Data-Driven Models for Microscopic Vehicle Emissions

Hajar Hajmohammadi[*,a], Giampiero Marra[b], Benjamin Heydecker[a]

[a] Centre for Transport Studies, Department of Civil, Environment and Geomatics Engineering,

University College London, Gower Street, London, WC1E 6BT, UK

[b] Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK

Abstract

In this paper, a new approach for describing the relationship between tailpipe emissions and vehicle movement variables is presented, called generalized additive model for location, scale and shape (GAMLSS). The dataset for this model is second-by-second emission laboratory measurements, following a real driving cycle that were recorded in urban, suburban and motorway areas of London. The GAMLSS emission model is the model to estimate each of CO_2, CO and NO_x in each second for two different vehicle types (petrol or diesel) using instantaneous speed and acceleration as the explanatory variables. Comparing the results with current emission models indicates substantial improvement in accuracy and quality of estimation by this approach.

Keywords: vehicle emission modelling, GAMLSS approach, air pollution.

[*] Corresponding author, E-mail: Hajar.hosseinabadi.14@ucl.ac.uk

1. Introduction

Traffic management and control policies are generally developed to reduce traffic congestion, while their effects on the air pollution from road transport are not a main consideration. Vehicle emission modelling is an approach to estimate these effects on air quality. By using these models, traffic management policies can be developed considering vehicle emissions as well as other criteria.

Exhaust or tailpipe emission from road transport depends upon many factors that can be described as falling into two groups:

- Vehicle-related factors, such as model, engine size, fuel and catalyst type, mileage and technology level.
- Operational factors, such as driving style, road type and design, and traffic controls.

The focus of this paper is to introduce a predictive microscopic vehicle emission model that considers vehicle-related and operational factors. In this model, the nonlinear and intricate relationship between vehicle movement variables (speed and acceleration) and tailpipe emission (CO_2, CO and NO_x) is described using a data-driven approach. This approach uses spline functions to capture as much information as possible from the explanatory variables without preconception of the form of the relationship as in parametric models. A consequence of this is that the model estimates emissions substantially better than generalized linear models (GLM), especially for the principal noxious emissions such as CO in the petrol and NO_x in the diesel vehicles. These pollutants are more complicated to estimate compared to CO_2 because of the effects of catalyst and are important in the series of Euro standards that have become progressively more stringent since their introduction in 1992.

In this paper, two vehicle types (petrol and diesel) are selected to develop the model. The evaluation process shows that the goodness of fit (BIC) of the proposed model is substantially better than the parametric ones including classified log- polynomial regression and generalized linear model that are the basis of the CADI and GLM emission models, respectively. This improvement is more remarkable for CO estimation in the petrol and NO_x in the diesel vehicles.

The paper is organized as follows: the background of the vehicle emission modelling is discussed in the next section, after which the dataset used for developing the proposed model is described in section 3. Section 4 is devoted to the methodology and model framework and after that the results are presented in section 5. The final section presents discussion and conclusions.

2. Background

A wide range of vehicle emission models have been developed up to now, mainly based on the type, amount and quality of available data. Type of explanatory variables and modelling approach of these models are discussed in this section.

2.1. Explanatory Variables

Burning fuel in the engine provides required power for vehicle movement, and fuel burning is the source of vehicle tailpipe emission. There are many factors from engine to the tailpipe that can affect the amount and type of emitted pollutants. The choice of explanatory variables to describe all these factors and engine mechanisms is a crucial part of a vehicle emission model.

The main explanatory variables for most microscopic emission models stem from the driving cycle or speed profile of the vehicle. These are instantaneous speed, v, and acceleration, a. Another explanatory variable for emission modelling is engine power demand. Usually the

power per unit of mass or vehicle specific power (V) is calculated from v and a to represent the engine power demand:

$$V = \left(\frac{\alpha_1 + \alpha_2 v + \alpha_3 v^2}{M} + a + g \sin\theta \right) v, \quad (1)$$

where coefficients α_1, α_2 and α_3 represent rolling, rotating and drag resistance, respectively. M is the mass, θ is the road gradient and g is the acceleration due to gravity. This definition was first presented in [1] which was found to have strong correlation with vehicle emission.

Examples of using speed and acceleration for emission modelling are MODEM and DGV [2]. These two models are similar in concept, both producing emission maps based on speed and acceleration. MODEM, produced during the European Commission's DRIVE research programme, uses two-dimensional look-up table for a particular vehicle type and pollutant according to the vehicle speed, v, and $v \times a$. The estimated mean rate of emissions is the arithmetic means of all observations in the cells, weighted according to their frequency in the driving cycle. The DGV model follows the same methodology, except the look-up table is formed by speed, v, and acceleration, a.

Vehicle specific power (V) is a key contributing factor in the emission models such as MOVES2004, developed by Koupal et al. [3]. In this model, a unidimensional look-up table for emission (or fuel consumption) is indexed by values of V. The vehicle emission (or fuel consumption) is then estimated as the frequency- weighted arithmetic mean according to the driving cycle.

2.2. Modelling approach

Multiple linear, log-linear, log-polynomial and generalized linear model (GLM) are widely used in current emission models such as VERSIT+ [4], VT-micro [5-7] and EMIT [8].

VERSIT+ was developed by the Netherlands Organisation for Applied Scientific Research. This model uses speed as an explanatory variable and GLM regression approach to estimate emission for levels of local, regional and national for specific vehicle types. While VERSIT+ uses a single model for the relationship between speed and tailpipe emission, other models such as VT-micro and CMEM[1] [9-11] implement classification based on the driving mode: Cruise, Accelerate, Decelerate, Idle (CADI), to improve undifferentiated models. The Virginia Tech microscopic model or VT-micro uses a classified log-polynomial model based on speed, v, and acceleration, a:

$$E = \begin{cases} \exp\left(\sum_{i=0}^{3}\sum_{j=0}^{3}\alpha_{i,j}v^i a^j\right) & (a \geq 0) \\ \exp\left(\sum_{i=0}^{3}\sum_{j=0}^{3}\beta_{i,j}v^i a^j\right) & (a < 0), \end{cases} \quad (2)$$

where E is tailpipe emission, v is vehicle speed with exponent i ($0 \leq i \leq 3$), a is acceleration with exponent j ($0 \leq j \leq 3$), and α and β are the model coefficients for acceleration and deceleration, respectively.

In this classification, deceleration mode is identified by $a < 0$ whilst the other CADI modes are merged into the $a \geq 0$ case. For a complete CADI classification, speed should be involved ($a \approx 0$ and $v > 0$ for cruising and $a = 0$, $v = 0$ for idling). In the CMEM, each CADI operating mode is considered separately to estimate exhaust emission and fuel consumption. This model first estimates engine-out emission (E^o) based on vehicle specific power, V, then the tailpipe emission is calculated as a function of engine-out emission, fuel consumption rate and catalyst pass fraction.

[1] -Comprehensive Modal Emissions Model

EMIssions from Traffic or EMIT is another classified model for estimating vehicle emission and fuel consumption of light duty vehicles [8]. The form of this model for estimating E^o from speed and acceleration is the classified polynomial form:

$$E^o = \begin{cases} \omega_0 + \omega_1 v + \omega_2 v^2 + \omega_3 v^3 + \omega_4 va & (V > 0) \\ \omega_0' & (V = 0) \end{cases} \quad (3)$$

where $\boldsymbol{\omega}$ is the model coefficient are the model coefficients, and V is the vehicle specific power.

This corresponds to classification according to vehicle specific power rather than acceleration. In terms of explanatory variables, the focus of the present study is to identify a simple set of variables based on the driving cycle to estimate tailpipe emission. Separate estimation of engine-out and tailpipe emission such as CMEM and EMIT increases the number of parameters required and adds extra complexity to the model, so is avoided here. This study proposes a model in the form of a single versatile model, using data-driven approach, rather than a parametric classified one such as VT-micro.

3. Dataset

The data used in the present study were generated in laboratory emission tests that were performed by Millbrook Laboratory using a chassis dynamometer, a full-scale dilution tunnel, and exhaust gas analysers. The chassis dynamometer should be capable of simulating the transient inertial load, aerodynamic drag and rolling resistance associated with normal operations of the vehicle [12] within a static laboratory.

3.1. Real London driving cycle

Vehicle emission tests are performed based on a driving cycle or a time series of vehicle speed, that are sometimes intended to represent more or less typical driving patterns. A visual display

of the desired and actual vehicle speed is provided for the driver, so they can follow the prescribed driving cycle. For this study, real driving cycle data were logged second-by-second during three separate drives on certain routes in London (Figure 1) on each of two days to allow for an averaging effect in case of any atypical traffic conditions encountered. The recorded data were then used as the target driving cycles in the laboratory emissions measurements. A total of 9 real driving cycles were defined according to the type of the road (urban, suburban and motorway), and traffic conditions (AM peak, inter peak and free flow) for the emission tests. In Table 1 statistical information of the real driving cycles is summarized. The driving cycle of free flow in urban, suburban and motorway is shown in Figure 2.

