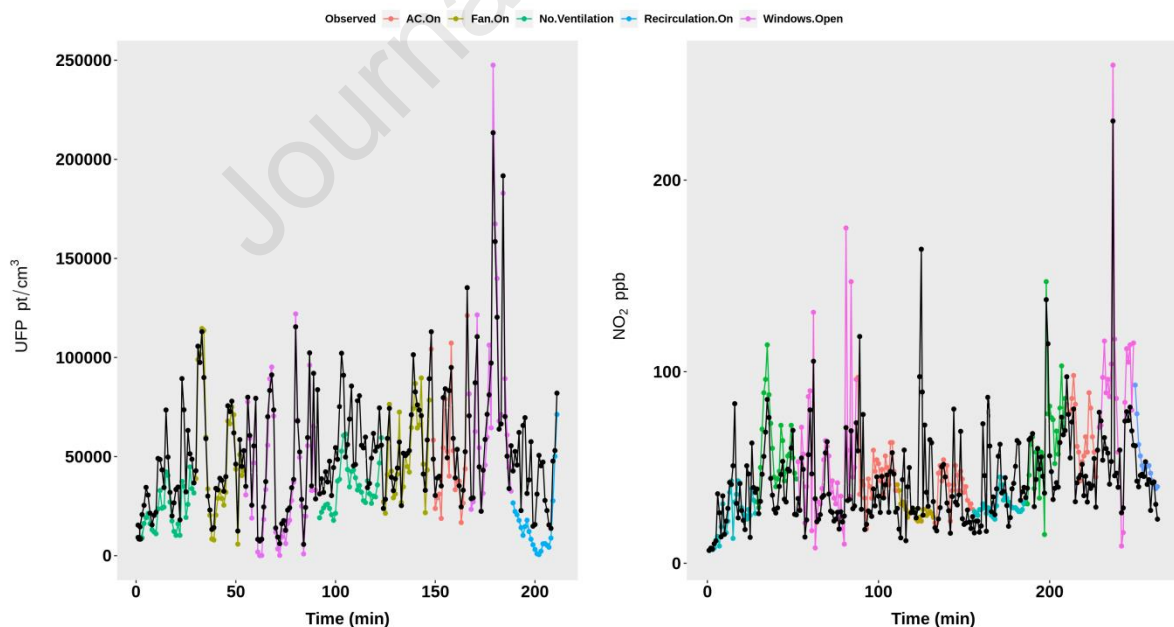
380 3.2.1 Mass-Balance model simulations

381 Figure 3 compares the timeseries of mass-balance (MB) model predictions and measured
 382 levels of (within-vehicle) UFP and NO_2 from one of the test vehicles. For UFP, the model performs well

383 under windows-open, fan-on and AC-on modes, but overpredicts the observed levels under the no-
 384 ventilation and recirculation modes. For NO_2 , the MB model performs well under no-ventilation and
 385 recirculation conditions but underestimates the observations for windows-open and AC-on, and
 386 overestimates for fan-on and AC-with-recirculation.

