*Paper Number

A NOVEL STATISTICAL MODEL FOR THE EVALUATION OF VEHICLE EMISSION FACTORS. APPLICATION TO A EURO III GASOLINE CAR FLEET

M. Rapone*, M.V. Prati*, G. Meccariello°, L. Della Ragione*, M.A. Costagliola°.

* Researcher CNR Istituto Motori

° Under CNR Research Contract

2005 SAE_NA section

ABSTRACT

A novel model has been developed for the analysis and the evaluation of average vehicle emissions in a real driving cycle (emission factors) from data in an emission data base. The model assumes that emission variation can be explained by parameters determined from dynamic vehicle equation and by the frequency of acceleration events at different speed. Because the number of resulting X-variables is large, and variables are correlated, a regression method based on principal components, the Partial Least Squares (PLS) method actually, has been adopted. In this paper, model potentiality is illustrated by an application to a case study taken from the data base built within the UE V Framework Project ARTEMIS. Data are relative to tests performed under hot conditions with a sample of EURO III 1.4-2.0 l gasoline passenger cars. A set of real driving cycles was utilized as representative of urban, rural and motorway operating conditions detected in different European countries. Results for PLS model fit are good for CO2, less than sufficient for CO, HC and NOX; this last result, mostly due to data spread out, is analyzed in the paper by estimating the percentage vehicle's effect.

INTRODUCTION

Average emission factor models commonly used, like COPERT III developed in Europe [1], as well MOBILE6 [2] and EMFAC [3] models developed in the United States, are based solely on the average trip speed to predict emissions. They are intended to evaluate emission inventory of a large road network, where average speeds are generally obtained by traffic assignment models. These macro-scale models cannot differentiate emission rates of trips with the same average speed but with different speed profile, thus they are not sensitive to variations of vehicle's instantaneous speed and acceleration, which have a strong effect on emissions and fuel consumption [4].

The average emission factor model proposed in this paper attempts to improve the sensitivity of macro-scale emission factor models including acceleration related terms in the regression equation. The model is intended to predict the average emission factor relative to a driving cycle (DC), considering the emission data base collected within the UE V Framework Project "Assessment and reliability of transport emission models and inventory systems" (ARTEMIS) [5]. Emission data are relative to a collection of vehicles of different fuel, technology, homologation, and size, tested on a consistent scenario of real driving cycles. To explain the resulting emission variability, a hierarchical statistical approach has been adopted to analyze and predict the pattern of various emissions as a function of the composite set of parameters used to characterize a driving cycle [6].

METHODOLOGY

The aim of research is to develop a statistical model based on emission measurements and concepts of mechanisms of emission production. A first multidimensional model was developed considering kinematic parameters generally used to characterize a driving cycle and the PLS regression method applied to build the prediction model [7]. Further study suggested to improve the statistical modeling approach by a deepened analysis of emission production mechanisms and of characteristics of emission measures [6]. The novel model presented here considers two potential sources of emission variation: the change of total exhaust mass produced in a driving cycle, the change of the distribution of the frequency of acceleration events at different speeds.

Consequently, a first block of parameters was identified considering that exhaust mass is proportional to fuel consumption, which in turn can be obtained by the integral of instantaneous power spent by vehicle in the driving cycle. Starting by the dynamic vehicle equation, it can be shown that exhaust mass is a function of running mean speed (MV), mean of square speed (MV2), mean of cube speed (MV3), running time (T_RUNNING), mean of instantaneous values of product ($a(t)\bullet v(t)$) when $v(t)>0$

and a(t)>0 (M_VA_POS). Moreover, idling time (T_IDLE), considers the emission production during idling time, and the reciprocal of driving cycle length (INVDIST) takes into account that unit emission mass (calculated as emission mass in a test divided DC's distance and expressed in g/km) is to be predicted.

Acceleration events are considered in the model by the joint empirical distribution of second by second instantaneous speed/acceleration {v(t),a(t)} of driving cycle. Hence, the second block of X-variables is obtained by the (cumulated) frequency of {v(t),a(t)} in each of forty two cells obtained considering the intersection of six speed classes and seven acceleration classes. Because frequencies are compositional data, to avoid bias in the estimates, they have to be centered respect to geometric mean and then log-transformed, they are indicated as FS_V20a1, FS_V20a2,…, FS_V101a7 [8].