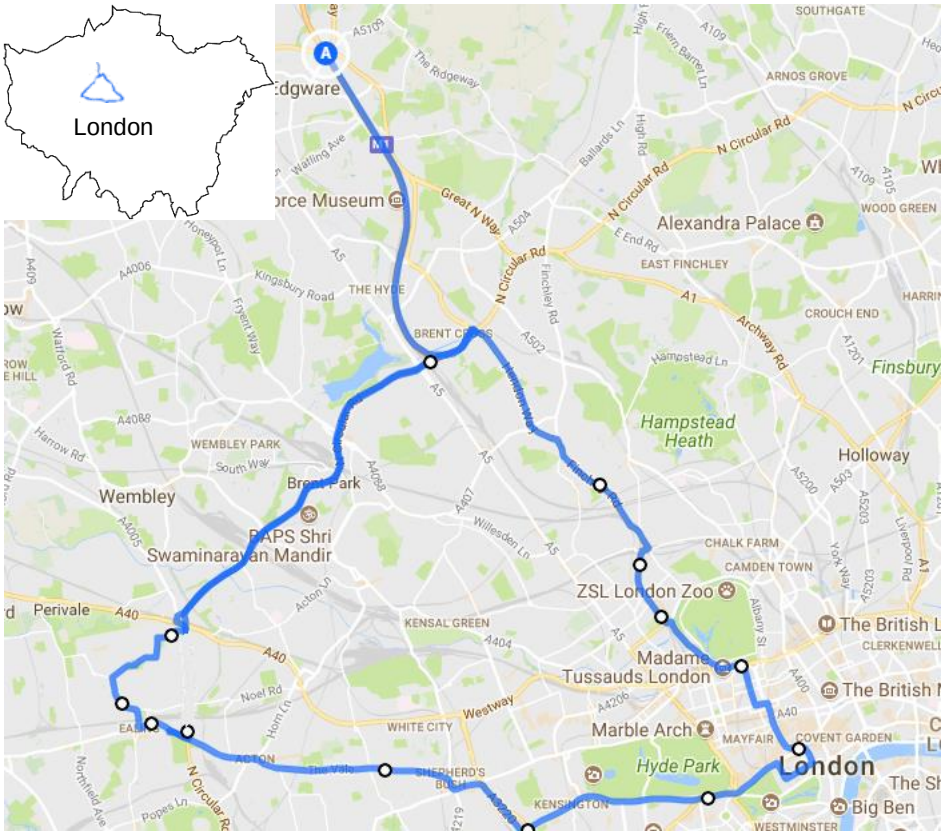


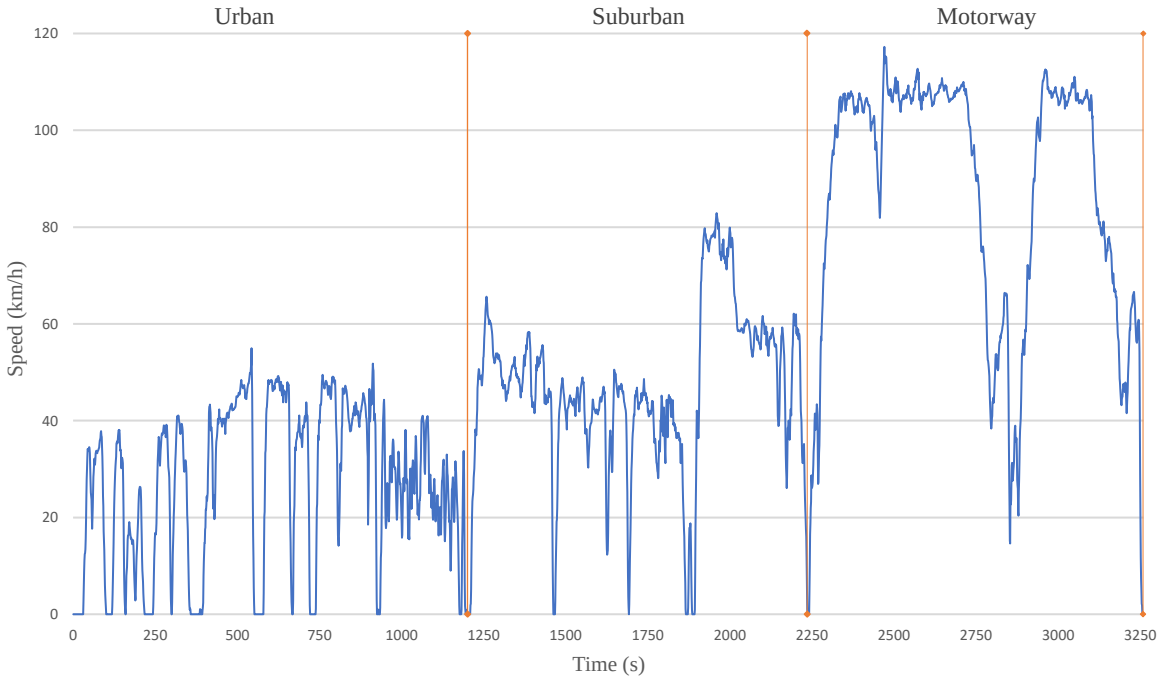
Figure 1: Map illustrating the route for data logging

135

Table 1: statistical information of the real driving cycles

	Average speed (km/h)	Maximum speed (km/h)	Std. Dev. speed (km/h)	Maximum acceleration (m/s^2)	Minimum acceleration (m/s^2)	Std. Dev. acceleration (m/s)	Duration (sec)	Length (km)
Urban								
Free Flow	26.53	52.85	16.24	1.91	-2.87	0.63	1202	8.86
AM peak	15.57	50.79	15.31	1.78	-2.43	0.58	2048	
Inter peak	13.80	49.48	14.58	2.36	-2.85	0.56	2310	
Suburban								
Free Flow	46.04	82.53	16.94	2.06	-2.92	0.49	1036	13.25
AM peak	25.15	78.42	22.06	2.36	-2.62	0.57	1867	
Inter peak	30.13	80.65	23.27	2.1	-2.64	0.57	1597	
Motorway								
Free Flow	86.18	113.32	27.21	1.69	-3.09	0.41	1025	24.54
AM peak	46.84	113.94	39.60	1.93	-2.72	0.51	1884	
Inter peak	85.77	113.03	31.75	2.21	-2.43	0.41	1030	

136



137

138 *Figure 2:real driving cycle for urban, suburban and motorway, free flow traffic conditions*

4. Model framework

The present emission model is a Generalized Additive Model for Location, Scale and Shape (GAMLSS) [13, 14], that is an extended class of generalized additive model (GAM) [15]. An outline of the model building process is shown in Figure 3. The calculations presented here were undertaken using the model as implemented in the GJMR [16, 17] R package (version 2). The R codes used for this study are presented in Appendix. The components of this model are explained in the following.

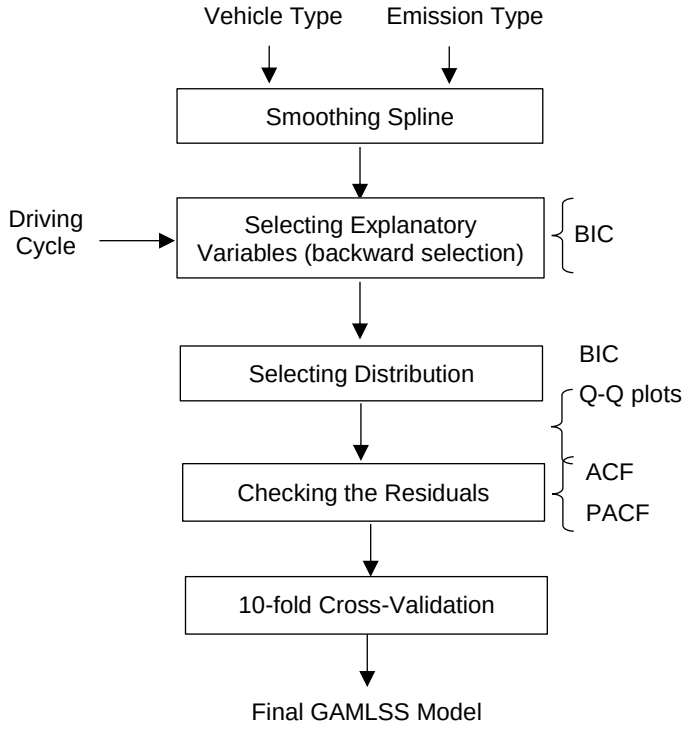


Figure 3: GAMLSS model building process

4.1. Spline smoothing function

A smooth function is adopted to summarize the trend of a response variable in respect of one or more explanatory variables. It is formed piecewise of polynomial (of order M) sections that are $M-1$ times continuously differentiable at the joints, resulting in a smooth function that can follow relationships in the data.

The method of splines, as applied in this paper, consists of dividing up the range of each explanatory variable x into segments with join points at the boundaries which are known as knots. A polynomial of a fixed degree is then fitted to the observed values of the response variable within each segment with constraints applied at the knots to ensure appropriate continuity there. Hence, the spline function $s(x)$ of order M (degree $M-1$) with knots at $x = \xi_1,...\xi_K$ $\left(\text{where } \xi_1 < \xi_2 < ... < \xi_K\right)$ and a domain $[a,b]$, is defined as a function with the following properties:

1) In each of the intervals:

   $a < x < \xi_1, \quad \xi_{j-1} \le x \le \xi_j$ $(j = 2,3,...,K)$, and $\xi_K \le x \le b$, $s(x)$ is a polynomial of degree $M-1$ at most.

2) $s(x)$ and its derivatives up to order $M-2$ are continuous at each of the knots.

The univariate spline of order M, $s_M(.)$, can be represented analytically in the form:

$$s_M(x) = \sum_{h=0}^{M-1} \alpha_h x^h + \sum_{j=1}^{K} \beta_j \left(x - \xi_j\right)_+^{M-1},$$

where $u_+ = \begin{cases} u & (u \ge 0) \\ 0 & (u \le 0). \end{cases}$ (4)

This shows clearly the continuity of derivatives up to $M-2$ at all values of the explanatory variable x, including the knots. However, other representations are better conditions for computation, and so are preferred for that. Duchon [18] extended this approach to one or more dimensions of explanatory variables, which is known as thin plate regression splines (TPRS).