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

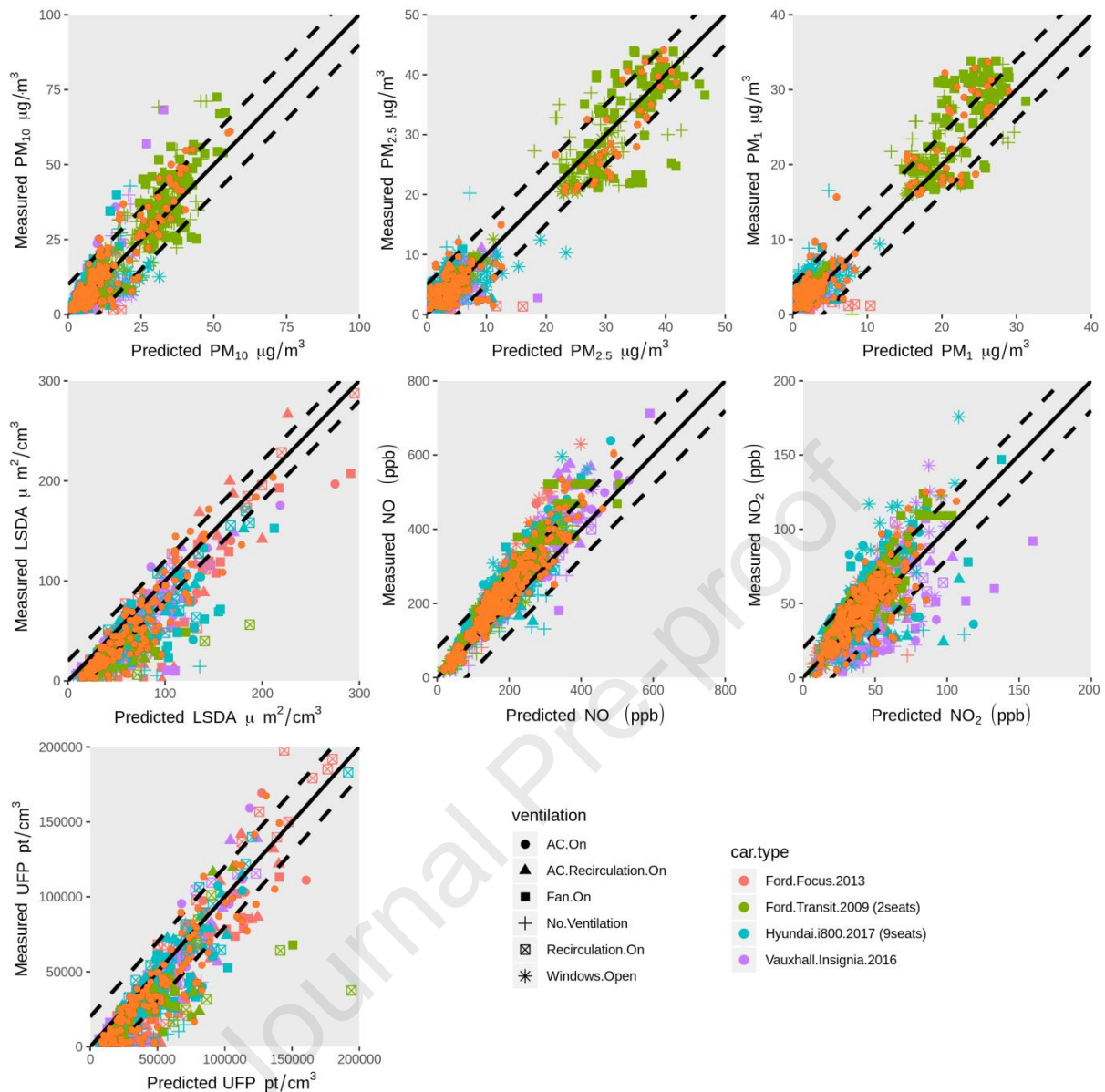
399



400

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

403



404

405 Figure 4. Measured vs MB and ML model within-vehicle concentrations of PM₁₀, PM_{2.5}, PM₁, LSDA, NO,
 406 NO₂ and UFP. The orange dots indicate the ML predictions for 20% of randomly excluded data. The
 407 solid line denotes the perfect model 1:1. The dashed lines indicate the ±10% of the perfect model.

408

409 3.2.2 Machine Learning (ML) model predictions

410 The machine learning (ML) model training method (80:20) is by definition expected to yield
 411 generally good predictions. In Figure 4 the orange dots also show the comparison between the
 412 observed and the ML modelled values for the 20% of measurements excluded from the training
 413 dataset. The ML model shows similar performance to the MB model and in some cases, such as for
 414 NO₂, it improves upon the MB model predictions. Most of the ML model predictions in almost all the

415 species are equally spread around the 1:1 line, however, an under-prediction still occurs in the LSDA
416 and UFP species.

417 Table 8 summarises the ML and MB model performance statistics against the observations
418 (20% of withheld data in the ML case) respectively. It can be seen both models show good skill in
419 predicting within-vehicle concentrations for all species. Pearson correlation coefficients for the ML
420 model between ML predicted and observed values are higher than 0.80, while an IOA (index of
421 agreement) is greater than 0.69 for all the species. For the MB model, the two indices between MB
422 predicted and observed concentrations were slightly worse, varying between 0.45 – 0.82 and 0.48 -
423 0.83 for Pearson correlation coefficient and IOA respectively. However, values of IOA greater than 0.5
424 in general indicate good model predictions (Hurley et al., 2005; Matthaios et al., 2017). The mean
425 gross error (MGE) of the ML and MB model's performance was less than 2.4 and 3.4 $\mu\text{g m}^{-3}$ respectively
426 for all the particle classes (PM₁₀, PM_{2.5} and PM₁) and 10.4 and 14.1 *ppb* for NO₂. The biggest error is
427 evidenced in NO and UFP, which is almost the same as the mean bias. The model's fraction of
428 predictions within a factor of two of observations (FAC2) is also in good agreement with observations
429 for the ML model (higher than 0.66 for all the species), while noteworthy is the fact the ML model's
430 FAC2 score is very high (0.89) for NO₂. For the MB model the FAC2 factor shows low prediction values
431 for LSDA and UFP. NO had FAC2 greater than 1 values which indicates overprediction. The mean bias
432 indicates that the ML model under-predicts the particulate species by less than $< 1 \mu\text{g m}^{-3}$ and the NO₂
433 by less than $< 5 \text{ ppb}$, while slightly greater mean bias for these species is observed for the MB model.
434 The biggest under-prediction occurs for UFP and NO. For NO the ML has a mean underprediction of
435 26 *ppb* while the MB model has a mean overprediction of 35.4 *ppb*. Events such as overtaking or
436 congestion that can result in greater NO outside and consequently inside, and particle leaks from the
437 engine or generation of already deposited particles (in the seats or fabrics) due to vibration or
438 movement cannot be captured in the MB model and can generate tails and cause skewness in the
439 data. kNN algorithms are known to suffer from skewed distributions if those observations are very
440 frequent in the data (Aha et al., 1991). Overall it can be stated that both MB and ML models showed
441 good skill in predicting the measurement data however better predictions were observed in the ML
442 model most likely due to the way the algorithm incorporates the data. The fact that ML improves the
443 model's performance was also found in other studies (Ozcift and Gulten, 2011; Aquilina et al., 2018).