Because emissions have positive values with not small coefficient of variations, and result generally distributed as a log-normal distribution, a log-transform is applied to emission data.
Thus the following regression equations are defined for the two blocks of variables and for each response Y (CO, CO2, HC, NOX expressed in g/km):

$$\ln Y = a_0 + a_1 MV + a_2 MV2 + a_3 MV3 + a_4 MVA\_POS + \\ + a_5 Trunning + a_6 Tidle + a_7 INVDIST + \varepsilon \quad (1)$$

$$\ln Y = b_0 + b_1 FS\_V20a1 + b_2 FS\_V20a2 + .. \\ .. + b_{42} FS\_V101a7 + \varepsilon \quad (2)$$

where random noise ε is assumed to be a random variable normally distributed ε ~ N$(0, \sigma_\varepsilon^2)$

Considering the number of X-variables, the most of which are correlated, it is convenient to utilize a regression method based on principal components (PC), which are latent variables function of original variables and orthogonal each other. In particular, the sparse matrix of data and the presence of missing values suggested to apply the Partial Least Square method and the NIPALS algorithm to estimate the regression model. Moreover, because response variables Y's may be correlated, a multivariate response **Y** (whose components are CO, CO2, HC, NOX) was considered and a multivariate PLS method applied [9,10].

Ultimately, regression equation becomes for each response component Y (CO, CO2, HC and NOX) and for each block :

$$\ln Y = c_1 t_1 + c_2 t_2 + .. + c_k t_k + \varepsilon \quad (3)$$

where t_i is given as a function of original variables x_j j=1,..,p by:

$$t_i = w^*_{i1} x_1 + w^*_{i2} x_2 + .. + w^*_{ip} x_p \quad \text{i=1,k} \quad (4)$$

where c and w* are y-loadings and x-weights respectively. The number (k) of principal components t_k to be retained in the model fit is determined by cross validation method [9,10].

To consider both the contributes of the two blocks in one model, because variables are separated into two conceptually meaningful blocks, a Hierarchical Multi-block PLS method [11] is adopted. Following this approach, a set (t_1, t_2,…, t_k) of principal components (X-scores) is estimated separately for each block of variables, fitting a PLS base model to each block. Then, the super-block regression model (named top-model) is built, by applying the PLS regression of Y-variables on super-scores made by the union of scores of the two base models. Thus the top model estimates the coefficients and predicted values of regression of **Y** on the full set of X-variables made of the two blocks. Mathematical details and estimation algorithms are presented in [11]. Finally, a flexible and comprehensive approach is obtained to analyze emission pattern and predict emission factors. In fact two PLS models are obtained at lower level (base models) showing the details of each block, and at the upper level the full regression relationship is modeled and predicted values estimated. Thus, in a data base analysis one can choose for each emission case study the model with the best results, among the two base models and the top model.

EXPERIMENTAL CONDITIONS

DRIVING CYCLES

Emission data are relative to measurements performed under stabilized hot conditions, with driving cycles determined in the framework of the Artemis project and called ARTEMIS URBAN, ROAD and MOTORWAY, specifically each driving cycle is divided into sub-cycles and sub-cycle emission quantity is computed by the integral of instantaneous emission record over sub-cycle duration. The method of determination and detailed characteristics of driving cycles are reported in [12]. In the remainder a sub-cycle will be named as a driving cycle (DC). Then, emission data will refer to the following sub-cycles: Artemis – Motorway (4 sub-cycles Motor 1-4), Artemis – Rural (5 sub-cycles Rural 1-5), Artemis – Urban (5 sub- cycles Urban 1-5). Each emission observation is the quantity related to each sub-cycle in the data base for a specific vehicle.

FLEET SELECTION

The case study refers to data of ARTEMIS data-base relative to a sample of Euro III homologation passenger cars with engine displacement in the range 1.4-2.0 l, all of them equipped with a three way catalyst. The list of models is shown in table 1.

Make	Model
ALFA ROMEO	156 J TS 2000

ALFA ROMEO	156 1.6
ALFA ROMEO	147 1.6 TNO
ALFA ROMEO	147 1.6 TUG
ALFA ROMEO	147 2.0 TWIN SPARK 16V
ALFA ROMEO	147 1.6
PEUGEOT	206 XS16S
PEUGEOT	306 1.8 16V
BMW	316I
SAAB	95 ESTATE 2.0
HONDA	ACCORD 2.0I VTEC
OPEL	ASTRA CARAVAN 1.6
TOYOTA	COROLLA TS
MAZDA	DEMIO
FORD	FOCUS
FORD	FOCUS 1.6 16V AUTO
VOLKSWAGEN	GOLF VARIANT 1.6 5D
RENAULT	MEGANE 1.6 16V
RENAULT	MEGANE SCENIC 2.0
FORD	MONDEO 2.0
NISSAN	PRIMERA 2.0 CVT
CHRYSLER	PT CRUISER
FIAT	PUNTO 1.8 HGT
RENAULT	SCENIC 1.6 16S
OPEL	ZAFIRA 1.8 16V AUTO

Table 1: Fleet selection.