A penalized maximum likelihood criterion is then used for fitting the TPRS, s_{tp}, by solving the joint optimisations:

$$\lambda^* = \arg\min_{\lambda} \; A(\lambda) = -2\mathcal{L}(\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda \mid \mathbf{x}) + 2\, e(\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda, \lambda)$$

where, (5)

$$\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \; B(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{x}) + \lambda \mathrm{J}_p\left(s_{tp}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right)$$

where $\mathcal{L}(.)$ is the log-likelihood of the spline, $\mathrm{J}_p(.)$ is a penalty function for roughness of the spline, and $e(.)$ is the effective degrees of freedom of the spline (Wood, 2017, p269). The first line corresponds to a performance optimisation to determine the value of the smoothing parameter λ that optimises the Akaike information criterion $A(.)$ of the fitted model with effective degrees of freedom e while the second line determines the associated model parameters $\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda$ that optimise the λ-penalised likelihood B.

The penalty function for a TPRS with one explanatory variable and 2 continuous derivatives is defined by:

$$\mathrm{J}_{p,1} = \int_a^b \left[s''_{tp}(x) \right]^2 dx \,. \tag{6}$$

And the penalty function for s_{tp} with two explanatory variables is:

$$\mathrm{J}_{p,2} = \iint \left(\frac{\partial^2 s_{tp}}{\partial x_1^2} \right)^2 + 2\left(\frac{\partial^2 s_{tp}}{\partial x_1^2 \partial x_2^2} \right)^2 + \left(\frac{\partial^2 s_{tp}}{\partial x_2^2} \right)^2 dx_1\, dx_2 \tag{7}$$

The effect of the smoothing parameter is to reduce the curvature of the spline function, with a consequent reduction in the effective degrees of freedom e corresponding to the number of free parameters in the model: without this, use of the maximal set of knots would lead to an underspecified spline function with a potentially high degree of roughness.

4.2. Selecting explanatory variables

The driving cycle is the source of explanatory variables for the present emission models. In order to develop the model, the initial set of explanatory variables was selected as:

$\mathbf{x} = \begin{bmatrix} v & v^2 & v^3 & a & av \end{bmatrix}^T$. These explanatory variables represent the physical concept of vehicle specific power (V) components from equation 1:

- Power to accelerate: av
- Rolling resistance: v
- Rotating resistance: v^2
- Air resistance: v^3

The model initially was run using all of these variables, and then variables were considered for elimination according to the backward selection procedure. The Bayesian Information Criterion (BIC) was used to assess model performance at each stage:

$$BIC = -2\mathcal{L} + \log_e(n)\,p \tag{8}$$

where p is the effective degrees of freedom, corresponding to the number of free parameters in the model, n is the number of observations and $\mathcal{L}$ is the log-likelihood of the model. Models with smaller values of BIC are preferred: this provides a balance between lack of fit (represents by lower log-likelihood) and model complexity (represented by the number p of parameters used). This criterion penalizes additional parameters according to the number n of observations used to fit the model. Whenever $n \geq 8 \approx e^2$ the entry criterion for additional parameters is more stringent than that of Akaike's information criterion $AIC = -2\mathcal{L} + 2p$, thus leading to less complicated models.

By applying this process, the final explanatory variables in the model were selected as: speed (v), acceleration (a) and their product (va). Other variables were eliminated from the model as they could not be justified as efficient in improving the estimation based on the BIC values.

The resulting emission models for two different vehicle types. The characteristics of these vehicles are presented in Table 2.

Table 2: Characteristics of tested vehicles

code	Type	Fuel	Transmission	Engine size (cc)	Mass (kg)
A	Compact	Petrol	Manual	1000	900
B	Supermini	Diesel	Manual	1400	1200

4.3. GAMLSS emission model

GAM models explained the relationship between response and explanatory variables only in terms of smooth functions, rather than the more limited specified parametric regressions, such as Genrelized Linear Model (GLM). The extended class of GAM is introduced by Rigby and Stasinopoulos [19] as the general additive model for location, scale and shape, GAMLSS. the model has three components:

1) Systematic component, that contains smooth functions of explanatory variables
2) Response variable distribution or error structure
3) Link function, that connect the parameters of the response variable distribution to the systematic component

By extension to GAM, in GAMLSS the outcome distribution is not restricted to belong to the exponential family. The systematic part of the GAMLSS is expanded to allow modelling of not only the mean (or location) but also the other parameters (scale and shape) of the response variable distribution. The systematic part of the GAMLSS emission model is the sum of splines of the explanatory variables (hence the epithet “additive”):

$$\eta_{i,j} = S_{tp1_j}(v_i) + S_{tp2_j}(a_i) + S_{tp3_j}(v_i, a_i) \quad (9)$$

where $\eta_{i,j}$ is the systematic component of observation i for estimating parameter j of the statistical distribution.

For each emission type, the GAMLSS emission models were tested by using twelve possible one, two and three-parameter distributions, presented in detail in [14]. The BIC values of the models, as well as the Q-Q (quantile-quantile) plots of normalized quantile residuals were used to assess the goodness of fit of the models according to each of these distributions.

According to this process, two distributions were identified for emission modelling: Fisk and the extended version of that, Dagum. These are defined as follows:

Fisk distribution:

$$f(y;\mu,\sigma)=\frac{\sigma y^{\sigma-1}}{\mu^{\sigma}\left[1+\left(y/\mu\right)^{\sigma}\right]^{2}} \tag{10}$$

for $y>0, \mu>0, \sigma>0$.

For moment k of this distribution to exist, the parameter σ is restricted as $\sigma > k$. Subject to $\sigma > 2$, the first and second moments are:

$$\mathrm{E}(Y)=\frac{\mu\pi/\sigma}{\sin\left(\pi/\sigma\right)} \text{ and } \mathrm{VAR}(Y)=\mu^{2}\left[\frac{2\pi/\sigma}{\sin\left(2\pi/\sigma\right)}-\left(\frac{\pi/\sigma}{\sin\left(\pi/\sigma\right)}\right)\right]^{2}.$$

The Dagum distribution is:

$$f(y;\mu,\sigma,\nu)=\frac{\sigma\nu}{y}\left[\frac{\left(y/\mu\right)^{\sigma\nu}}{\left[1+\left(y/\mu\right)^{\sigma}\right]^{\nu+1}}\right], \tag{11}$$

for $y>0$ and location, scale and shape parameters, respectively $\mu>0, \sigma>0, \nu>0$.

Provided that $\sigma > 2$, the first and second moments are:

$$\mathrm{E}(Y)=\frac{-\mu\,\Gamma\left(\frac{-1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\sigma\,\Gamma(\nu)},\ \mathrm{Var}(Y)=-\left(\frac{\mu}{\sigma}\right)^{2}\left[2\sigma\frac{\Gamma\left(\frac{-2}{\sigma}\right)\Gamma\left(\frac{2}{\sigma}+\nu\right)}{\Gamma(\nu)}+\left(\frac{\Gamma\left(\frac{-1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}\right)^{2}\right].$$

Based on the range of the y in both Fisk and Dagum distribution, the link function for all their parameters is the logarithm. Hence, the GAMLSS models for each of the different emittants have distributions with parameters defined by:

for the Fisk distribution model:

$$\mu_i=\exp\eta_{i,1},\ \sigma_i=\exp\eta_{i,2} \tag{12}$$

and for the Dagum distribution model:

$$\mu_i=\exp\eta_{i,1},\ \sigma_i=\exp\eta_{i,2},\ \nu_i=\exp\eta_{i,3}. \tag{13}$$

5. Results

The results of estimation, evaluation and cross validation of the GAMLSS emission model is presented in this section. The results are presented for each vehicle type (A and B) and emission type (CO_2, CO and NO_x) separately.

5.1. Fitted distributions

According to the process described in section 4.3, twelve possible distributions were fitted to the emission and the best distribution selected according to the BIC values and Q-Q plots of normalized quantile residuals (Table 3). The Q-Q plots of fitted distribution are presented in Figure 4.

In these plots, sample quantiles are the emission observation (CO_2, CO and NO_x) and the theoretical quantiles are calculated from the selected distributions. The reference red line indicates the case in which the both quantiles are from the same distribution.