444

445

446

Journal Pre-proof

448 Table 8. Model evaluation statistics against 20% random observation data after the machine learning approach. n: indicates the number of compared values.
 449 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
 450 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
 451 – 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:
 452 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

453

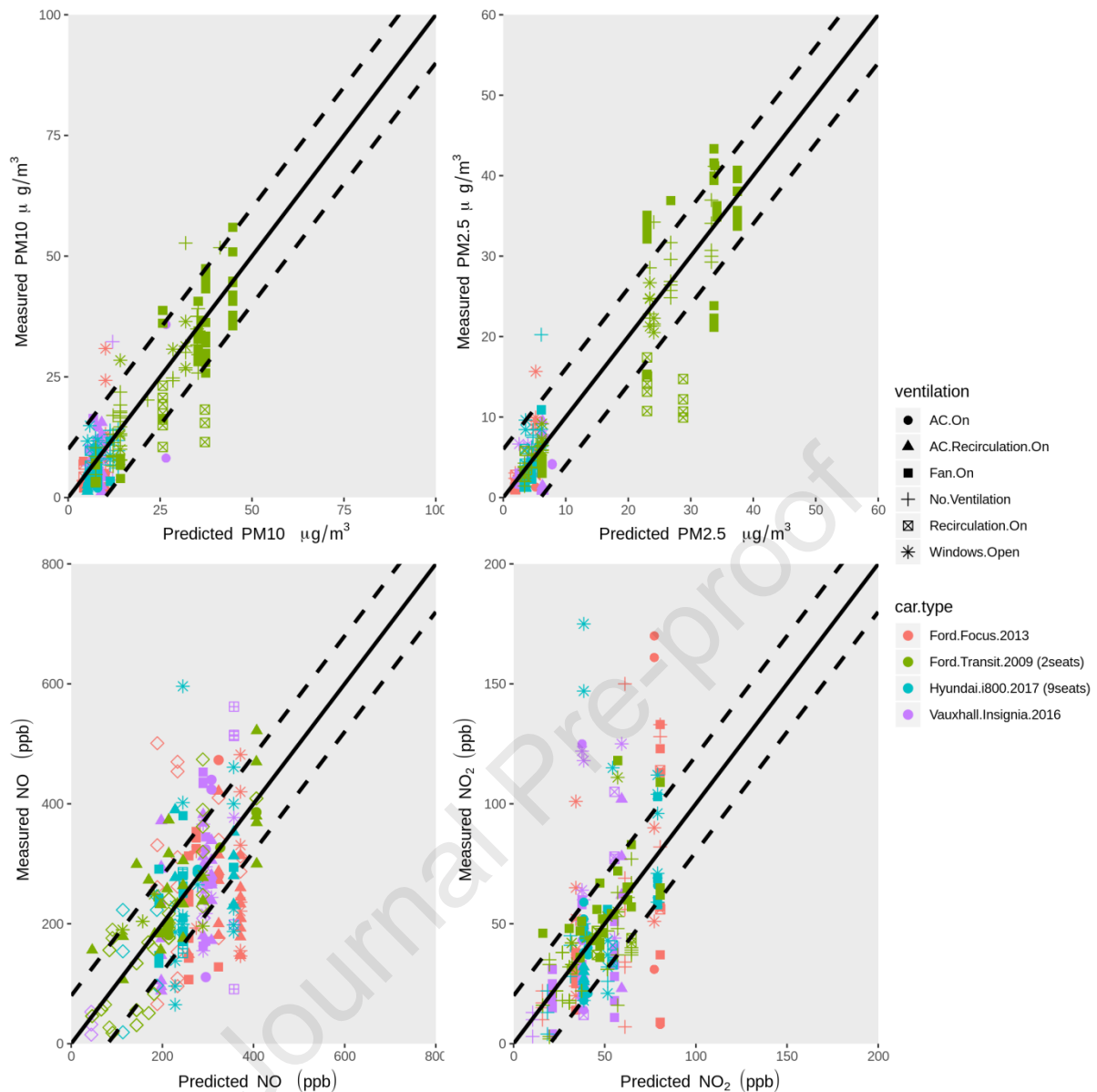
454

455 3.2.3 Extended application of ML model using data from monitoring stations

456

457 In the predictions discussed above, each model utilized external concentrations of air
458 pollutants measured directly outside the study vehicle, to either to drive the calculated pollutant
459 exchange (MB model), or as input for the ML model. However, in order to explore the ML model's
460 potential wider application under real world circumstances we explored case (ii) where both within
461 and directly-outside vehicle pollutant concentrations are unknown and the only data available is from
462 nearby air quality monitoring stations (see 2.5). In this case, the ML model used a median within-
463 vehicle level from all vehicles and hourly outdoor air quality measurements. The air quality levels from
464 the monitoring sites were taken from urban-traffic locations representing different locations of the
465 testing route. Figure 6 shows the case (ii) comparison of the ML model predicted within-vehicle
466 pollutant concentrations, vs those measured. The results generally show some notable discrepancies
467 for NO greater than 260 *ppb* and NO₂ greater than 60 *ppb*, of the within-vehicle air quality for a given
468 air quality value, however the applicability of the method provides an indication of within-vehicle
469 exposure without the need for directly-outside measurement. The ML predictions would have been
470 more representative of the actual exposures in case where more information of the accurate
471 representation of the ventilation system, filtration and air exchange, vehicle number and fleet
472 composition were available.

473



474

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

478

479 4 Comparison with other studies and limitations

480 The study investigated in-vehicle air pollution exposure with novel complementary modelling
 481 techniques using mass balance and machine learning approaches. Studies that used ML algorithms to
 482 predict in-vehicle air quality typically used low-cost sensors to calculate an air quality index that
 483 involved CO₂ and PM_{2.5} and tested the performance of supervised ML algorithms against traditional
 484 regression techniques and deep-learning techniques (Sukor et al., 2022; Goh et al., 2021). Similarly,