RESULTS

A hierarchical PLS model was calculated, which implied to build the *basemodel1* (exhaust mass model MG, the *basemodel2* (speed/acc. distribution model MVA) and then the *supermodel* (topmodel MT).

Results of model fit to data are reported in this paragraph. PLS estimates and diagrams were obtained by using Simca P © and SAS System© software.

Diagnostics of block and super level models are reported in the tables (2-4), where percentage amount of explained variance R2Y and cross validated prediction variance Q2 are presented for the three models. R2Y and Q2 values are cumulated respect to PC's retained in the model: 3 for MG, 6 for MVA, 2 for MT. The cumulated values of these indexes present good values only for CO_2 emission, meaning that models explains more than 77% of emission variation. For other pollutants prediction is unsatisfactory (R2Y and Q2 ≈ 0.2) in both blocks (MG, MVA) and also in the Top Model (MT), even if a little improvement results for HC and NOx.

Ln Emission	***MG.R2VY[3](cum)***	***MG.Q2VY[3](cum)***
ln CO (g/km)	0,211758	0,202649
ln CO_2 (g/km)	0,776434	0,768253
ln HC (g/km)	0,17297	0,163934
ln Nox (g/km)	0,189783	0,171906

Table 2: MG Summary Y overview.

Ln Emission	***MVA.R2VY[6](cum)***	***MVA.Q2VY[6](cum)***
ln CO (g/km)	0,236887	0,200313
ln CO_2 (g/km)	0,800415	0,773302
ln HC (g/km)	0,201174	0,174054
ln Nox (g/km)	0,229851	0,193914

Table 3: MVA Summary Y overview.

Ln Emission	***MT.R2VY[2](cum)***	***MT.Q2VY[2](cum)***
ln CO (g/km)	0,229657	0,218149
ln CO_2 (g/km)	0,819361	0,815974
ln HC (g/km)	0,195617	0,191941
ln Nox (g/km)	0,210738	0,204275

Table 4: MT Summary Y overview.

To make a comparison with a classical multiple regression approach, based solely on average speed, a multiple regression model (GLM) based on quadratic polynomial equation of sub-cycle overall mean speed was fitted to same data set. GLM goodness of fit resulted much poorer than PLS ones, as one can see in table 5, especially for CO and HC.

Emission	***R2***
CO (g/km)	0.068001
CO_2 (g/km)	0.636331
HC (g/km)	0.082305
NOx (g/km)	0.168392

Table 5: GLM Summary Y overview.

PLS predicted quantities as lnY were retransformed in original scale to get emission factors, taking into account the retransformation bias [13]

Figures 1-4 report observed data (gray dot), PLS (blue rectangle) and GLM (red diamond) predicted data of emissions versus cycle mean speed. In particular, one can observe the large spread out of experimental data used to fit the regression model, the trend detected by the GLM model for the specific DC (blue continuous curve) and the predicted values obtained by the PLS model (red broken curve). Overall trend of PLS and GLM agreed, PLS predicted values follow the sharp-toothed pattern of emissions versus average speed, being capable to take into account the peculiar effect of each driving cycle, explained by other variables besides average speed. Thus in some cases predicted values can be different for DC's having similar average speed. Moreover, log-predicted values, when retransformed, tend to geometrical mean of observations, which is less influenced by extreme observed values than arithmetical mean, to which GLM predicted values tend. This is better outlined for CO_2 (figure 2), where emission variance due to driving cycle effect (and explained by model) is relevant respect to residual variance. Different patterns result for different emissions: CO_2 and NOx show a decreasing trend with average speed, from urban to motorway DC's, CO shows an opposite trend, while HC shows quite a symmetric trend with minimum values for intermediate speed (rural DC's).