Table 3: Results of distribution selecting for each vehicle type and emission type

	Selected Distribution		
	CO_2	CO	NO_x
Vehicle A	Fisk	Fisk	Dagum
Vehicle B	Dagum	Dagum	Fisk

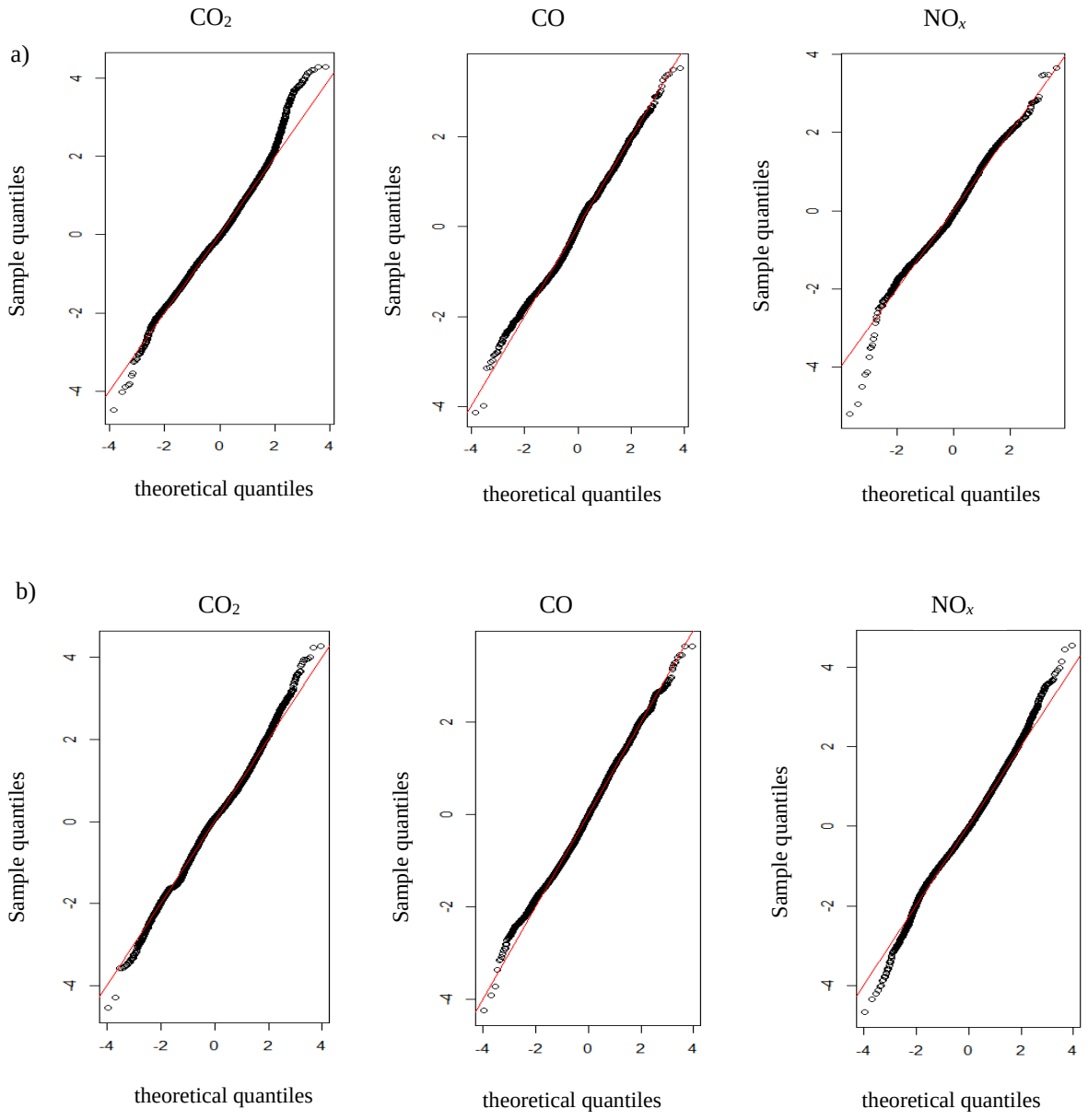


Figure 4 : analysis of quantile residuals for a) vehicle 1 and b) vehicle 2

5.2. Estimation

For each vehicle and emission type, GAMLSS approach (models (12) and (13)) were applied to the dataset. The results are presented in Figure 5- Figure 10.

Plot (a) in these figures is the smooth function of speed, $S_{tp1}(v)$, plot (b) is the smooth function of acceleration, $S_{tp2}(a)$, and plot (c) is the interaction of speed and acceleration, $S_{tp3}(v \times a)$.

The shape of smooth functions for petrol (A) and diesel (B) vehicle types are different in most cases. In vehicle A, $S_{tp1}(v)$ approximately has a constant positive slop particularly when $v \geq 60$ km/h for CO_2. This increasing trend is repeated for CO, with a constant positive slop for $10 \leq v \leq 40$ km/h. For NO_x, the smooth function of speed has approximately no effect on the emission for $20 \leq v \leq 60$ km/h, and after that it has an increasing trend with a limited slop. The reason of that comes from low variation in the NO_x values in the petrol vehicle.

The smooth function of speed for vehicle B can be divided into four categories. Decreasing trend for $v \leq 20$ km/h, increasing trend for $20 \leq v \leq 40$ km/h, decreasing trend for $40 \leq v \leq 60$ and increasing trend for $v \geq 60$. These fluctuations for speed is approximately repeated for CO_2, CO and NO_x, with little differences in CO.

The smooth function of acceleration, $S_{tp2}(a)$, has a changing point close to the $a \approx 0$ m/s^2 for some of the emission types. That could be interpreted as the different effects of $S_{tp2}(a)$ on the emission in acceleration and deceleration driving modes. Smooth functions of acceleration for CO in vehicle B and NO_x in vehicle A have limited effects on the emission, due to the low variations of CO and NO_x in the diesel and petrol vehicles, respectively. For other emission types, the increasing trend when $a < 0$ m/s^2 and then decreasing trend when a is positive can be observed for vehicle B.

334 The highest effect of interaction between speed and acceleration on CO_2 and CO for vehicle A
335 is when $40 \le v \le 60$ km/h and $-3.5 \le a \le 2.5$ m/s^2 (medium speed and harsh deceleration). For
336 NO_x, it is when the speed is low ($v \approx 20$ km/h) and $-2 \le a \le -1$ m/s^2 (deceleration). For vehicle
337 B, the highest impacts of interaction on CO and NO_x is when $v \le 20$ km/h (low speed) and
338 $a \approx -3$ m/s^2 (harsh deceleration). For CO_2 of this vehicle, the interaction effects is high when
339 $40 \le v \le 80$ km/h (medium speed) and $a \le -1$ m/s^2 (deceleration).