485 Lohani et al., 2022 compared traditional auto-regressive integrated moving average (ARIMA) and ML
486 support vector regression (SVR) to investigate their performance against in-vehicle CO₂ levels. Chung
487 and Kim, (2020), developed an anomaly detection system inside cars based on ML algorithms to
488 prevent fatigue and drowsiness due to CO₂ and reduction in PM_{2.5} exposures. Baldi et al., (2022),
489 measured the performance of several ML algorithms against observations of PM₁₀, PM_{2.5}, PM₁, CO₂
490 and formaldehyde and found good results. Our study, apart from the application of ML to predict in-
491 vehicle exposures, it offered novel expansion upon real-world applications with the implementation
492 of air quality data from nearby monitoring sites. Several MB models have been reported for the
493 prediction of within-vehicle concentrations of air pollutants, albeit focusing on different aspects of the
494 problem, for example the models of Hudda et al., (2012); Knibbs et al., (2010) or Xu and Zhu, (2009).
495 The model developed by Hudda et al., 2012, used measured data from a large number of vehicles and
496 multi linear regression approaches and generalized estimating equations to estimate within-vehicle
497 concentrations of UFP, while the models of Knibbs et al., (2010) and Xu and Zhu, (2009) are mass-
498 balance based models. The differential equations applied in this work build on the mass balance
499 studies of Knibbs et al., 2010 and Xu and Zhu, (2009), with some modifications in the equations,
500 including incorporation of key aspects of chemical processing. The reason for the difference in some
501 modelled vs observed levels is likely due to values such as deposition coefficients, filtration efficiency
502 and penetration factors were taken from literature and often from experiments conducted in houses
503 which are larger volumes than vehicle cabins and do not reflect actual within-vehicle values. Another
504 reason might be due to our simplified approach of not having a speed dependent pressure difference
505 penetration factor. As highlighted in Lee et al., (2015a), those factors depend on the combined effects
506 of the ventilation conditions (i.e., ventilation mode and fan settings) and the aerodynamic changes on
507 the vehicle envelope (i.e., driving speed and vehicle shapes) which have not yet been incorporated in
508 this model. It should be further noted that the importance of physical air exchange processes of the
509 outside measurements often dominate comparing to the other indoor sinks and when a rapid change
510 of the outdoor concentrations (i.e. vehicle overtaking, high emitters etc) occurs it has implications for
511 the modelling of within-vehicle NO and NO₂. This is likely the reason that the model underpredicts the
512 high levels of within-vehicle NO₂.