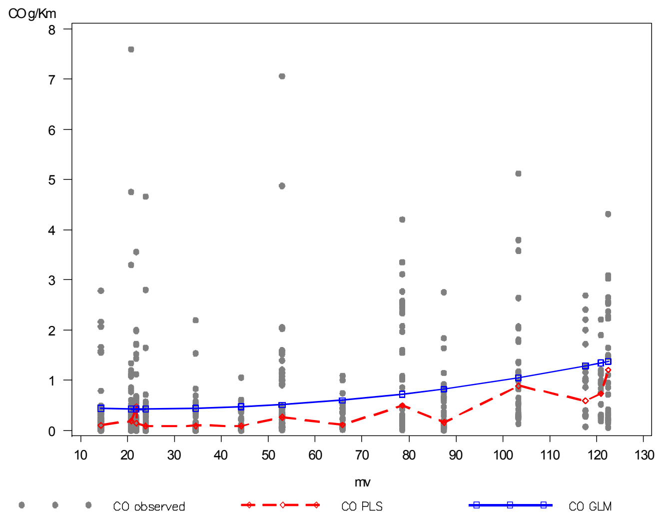


Figure 1: Comparison between experimental data, PLS and GLM predicted CO emissions versus MV (km/h)

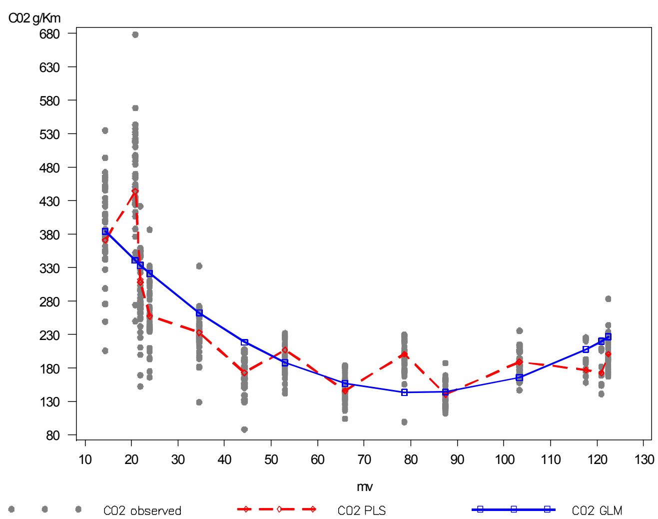


Figure 2: Comparison between experimental data, PLS and GLM predicted CO2 emissions versus MV (km/h)

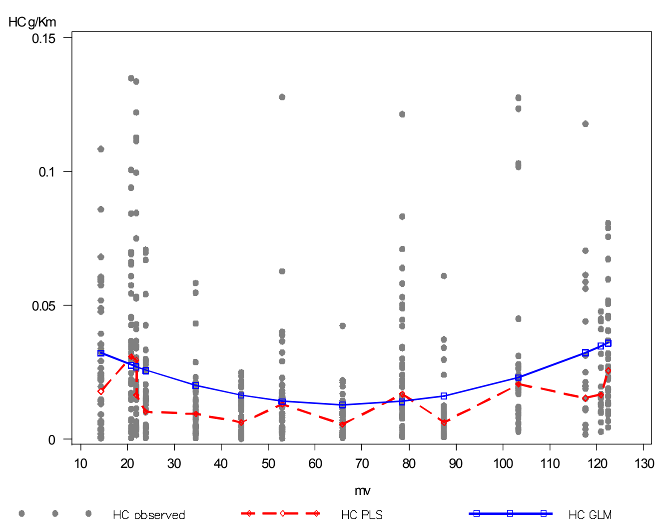


Figure 3: Comparison between experimental data, PLS and GLM predicted HC emissions versus MV (km/h)

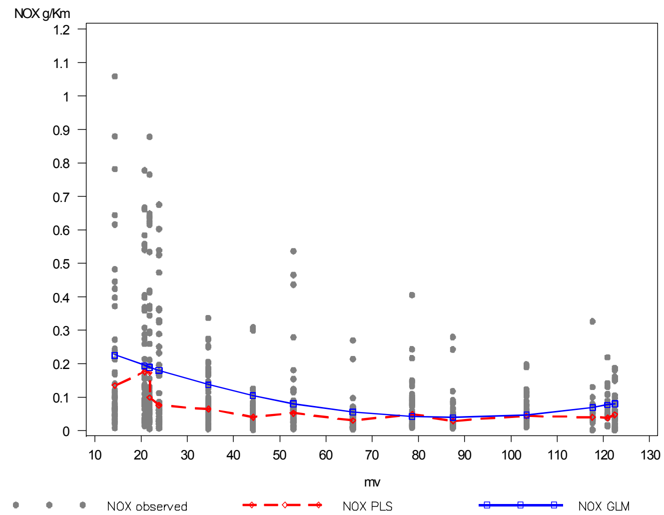


Figure 4: Comparison between experimental data, PLS and GLM predicted NOx emissions versus MV (km/h)

Results of models can be analyzed in the terms of principal components and variables, to show the model capability into explaining emission trends, irrespectively of goodness of fit, dependent on the particular case study.