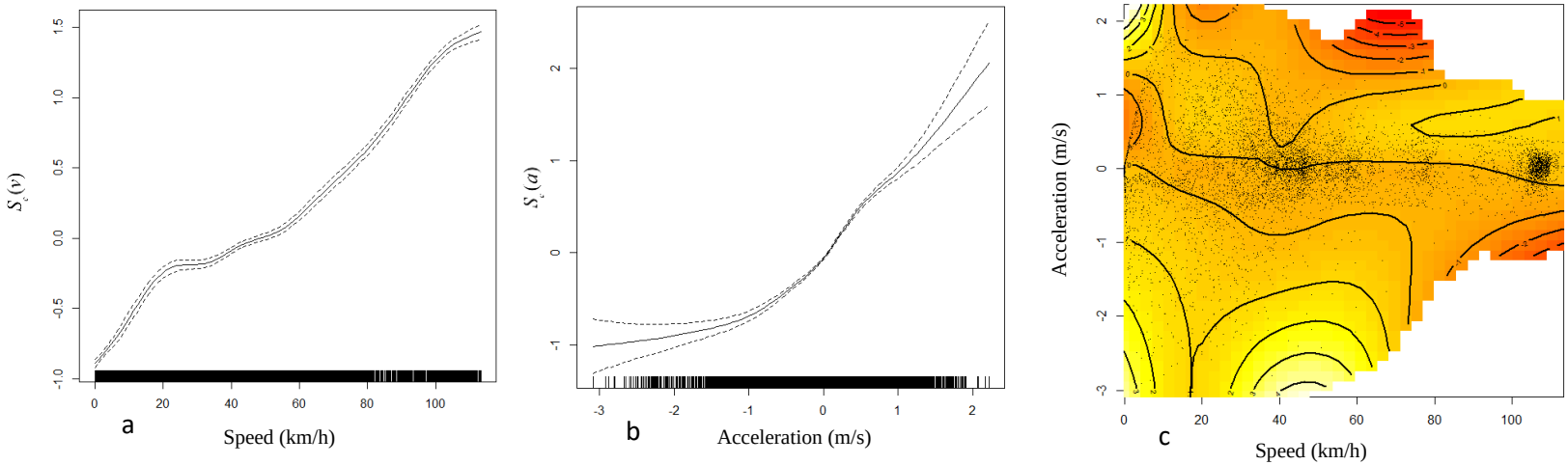


Figure 5: CO_2 , vehicle A: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

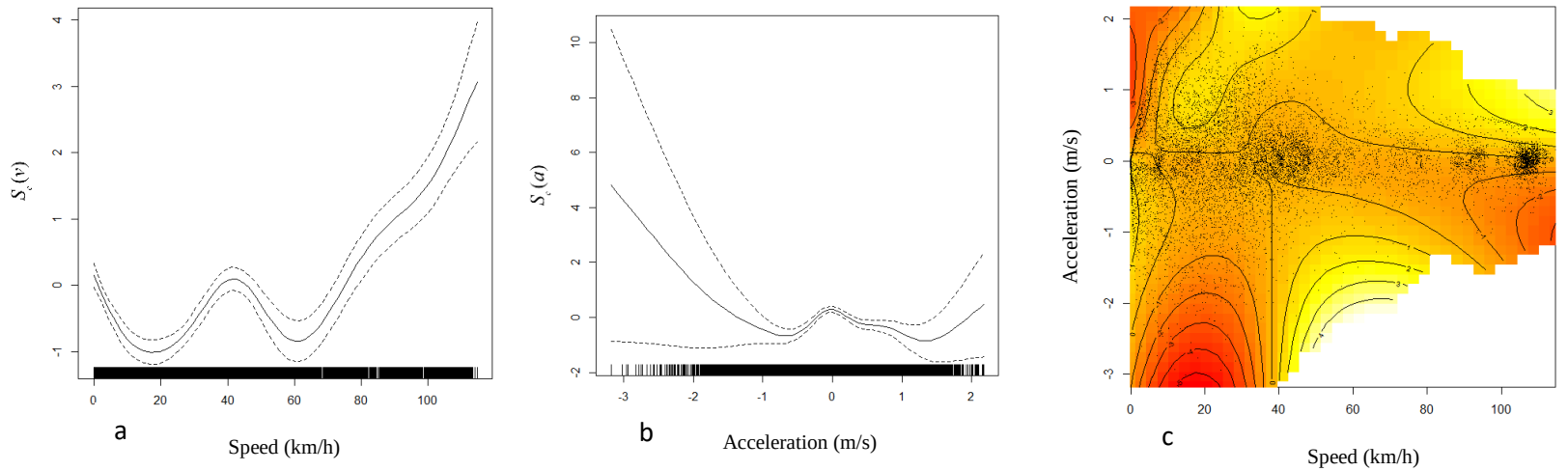


Figure 6: CO_2 , vehicle B: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

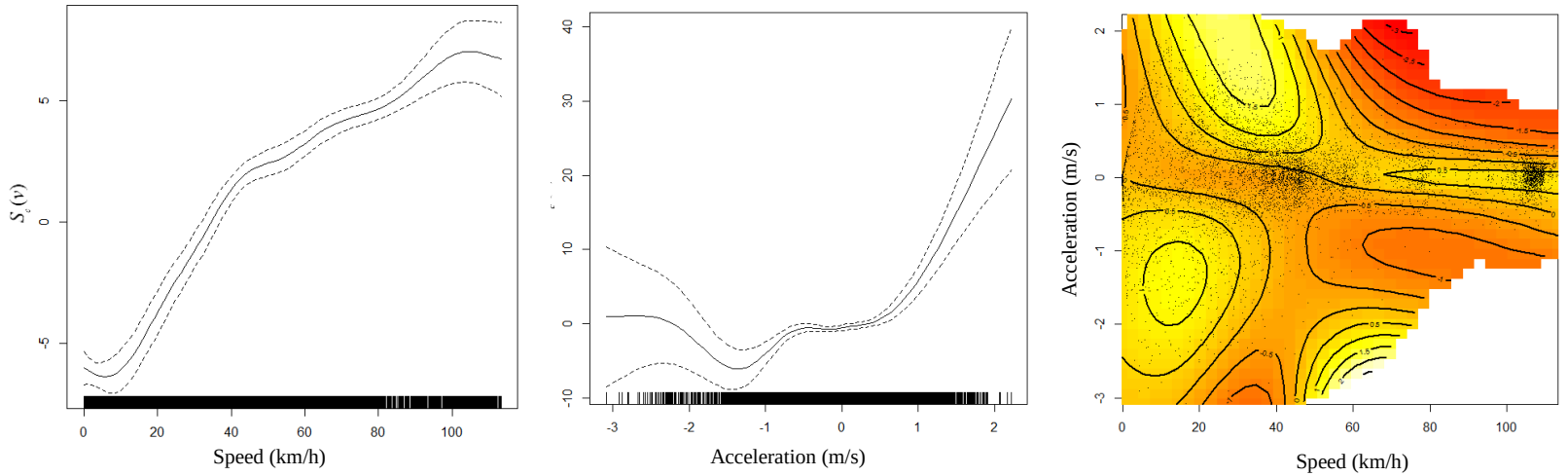


Figure 7: CO, vehicle A: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

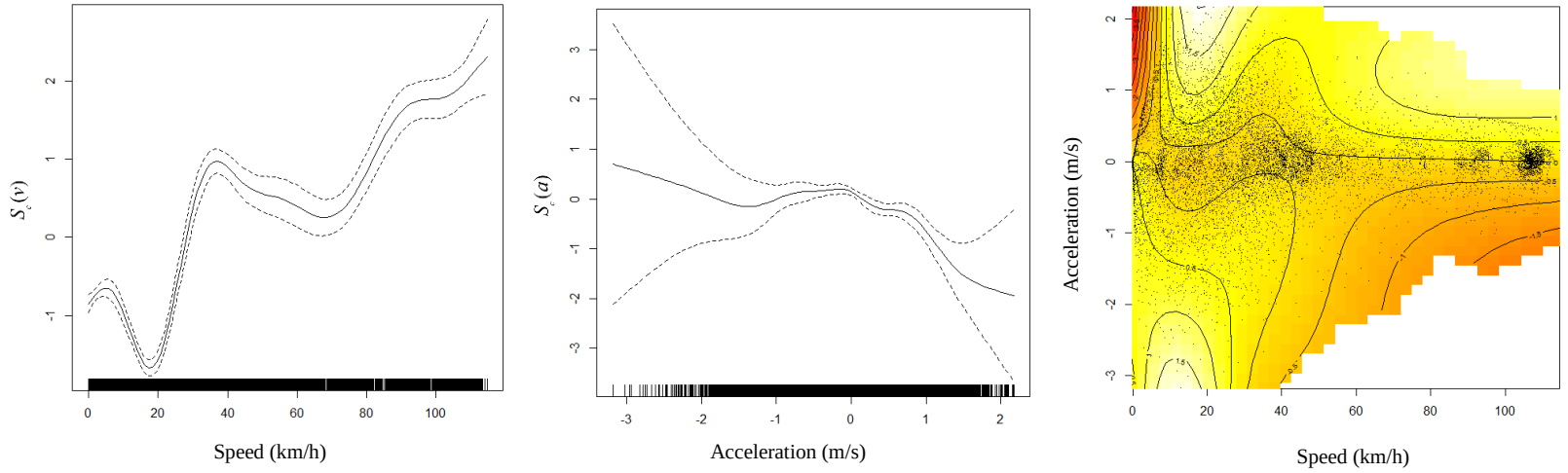


Figure 8: CO, vehicle B: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

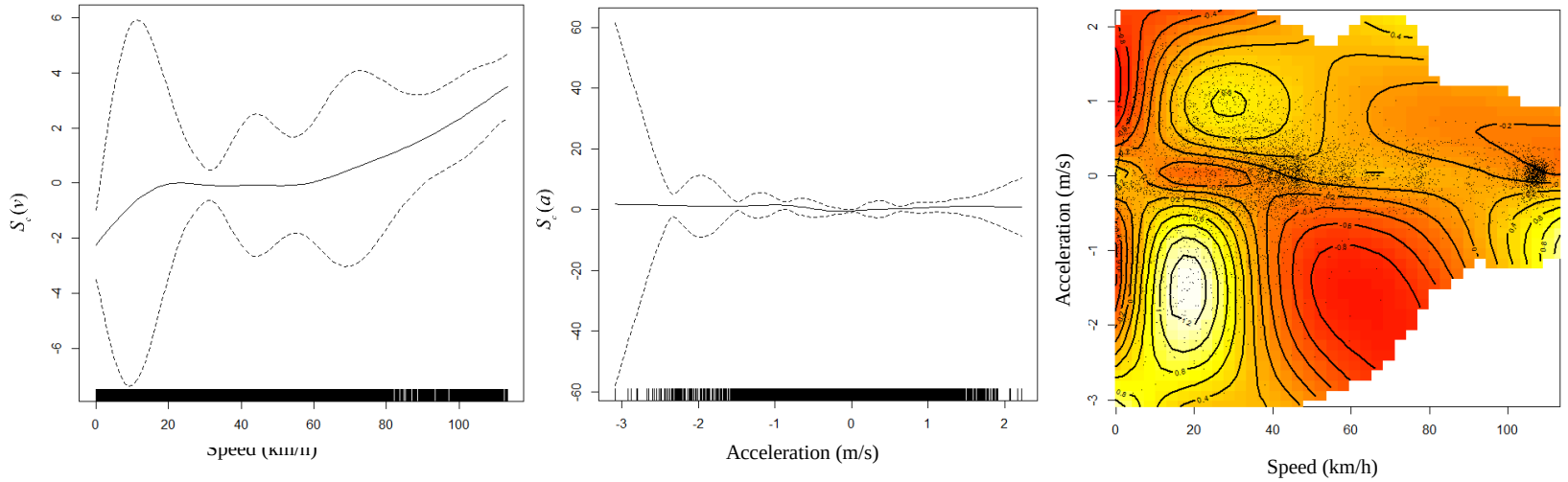


Figure 9: NO_x, vehicle A: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

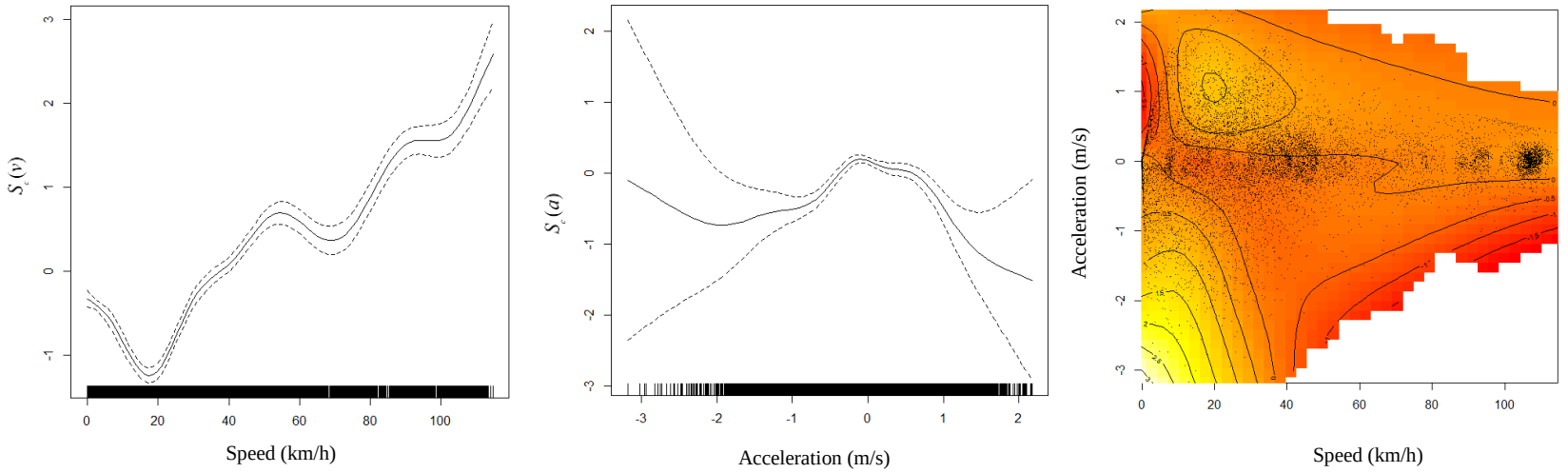


Figure 10: NO_x, vehicle B: a) smooth function of speed, b) smooth function of acceleration and c) smooth function of interaction between speed and acceleration

CO_2 observation (real) against fitted values for vehicle A and B are shown in Figure 11. The reference red line indicates the case in which the fitted would be the same as the observed values. The points are well located along this line, indicate the ability of the GAMLSS model for estimation.