513 The current MB model and methodology likely has limitations in the prediction of other more
514 reactive species within-vehicles, where chemical processing is more important (relatively) to ingress
515 and deposition and needs to be considered for those species; this also implies a more sophisticated
516 treatment of physical conditions (including photolysis frequencies). The MB model assumes a well-
517 mixed (within-vehicle) microenvironment, which may not reflect reality. Furthermore, the MB and ML
518 models are dependent upon the initial parameters (e.g. vehicle characteristics, fan power and other

519 within-vehicle parameters to build the model) and therefore they might be case-dependent and their
520 applicability needs to be tested in other cases. In the model the leakage rate/passive ventilation was
521 calculated using the equations of Hudda et al., (2012). However, since that method uses generalized
522 regression models based on vehicle age, driving speed, and fan strength, the method may impose
523 uncertainty across different vehicle models and other approaches to calculate the leakage
524 flow/passive ventilation, for example based on the pressure difference (Lee et al.,2015b), or using an
525 explicit CO₂ tracer, may be tested for suitability. Engine/fuel leaks can generate gaseous and
526 particulate pollution and other organic gas compounds such as, benzene, toluene, xylene, and methyl-
527 tertiary butyl ether (Faber et al., 2013; Fedoruk and Kerger, 2003; Jo and Park, 1998; Duffy and Nelson,
528 1997) that can enter the interior of the vehicles via the ventilation system. This source is not currently
529 included in the model of this study. Finally, carcinogenic/toxic species such as volatile organic
530 compounds which are released from plastics and fabrics on exposure to sunlight and heat (Yoshida
531 and Matsunaga, 2006, You et al., 2007) and heterogeneous surface reactions or reactions of peroxy
532 radicals with NO, can play a role in the within-vehicle chemistry and improve NO₂ predictions. The
533 model currently is limited in omitting representation of such detailed chemistry, secondary aerosol
534 formation and other particle physics processes.

535 **5 Implications**

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

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

544 Future work should focus on developing more comprehensive exposure predictive models for
545 car passengers. These models will need to account for various driving conditions (e.g., urban and
546 motorway driving), driving durations, passenger characteristics (e.g., differing breathing rates,
547 metabolism, sex, weight), and pathways for pollutant infiltration and penetration, including the
548 assessment of potential in-cabin sources like engine leaks. Such information will be critical for the
549 application of air quality management policies and new technologies such as within-vehicle air
550 purifiers or high selectivity air cabin filters to reduce air pollution exposure. In conclusion, our study
551 presents a novel method to predict within-vehicle air pollution exposure, which has far-reaching
552 implications for public health and environmental research. The study has successfully demonstrated

553 the effectiveness of the approach in providing real-time exposure estimates and mapping. We believe
554 that this work serves as a foundational contribution to the field of real-time air pollution exposure
555 assessment, offering a path towards cleaner and healthier urban environments. While our study is a
556 significant step forward, we acknowledge that further research is essential to refine our approach and
557 enhance its accuracy.

558

559 **Data availability**

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

562

563 **Acknowledgments**

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

569

570 **References**

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

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

577 Aha, D. W., Kibler, D., Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1),
578 37–66. doi:10.1007/bf00153759

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

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

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

- 586 Benzene. *Environmental Science & Technology* 2018 52 (19), 11215-11222. DOI:
587 10.1021/acs.est.8b03328
- 588 Baldi, T., Delnevo, G., Girau, R. and Mirri, S., 2022, August. On the prediction of air quality within
589 vehicles using outdoor air pollution: sensors and machine learning algorithms. In *Proceedings of the*
590 *ACM SIGCOMM Workshop on Networked Sensing Systems for a Sustainable Society* (pp. 14-19).
- 591 Carslaw N., 2007. A new detailed chemical model for indoor air pollution. *Atmos. Environ*, 41 (6),
592 1164–1179.
- 593 Carslaw, D.C. and K. Ropkins, (2012). *openair* — an R package for air quality data analysis.
594 *Environmental Modelling & Software*. Volume 27-28, pp. 52-61.
- 595 Carslaw, D.C. (2019). *The openair manual — open-source tools for analysing air pollution data*. Manual
596 for version 2.6-6, University of York.
- 597 Chen, C., Zhao, B., 2011. Review of relationship between indoor and outdoor particles: I/O ratio,
598 infiltration factor and penetration factor. *Atmos. Environ.* 45, 275–288.
599 <https://doi.org/10.1016/j.atmosenv.2010.09.048>.
- 600 Chung, J.J. and Kim, H.J., 2020. An automobile environment detection system based on deep neural
601 network and its implementation using IoT-enabled in-vehicle air quality sensors. *Sustainability*, 12(6),
602 p.2475.
- 603 Cover T., Hart P., 1967. Nearest neighbor pattern classification. In *IEEE Transactions in Information*
604 *Theory*, IT-13, pages 21–27.
- 605 Delgado-Saborit, J.M., 2012. Use of real-time sensors to characterise human exposures to combustion
606 related pollutants. *J. Environ. Monit.* 14, 1824–1837.
- 607 Delfino, R.J., Malik, S., Sioutas, C., 2005. Potential role of ultrafine particles in associations between
608 airborne particle mass and cardiovascular health. *Environmental Health Perspectives* 113 (8), 934-946.
- 609 De Hartog, J.J., Ayres, J., Karakatsani, A., Analitis, A., ten Brink, H., Hameri, K., Harrison, R.,
610 Katsouyanni, K., Kotronarou, A., Kavouras, I., Meddings, C., Pekkanen, J., Hoek, G., 2010. Lung function
611 and indicators of exposure to indoor and outdoor particulate matter among asthma and COPD
612 patients. *Occupational and Environmental Medicine* 67, 2-10.
- 613 DfT, Department for Transport, National Travel Survey: England 2016, 2017, July
614 2017 <https://www.gov.uk/government/statistics/national-travel-survey-2016>. (accessed July 2019)
- 615 Dons, E., Int Panis, L., Van Poppel, M., Theunis, J., Willems, H., Torfs, R., Wets, G., 2011. Impact of
616 time–activity patterns on personal exposure to black carbon. *Atmos. Environ.* 45, 3594–3602.
617 <https://doi.org/10.1016/j.atmosenv.2011.03.064>.
- 618 Frederickson LB, Lim S, Russell HS, Kwiatkowski S, Bonomaully J, Schmidt JA, Hertel O, Mudway I,
619 Barratt B, Johnson MS. Monitoring Excess Exposure to Air Pollution for Professional Drivers in London
620 Using Low-Cost Sensors. *Atmosphere*. 2020; 11(7):749. <https://doi.org/10.3390/atmos11070749>
- 621 Fruin S. A., Hudda N., Sioutas C., Delfino R. J., 2011. Predictive Model for Vehicle Air Exchange Rates
622 Based on a Large, Representative Sample. *Environmental Science and Technology*, Vol. 45, pp. 3,569-
623 3,575.

- 624 Goh, C.C., Kamarudin, L.M., Zakaria, A., Nishizaki, H., Ramli, N., Mao, X., Syed Zakaria, S.M.M.,
625 Kanagaraj, E., Abdull Sukor, A.S. and Elham, M.F., 2021. Real-time in-vehicle air quality monitoring
626 system using machine learning prediction algorithm. *Sensors*, 21(15), p.4956.
- 627 Gong L., Xu B., Zhu Y., 2009. Ultrafine particles deposition inside passenger vehicles. *Aerosol Sci.*
628 *Technol.*, 43, 544–553
- 629 Hachem, M., Saleh, N., Bensefa-Colas, L. and Momas, I. (2021), Determinants of ultrafine particles,
630 black carbon, nitrogen dioxide, and carbon monoxide concentrations inside vehicles in the Paris area:
631 PUF-TAXI study. *Indoor Air*, 31: 848- 859. <https://doi.org/10.1111/ina.12779>
- 632 Hamra, G. B., Laden, F., Cohen, A. J., Raaschou-Nielsen, O., Brauer, M., Loomis, D. 2015. Lung cancer
633 and exposure to nitrogen dioxide and traffic: A systematic review and meta-analysis. *Environmental*
634 *Health Perspectives*, 123 (11), 1107–1112. <https://doi.org/10.1289/ehp.1408882>
- 635 Harrison, R.M., 2018. Urban atmospheric chemistry: a very special case for study. *npjClim. Atmos. Sci.*
636 1, 5. <https://doi.org/10.1038/s41612-017-0010-8>.
- 637 Heal M. R., Kumar P., Harrison R. M, 2012. Particles, air quality, policy and health. *Chem Soc Rev*
638 41:6606–30.
- 639 Hinds W. C., 1999. *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles.*
640 Wiley, New York.
- 641 Hudda, N., Eckel, S.P., Knibbs, L.D., Sioutas, C., Delfino, R.J., Fruin, S.A., 2012. Linking in-vehicle
642 ultrafine particle exposures to on-road concentrations. *Atmos. Environ.* 59,578–586.
643 <https://doi.org/10.1016/j.atmosenv.2012.05.021>.
- 644 IARC, (2014). Diesel and Gasoline engine exhausts and some nitroarenes. Volume 105 IARC
645 monographs on the evaluation of carcinogenic risks to humans. [https://monographs.iarc.fr/wp-](https://monographs.iarc.fr/wp-content/uploads/2018/06/mono105.pdf)
646 [content/uploads/2018/06/mono105.pdf](https://monographs.iarc.fr/wp-content/uploads/2018/06/mono105.pdf) (accessed June 2020)
- 647 Kotsiantis S. B., 2007. Supervised machine learning: A review of classification techniques. *Informatica*,
648 31, 249–268.
- 649 Knibbs, L.D., de Dear, R.J., Morawska, L., 2010. Effect of cabin ventilation rate on ultrafineparticle
650 exposure inside automobiles. *Environ. Sci. Technol.* 44, 3546–3551.
651 <https://doi.org/10.1021/es9038209>.
- 652 Knibbs, L. D.; de Dear, R. J.; Atkinson, S. E. Field study of air change and flow rate in six automobiles.
653 *Indoor Air* 2009, 303–313.
- 654 Kumar, P., Hama, S., Nogueira, T., Abbass, R.A., Brand, V.S., de Fatima Andrade, M., Asfaw, A., Aziz,
655 K.H., Cao, S.J., El-Gendy, A. and Islam, S., 2021. In-car particulate matter exposure across ten global
656 cities. *Science of the total environment*, 750, p.141395.
- 657 Kumar, P., Rivas, I., Singh, A.P., Ganesh, V.J., Ananya, M. and Frey, H.C., 2018. Dynamics of coarse and
658 fine particle exposure in transport microenvironments. *NPJ climate and atmospheric science*, 1(1),
659 p.11.
- 660 Lawin H., Fanou L. A., Hinson A. V., Stolbrink M., Houngebegnon P., Kedote N. M., Fayomi B., Kagima J.,
661 Katoto P., Ouendo E. M. D., Mortimer K., 2018. Health Risks Associated with Occupational Exposure
662 to Ambient Air Pollution in Commercial Drivers: A Systematic Review. *Int. J. Environ. Res. Public Health*,
663 15, 2039; doi:10.3390/ijerph15092039.