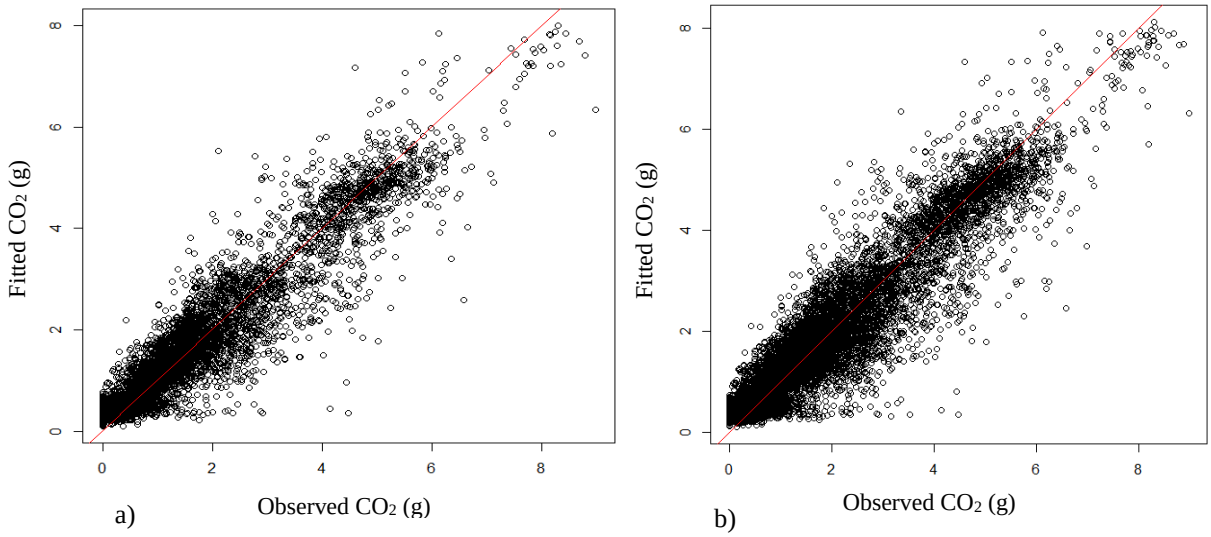


Figure 11: fitted against observations for CO_2: a) vehicle A and b) vehicle B

Autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to check the structure of the residuals. These functions for vehicle A are presented in Figure 12, that indicate clear residuals (white noise) for all the GAMLSS emission models. The ACF and PACF of the vehicle B show the same result as well.

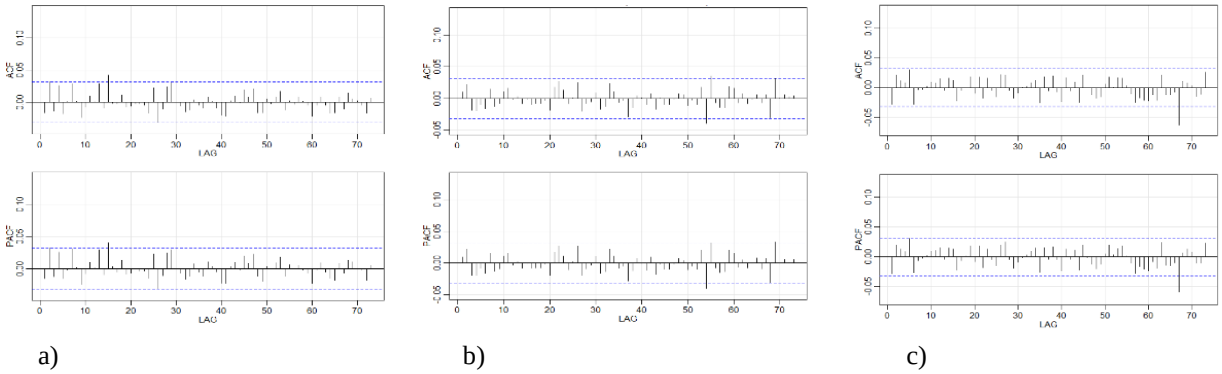


Figure 12: ACF and PACF of the residuals for GAMLSS emission model (vehicle A): a) CO_2, b) CO and c) NO_x

5.3. Evaluation

Two of the most common approaches for emission modelling are selected for comparison with the GAMLSS emission model. These approaches are:

- Generalized Linear Model (GLM)

Multiple linear regression with different explanatory variables as well as GLM are widely used in the literature for emission modelling. Here the GLM model with log link function and log-normal distribution for the response variable is used; this model is defined by:

$$\boldsymbol{\mu} = \exp\left(\mathbf{X}\boldsymbol{\gamma}\right), \quad \mathbf{X} = \begin{bmatrix} 1 & v_1 & v_1^3 & a_1 v_1 \\ \vdots & & \ddots & \\ 1 & v_n & v_n^3 & a_n v_n \end{bmatrix} \tag{14}$$

where $\boldsymbol{\mu} = \left(\mu_1, ..., \mu_n\right)^T$ is the n-vector of estimated mean values for the n observations, $\mathbf{X}$ is the design matrix of explanatory variables, and $\boldsymbol{\gamma} = \left(\gamma_0, \gamma_1, ..., \gamma_5\right)^T$ is the vector of parameters including constant.

- CADI classified model

Using driving mode: cruise, acceleration, deceleration and idling, to explain tailpipe emission during different traffic situations is the common classification in the emission models. This approach is described in section 2.2. The model (2) is applied to the dataset to evaluate the results.

Table 4 shows the results of the evaluation. The GLM and CADI classified model were applied to the dataset and for each vehicle and emission type, degree of freedom, BIC and log-likelihood (L) are reported. The effective degree of freedom e is calculated for GAMLSS models (Wood, 2017), which can be interpreted as an estimate of how many free parameters are needed to represent the spline. Due to penalization, the effective degrees of freedom, which

indicates the amount of non-linearity of the spline, may not be integer. If the effective degrees of freedom for a certain spline is (close to) 1, this means that the function is (close to) linear, whilst a greater value means that the function has a greater degree of non-linearity.

Substantially lower values of BIC were achieved in the GAMLSS model compared to the GLM and CADI classified approach. This indicates the strong advantage of adopting GAMLSS formulation in modelling emissions.

Table 4: BIC of the GAMLSS, GLM and CADI models with measures of their free parameters

CO_2							
	GAMLSS			**GLM (df=5)**		**CADI (df=32)**	
Vehicle	edf *e*	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC
A	58.2	-1241	2970	-3643	7304	-3584	7397
B	97.4	-1933	4727	-6171	12387	-6507	13258
CO							
Vehicle	**GAMLSS**			**GLM (df=5)**		**CADI (df=32)**	
	edf *e*	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC
A	63.8	31540	-62543	7907	-15733	7953	-15676
B	102.8	47256	-93603	28142	-56240	28498	-56751
NO_x							
Vehicle	**GAMLSS**			**GLM (df=5)**		**CADI (df=32)**	
	edf *e*	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC	$\mathcal{L}$	BIC
A	53.7	33844	-67234	25112	-50183	25257	-50284
B	63.2	32808	-65062	23516	-46988	23194	-46143

To complete the evaluation process, the ACF and PACF of the residuals of the GLM and CADI classified models were investigated. The strongly significant serial correlation shown by this shows that the residuals are not white noise and hence indicate that the systematic part of these models is insufficient to capture all the structure. One example of these plots is shown in Figure 13 for CO_2 .

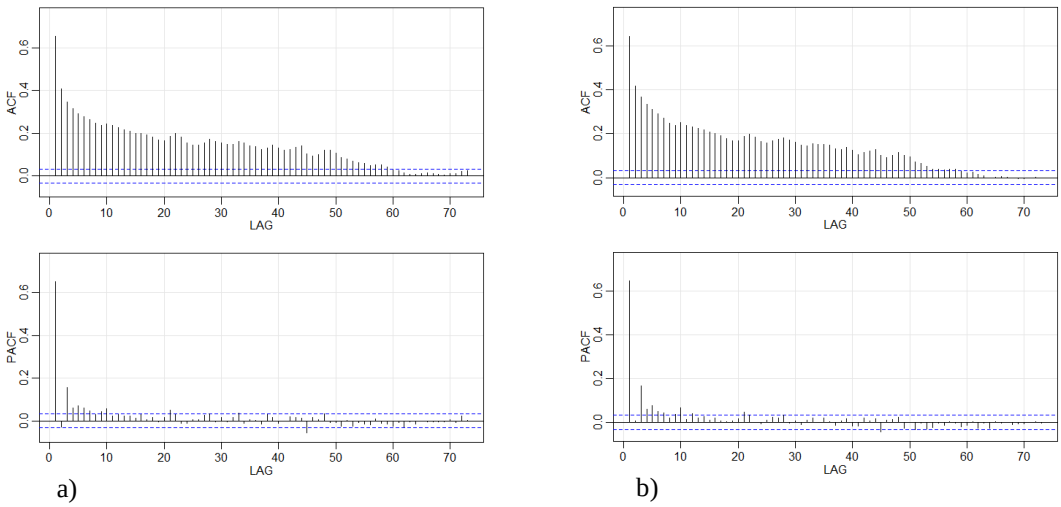


Figure 13: ACF and PACF of the residuals, vehicle 1, CO_2: a) GLM model and b) CADI classified model

5.4. Cross-Validation

Using a same dataset for evaluating the performance of a model as was used for fitting will usually have an unduly optimistic result. To avoid this, cross-validation (CV) procedures were developed in the early 1930s. In the CV, the statistical performance of the model is tested on a dataset that was not used in fitting it, hence giving a better understanding about the model capabilities [20].

In most cases, limitation of the available data is an issue. This leads to the idea of splitting the available data, part for fitting the model and the remainder for evaluating its performance. In this paper, the K-fold cross-validation approach is used to evaluate the performance of the GAMLSS model independent from fitting. In this procedure, the data is partitioned into K subsets or folds, each one of which is excluded in turn from the fitting dataset and then used for model validation as an independent test dataset. This leads to a total of K fitting and testing analysis in which each observation is used K-1 time in fitting and once in evaluation. Because the dataset of this study is a time series, two distinct approaches for partitioning the dataset are implemented: in the first one, the dataset is partitioned randomly without respect to the time order, while in the second approach, the time dependency of the dataset is considered by partitioning according to whole sections of the driving cycle.

- Random partitioning

The dataset was divided into 10 subsets randomly. In this procedure, each observation is used in a test set exactly once and in a training dataset 9 times. This approach fragments the time series of the data, so any advantage implicit in this due to temporal continuity of speed is reduced in fitting and largely eliminated in testing.

The test results for 10-fold cross-validation of vehicle A are presented in Table 5. In this table the test number shows that which of the subsets are used for the test (hence the others are used for fitting the model). The root mean square error (RMSE) s for each test is calculated by:

$$s = \sqrt{\frac{1}{n'}\sum_{i=1}^{n'}(y_i - \hat{y}_i)^2} \tag{8}$$

where n' is the number of observations in each fold (here $n' = \frac{n}{10}$ and $n = 7860$ is the total number of observations), y is the observed emission and $\hat{y}$ is the estimated emission from the model fitted to the dataset complement of the fold.

Table 5: 10-fold cross validation for vehicle A, random partitioning

	RMSE s (mg/s)			R^2		
Test number	CO_2	CO	NO_x	CO_2	CO	NO_x
1	420.7	20.25	0.285	0.83	0.75	0.63
2	435.9	22.58	0.279	0.88	0.78	0.68
3	435.9	22.14	0.288	0.84	0.79	0.63
4	424.3	20.74	0.286	0.87	0.78	0.61
5	412.3	20.49	0.279	0.88	0.79	0.68
6	424.3	22.14	0.275	0.86	0.69	0.71
7	489.9	21.21	0.276	0.85	0.75	0.62
8	883.2	22.80	0.289	0.36	0.63	0.59
9	433.6	22.80	0.285	0.86	0.72	0.53
10	430.1	22.14	0.283	0.85	0.68	0.63
Average	479.0	21.75	0.283	0.81	0.74	0.63

- Time series partitioning

Each of the 9 driving cycle (described in Table 1) is considered as a subset (fold) for CV. In each test, one of the driving cycles is held-out and a model is fitted to the remaining ones. The

hold-out driving cycle is then used for model evaluation. This systematic approach retains the time series structure in both the fitting and the testing dataset.

Each fold is used in a test exactly once and in the training dataset 8 times. The RMSE (Equation 8) is used as an evaluation score for each test. It should be noted that the number of observations in each driving cycle (n') is varies among different cycles. The results of time series cross- validation for vehicle A are presented in Table 6.

Table 6: 10-fold cross validation for vehicle A, time series partitioning

		RMSE s (mg/s)			R^2		
Test number	Driving cycle	CO_2	CO	NO_x	CO_2	CO	NO_x
1	Urban-Free flow	468.2	20.48	0.261	0.85	0.78	0.65
2	Urban-AM peak	473.1	21.15	0.259	0.89	0.74	0.66
3	Urban-Inter peak	471.2	22.01	0.255	0.84	0.77	0.64
4	Suburban-Free flow	432.1	19.63	0.244	0.87	0.76	0.67
5	Suburban-AM peak	449.2	19.33	0.245	0.85	0.77	0.63
6	Suburban-Inter peak	437.3	19.67	0.24	0.84	0.75	0.65
7	Motorway-Free flow	461.3	21.22	0.274	0.87	0.79	0.64
8	Motorway-AM peak	478.3	20.98	0.265	0.86	0.71	0.66
9	Motorway-Inter peak	475.6	20.87	0.271	0.85	0.76	0.65
	Average	460.7	20.60	0.257	0.86	0.74	0.65

In addition to that, the same procedure is applied on the 3 driving cycles (folds) instead of 9: three different traffic conditions (Free flow, AM peak, Inter peak) in each road type are added together to have the urban, suburban and motorway driving cycles (folds). The results of RMSE for each fold is presented in Table 7. This table shows that the suburban driving cycle has the lowest average RMSE compared to urban and motorway road types. This emphasises the importance of including urban and motorway driving cycles in emission modelling compared to the suburban ones.

Table 7: average RMSE in urban, suburban and motorway driving cycles

	Average RMSE s (mg/s)		
Driving cycle	CO_2	CO	NO_x
Urban	470.8	21.21	0.258
Suburban	439.5	19.54	0.243
Motorway	471.73	21.02	0.270

The average of the RMSE results for random and time series CV are presented in Table 8. In this table, corresponding fitting results for the full models (GAMLSS, GLM and CADI) are presented as well.

Table 8: Fitting results of models for Vehicle A on the full dataset

		RMSE s (mg/s)			R^2		
		CO_2	CO	NO_x	CO_2	CO	NO_x
Cross-validation	Random CV	479.0	21.75	0.283	0.81	0.74	0.63
	Time series CV	460.7	20.60	0.257	0.86	0.74	0.65
Model (fitted to the whole dataset)	GAMLSS	443.8	16.43	0.228	0.87	0.85	0.76
	GLM	603.3	26.32	0.413	0.76	0.61	0.21
	CADI	563.9	24.90	0.363	0.79	0.65	0.39

The results of CV approaches show that splitting the dataset without respect to the time series (random) is a stricter approach than time series partitioning, leading to an average of the RMSE that is slightly lower: 3.8%, 5.22% and 8.98% lower in CO_2, CO and NO_x, respectively. in addition, comparison CV results with the fully fitted GAMLSS model shows that although these results are somewhat weaker than the corresponding fully fitted ones, they remain preferable to the GLM and CADI even when those are fully fitted. The GAMLSS model out-performs the GLM and CADI models for each of the three emissions when fitted to the full dataset.

5.5. Prediction for NEDC driving cycle

To show the intended use of the developed method for prediction, the GAMLSS model is applied for NEDC driving cycle. NEDC was used as the reference cycle for emission tests of vehicles until Euro 6, in Europe and some other countries. As it is shown in Figure 14, it contains an Urban Driving Cycle part (UDC) that is repeated four times, and an Extra-Urban Driving cycle part (EUDC) after that.

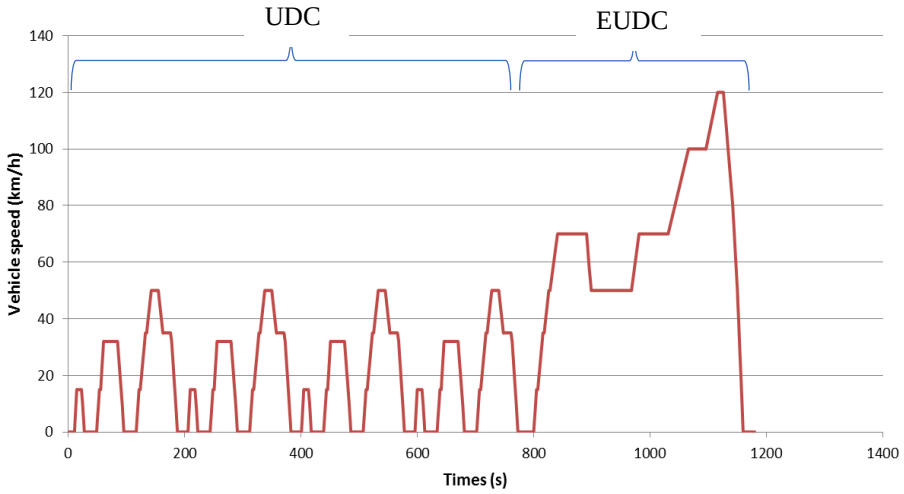


Figure 14: New European Driving Cycle (NEDC) [21]

As an example application, the model for vehicle 1 is applied to predict the CO_2 values of this vehicle while following the NEDC driving cycle (Figure 15).