- 664 Lee E. S., Stenstrom M. K., Zhu Y.F., 2015a. Ultrafine particles infiltration into passenger vehicles Part
665 I: experimental evidences. *Transp. Res. Part D.: Transp. Environ.* 38, 156–165.
- 666 Lee E. S., Stenstrom M. K., Zhu Y.F., 2015b. Ultrafine particle infiltration into passenger vehicles, Part
667 II: model analysis. *Transp Res D.*; 38: 144–155.
- 668 Lohani, D., Barthwal, A. and Acharya, D., 2022. Modeling vehicle indoor air quality using sensor data
669 analytics. *Journal of Reliable Intelligent Environments*, pp.1-11.
- 670 Shanon Lim, Benjamin Barratt, Lois Holliday, Chris J. Griffiths, Ian S. Mudway, Characterising
671 professional drivers' exposure to traffic-related air pollution: Evidence for reduction strategies from
672 in-vehicle personal exposure monitoring, *Environment International*, Volume 153, 2021, 106532,
673 <https://doi.org/10.1016/j.envint.2021.106532>.
- 674 Madronich, S.: The atmosphere and UV-B radiation at ground level. *Environmental UV Photobiology*,
675 Plenum Press, 1–39, 1993.
- 676 Martin, A.N., Boulter, P.G., Roddis, D., McDonough, L., Patterson, M., Rodriguez del Barco, M., Mattes,
677 A., Knibbs, L.D., 2016. In-vehicle nitrogen dioxide concentrations in road tunnels. *Atmos. Environ.*
678 144:234–248. <http://dx.doi.org/10.1016/j.atmosenv.2016.08.083>
- 679 Mathai, V.; Das, A.; Bailey, J.A.; Breuer, K. Airflows inside passenger cars and implications for airborne
680 disease transmission. *Sci.Adv.* 2021, 7, eabe0166.
- 681 Matthaios V. N., Triantafyllou A. G., Albanis T. A., Sakkas V., Garas S., 2017. Performance and
682 evaluation of a coupled prognostic model TAPM over a mountainous complex terrain industrial area.
683 *Theor Appl Climatol* 132:885–903. <https://doi.org/10.1007/s00704-017-2122-9>.
- 684 Matthaios N. V., Kramer J. L., Sommariva R., Pope D. F., Bloss J. W., 2019. Investigation of vehicle cold
685 starts primary NO₂ emissions from ambient monitoring data in the UK and their implications for urban
686 air quality. *Atmospheric Environment* 199, 402-414, DOI: 10.1016/j.atmosenv.2018.11.
- 687 Matthaios, V. N., Kramer, L. J., Crilley, L. R., Sommariva, R., Pope, F. D., Bloss, W. J., 2020.
688 Quantification of within-vehicle exposure to NO_x and particles: Variation with outside air quality,
689 route choice and ventilation options. *Atmospheric Environment*, 117810.
690 doi:10.1016/j.atmosenv.2020.117810
- 691 Matthaios, V.N., Rooney, D., Harrison, R.M., Koutrakis, P. and Bloss, W.J., (2023a). NO₂ levels inside
692 vehicle cabins with pollen and activated carbon filters: A real world targeted intervention to estimate
693 NO₂ exposure reduction potential. *Science of the Total Environment*, 860, p.160395.
- 694 V.N. Matthaios, R.M. Harrison, P. Koutrakis, W.J. Bloss, (2023b). In-vehicle exposure to NO₂ and PM_{2.5}:
695 A comprehensive assessment of controlling parameters and reduction strategies to minimise personal
696 exposure, *Science of the Total Environment*, <https://doi.org/10.1016/j.scitotenv.2023.165537>
- 697 Nazaroff, W.N., Cass, G.R., 1986. Mathematical modeling of chemically reactive pollutants in indoor
698 air. *Environmental Science and Technology* 20 (9), 924–934.
- 699 Ott W., Klepeis N., Switzer P., 2008. Air change rates of motor vehicles and in-vehicle pollutant
700 concentrations from secondhand smoke. *J. Exposure Sci. Environ. Epidemiol.*, 18, 312–325
- 701 Ozcift, A., Gulten, A., 2011. Classifier Ensemble Construction with Rotation Forest to Improve Medical
702 Diagnosis Performance of Machine Learning Algorithms. *Comput. Methods Programs Biomed.* 552,
703 104 (3), 443–451.

- 704 Postlethwait, E. M., and Bidani, A., 1990. Reactive uptake governs the pulmonary airspace removal of
705 inhaled nitrogen dioxide. *J. Appl. Physiol.* 68:594-603.
- 706 Qi, C.; Stanley, N.; Pui, D. Y. H.; Kuehn, T. H. Laboratory and on-road evaluations of cabin air filters
707 using number and surface area concentration monitors. *Environ. Sci. Technol.* 2008, 42, 4128–4132.
- 708 Song, Y., Liang, J., Lu, J., Zhao, X. (2017). An efficient instance selection algorithm for k nearest
709 neighbor regression. *Neurocomputing*, 251, 26–34. doi:10.1016/j.neucom.2017.04.01
- 710 Sukor, A.S.A.; Cheik, G.C.; Kamarudin, L.M.; Mao, X.; Nishizaki, H.; Zakaria, A.; Syed Zakaria, S.M.M.
711 Predictive Analysis of In-Vehicle Air Quality Monitoring System Using Deep Learning Technique.
712 *Atmosphere* 2022, 13, 1587. <https://doi.org/10.3390/atmos13101587>
- 713 Thatcher, T.L., Lunden, M.M., Revzan, K.L., Sextro, R.G., Brown, N.J., 2003. A concentration rebound
714 method for measuring particle penetration and deposition in the indoor environment. *Aerosol Sci.*
715 *Technol.* 37, 847-864.
- 716 TomTom, 2019. World traffic index, measuring congestion worldwide
717 https://www.tomtom.com/en_gb/trafficindex/list?citySize=LARGE&continent=ALL&country=ALL
- 718 Weinberger K. Q., Blitzer J., Saul L. K., 2006. Distance metric learning for large margin nearest neighbor
719 classification. In *NIPS*. MIT Press 2, 3
- 720 Wettschereck D., Aha D. W., Mohri T., 1997. A Review and Empirical Evaluation of Feature Weighting
721 Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review* 10:1–37.
- 722 Williams, R., Suggs, J., Rea, A., Sheldon, L., Rodes, C., Thornburg, J., 2003. The Research Triangle Park
723 particulate matter panel study: modeling ambient source contribution to personal and residential PM
724 mass concentrations. *Atmos. Environ.* 37, 5365-5378.
- 725 Wilks D. S., 2005. *Statistical Methods in the Atmospheric Sciences*, Volume 91, Second Edition
726 (International Geophysics). 2nd ed. Academic Press (cit. on p. 238).
- 727 Xu B., Zhu Y., 2009. Quantitative analysis of the parameters affecting in-cabin to on-roadway (I/O)
728 ultrafine particle concentration ratios. *Aerosol Sci. Technol.*, 43, 400–410.
- 729 Yamada, H., Hayashi, R., Tonokura, K., 2016. Simultaneous measurements of on road/in-vehicle
730 nanoparticles and NO_x while driving: actual situations, passenger exposure and secondary formations.
731 *Sci. Total Environ.* 563, 944-955.
- 732 Yamamoto, N., Shendell, D.G., Winer, A.M., Zhang, J., 2010. Residential air exchange rates in three
733 major US metropolitan areas: results from the relationship among indoor, outdoor, and personal air
734 study 1999-2001. *Indoor Air* 20, 85-90.
- 735 Yoshida, T., Matsunaga, I., 2006. A case study on identification of airborne organic com-pounds and
736 time courses of their concentrations in the cabin of a new car for private use. *Environ. Int.* 32:58–79.
737 <http://dx.doi.org/10.1016/j.envint.2005.04.009>.
- 738 You, K., Ge, Y., Hu, B., Ning, Z., Zhao, S., Zhang, Y., Xie, P., 2007. Measurement of in-vehicle volatile
739 organic compounds under static conditions. *J. Environ. Sci.* 19:1208–
740 1213. [http://dx.doi.org/10.1016/S1001-0742\(07\)60197-1](http://dx.doi.org/10.1016/S1001-0742(07)60197-1).

741 Zuurbier, M., Hoek, G., Oldenwening, M., Lenters, V., Meliefste, K., van den Hazel, P., Brunekreef, B.,
742 2010. Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel
743 type, and route. *Environ. Health Perspect.* 118, 783–789.

744

745

746

Journal Pre-proof

- 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 NO₂.
- Use of air quality monitoring data provides new capabilities for within-vehicle exposure predictions

Journal Pre-proof

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:

--

Journal Pre-proof