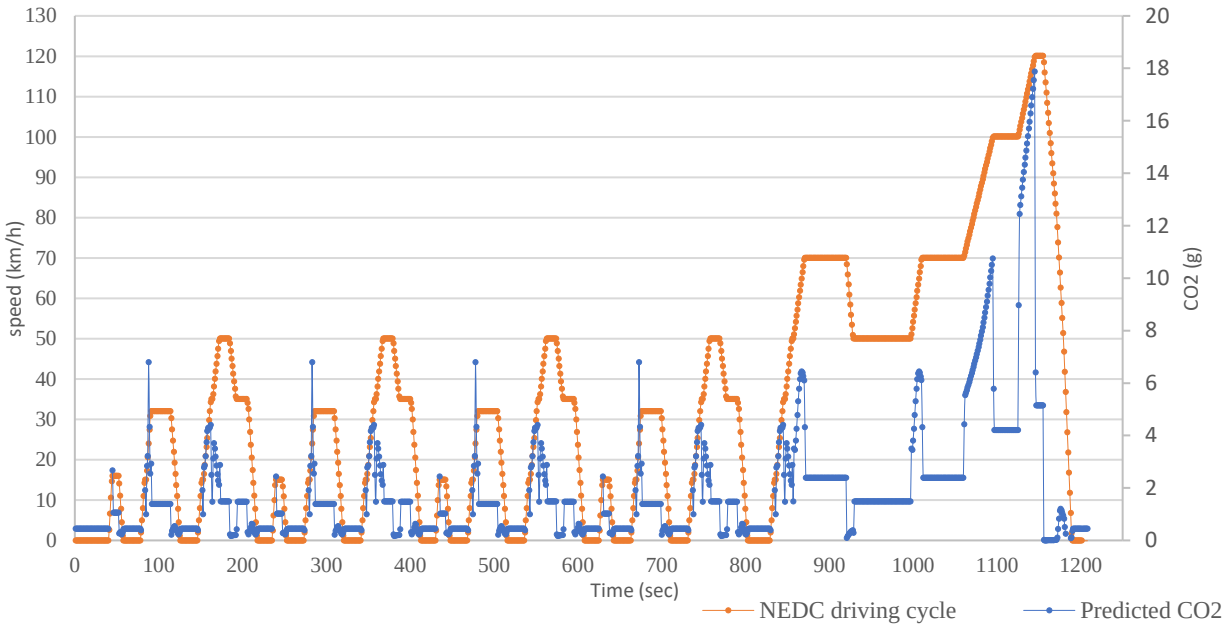


Figure 15: predicted CO_2 values for NEDC driving cycle

The total CO_2 emission per kilometre of NEDC driving cycle for this vehicle type is reported as 110 g/km. GAMLSS approach estimation is 115 g/km, which therefore represents good agreement with the reported emission.

6. Summary and Conclusion

This paper presents a novel statistical approach to build a vehicle emission model, that can describe the relationship between vehicle movements variables, speed and acceleration, with tailpipe emission. Speed, acceleration and their interaction are the explanatory variables of this model. Cubic smoothing spline function of these variables are added together to build a general additive model for location, scale and shape. In this form of model, the error structure is allowed to extend beyond the exponential family of distributions, hence, more flexible distributions can be fitted.

To evaluate these GAMLSS emission models, BIC values of that are compared with GLM and CADI classified model. The results indicate the substantial advantage of this model over GLM and CADI classified model. A 10-fold cross-validation approach was used to test the models. To show the ability of the model to predict emission from a driving cycle, CO_2 emission values are predicted with this model for NEDC driving cycle.

The reason of using cubic smoothing spline is that it offers greater flexibility in following the shape of the relationship between the explanatory variables and the emissions. The generalized linear or classified models are insufficient to describe the relationship between explanatory variables and emission, as shown by the figures of smooth splines of speed and acceleration, which indicate that this is more complicated. Furthermore, there is no requirement to classify vehicle operation regimes in this approach, as all variations can be estimated by a single model.

Acknowledgment

This research was funded by the Lloyd's Register Foundation (LRF), which helps to protect life and property by supporting engineering-related education, public engagement and the

application of research. The vehicle emission dataset is provided by the Transport for London (TfL).

References

[1] J. L. Jimenez-Palacios, "Understanding and Quantifying Motor Vehicle Emissions with Vehicle Specific Power and TILDAS Remote Sensing," PhD, Mechanical Engineering Massachusetts Institute of Technology, 1999.

[2] G. Boutler, S. McCrae, and T. Barlow, "A review of instantaneous emission models for road vehicles ", Transport Research Laboratory (TRL), No. 323-R041 ,2007.

[3] L. L. John Koupal, Edward Nam, James Warila, Carl Scarbro, Edward Glover, Robert Giannelli, "MOVES2004 Energy and Emission Inputs," *U.S. Environmental Protection Agency, Office of Transportation and Air Quality,* vol. Report EPA420-P-05-003, March 2005.

[4] R. Smit, R. Smokers, and E. Rabé, "A new modelling approach for road traffic emissions: VERSIT+," *Transportation Research Part D: Transport and Environment,* vol. 12, pp. 414-422, 8// 2007.

[5] K. Ahn, H. Rakha, A. Trani, and M. Van Aerde, "Estimating Vehicle Fuel Consumption and Emissions based on Instantaneous Speed and Acceleration Levels," *Journal of Transportation Engineering,* vol. 128, pp. 182-190, 2002/03/01 2002.

[6] H. Rakha, K. Ahn, and A. Trani, "Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions," *Transportation Research Part D: Transport and Environment,* vol. 9, pp. 49-74, 1// 2004.

[7] H. Rakha and K. Ahn, "Integration Modeling Framework for Estimating Mobile Source Emissions," *Journal of Transportation Engineering,* vol. 130, pp. 183-193, 2004/03/01 2004.

[8] A. Cappiello, I. Chabini, E. K. Nam, A. Lue, and M. A. Zeid, "A statistical model of vehicle emissions and fuel consumption," in *Proceedings. The IEEE 5th International Conference on Intelligent Transportation Systems*, 2002, pp. 801-809.

[9] F. An, M. Barth, J. Norbeck, and M. Ross, "Development of Comprehensive Modal Emissions Model: operating under hot-stabilized conditions," *Transportation research Record* vol. 1586, pp. 52-62, 1997.

[10] F. A. Matthew Barth, Younglove T, Scora G, Levine C, Ross M and Wenzel T, "Development of a comprehensive modal emissions model " NCHRP Project 25-11 Project 25-11, April 2000.

[11] B. Scora "Comprehensive Modal Emissions Model (CMEM)," University of California, Riverside, Center for Environmental Research and Technology 2006.

[12] "Chasis Dynamometer test procedures for approval of low emission adaptations " Clean Air Zone (CAZ)- Clean vehicle retrofit cretification (CVRC), Version 4 ,August 2017

[13] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 54, pp. 507-554, 2005.

[14] G. Marra and R. Radice, "Bivariate copula additive models for location, scale and shape," *Computational Statistics & Data Analysis,* vol. 112, pp. 99-113, 2017/08/01/ 2017.

[15] S N Wood, *Generalized Additive Models: an introduction with R* CRS Press 2017

[16] S. N. Wood. (2018). *mgcv: Mixed GAM computation vehicle with authomatic smoothness estimation*. Available: URL http://CRAN.R-project.org/package=mgcv

[17] G. Marra and R. Radice. (2018). *GJRM: Generelized Joint Regression Modelling*. Available: URL http://CRAN.R-project.org/package=GJRM

[18] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables*, Berlin, Heidelberg, 1977, pp. 85-100.

[19] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 54, pp. 507-554, 2005.

[20] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.,* vol. 4, pp. 40-79, 2010 2010.

[21] T. J. Barlow, S. Latham, I. S. McCrae, and P. G. Boulter, "A reference book of driving cycles for use in the measurement of road vehicle emissions " TRL2009.

Appendix: R codes for GAMLSS emission model

The R codes for building GAMLSS emission model for vehicle type A and emission type CO2 is presented here. The data frame "V.1" includes: "coo.v1" "s.v1" and "a.v1" that representing CO_2, speed and acceleration for vehicle A, respectively.

```
>library(mgcv)
>library(GJRM)
>library(astsa)
>library(xlsx)
>V.1 <- read.xlsx("V.1.xlsx")                                   #load the data from excel file for vehicle
1
>V.1 <- as.data.frame(V.1)
>f11.coo.v1 <- list(coo.v1 ~ s(s.v1) + s(acce.v1) + ti(s.v1, a.v1),          #define the systematic part of the
model
                   ~ s(s.v1) + s(a.v1) + ti(s.v1, a.v1) )
>Fisk.coo.v1   <- gamlss(f11.coo.v1,  data = V.1, margin = "FISK")          #build the GAMLSS emission model with Fisk
                                                                           distribution
> post.check(Fisk.coo.v1)                                                  #check the selected distribution
> summary(Fisk.coo.v1)                                                     #summary of the model
>plot(Fisk.coo.v1, eq = 1, scale = 0, pages = 1, scheme = 1)              #plot the fitted model based on the first
                                                                           equation (Location)
>plot(Fisk.coo.v1, eq = 2, scale = 0, pages = 1, scheme = 1)              #plot the fitted model based on the
                                                                           second equation (Scale)
>pred.Fisk.coo.v1 <- pred.mvt(Fisk.coo.v1, fun = "mean", n.sim = 100, prob.lev = 0.05, newdata = V.1)      #use the model
                                                                                                           for estimation
>acf2( V.1$coo.v1 - pred.Fisk.coo.v1$pred)                               #plot acf/pacf of the residuals

>plot(V.1$coo.v1, pred.Fisk.coo.v1$pred)                                 #plot fitted against observations
```