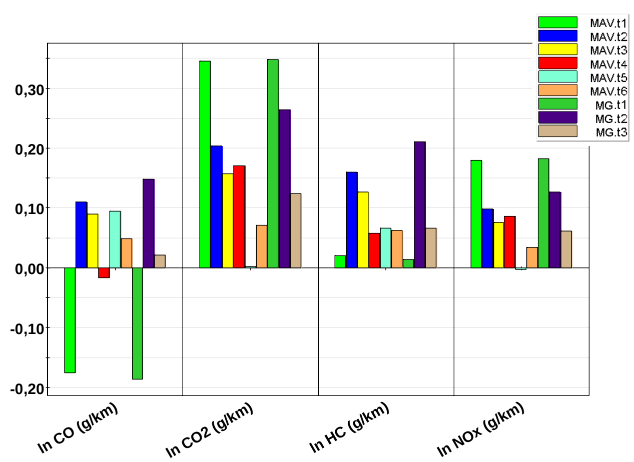


Figure 5: Coefficient overview of MT model.

In the figure 5, where the coefficients of top model are reported, the relative weight of block principal components can be recognized for each emission. It results that for CO, CO2 and NOX the first two components are more important, whereas for HC the second PC is the most important one. The two blocks have similar weights in the model for this case study.

Figure 6 (diagram of y-loadings (c) and x-weights (w*) for the first two PC's) show the relations among the variables of the block 1 (model MG). The representation of observations (labeled by the corresponding driving cycles) in the principal components plane is shown in the figure 7. It can be argued that CO2 and NOX are positively related with TIDLE and INVDIST, which have a positive correlation with t1 and characterize urban DC's (figure 7). Negative relations result with variables (MV, MV2, MV3, T_RUNNING, M_VA_POS) that characterize motorway and rural driving cycles. Similar relations but with opposite sign result for CO. HC resulted explained only by the second component t2, and thus positively related to idling time (TIDLE), and with a lower weight to INVDIST, MV3 and M_VA_POS. As a consequence, predicted HC are higher for DC's Urban 3 and motorway cycles, independently from mean speed.

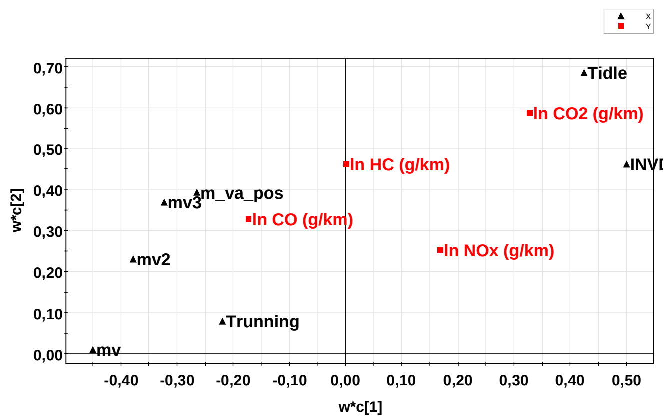


Figure 6: Superimposed PLS loadings of Y and X in MG

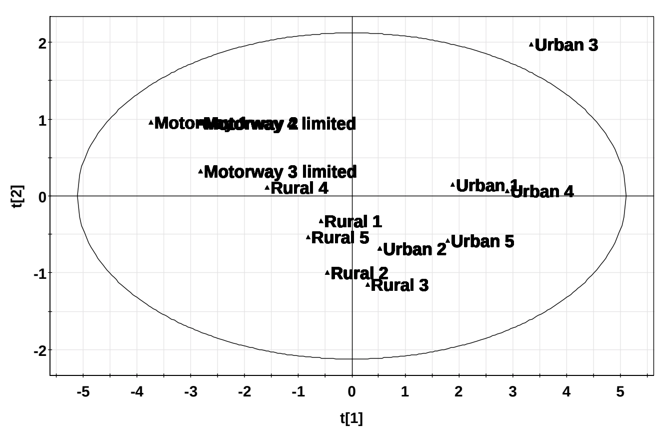


Figure 7: Score plot of t1 vs t2 - MG.

Relations between emissions and block 2 X-variables (speed/acceleration distribution) can be recognized in figure 8, where w*,c diagram of model MVA is reported.

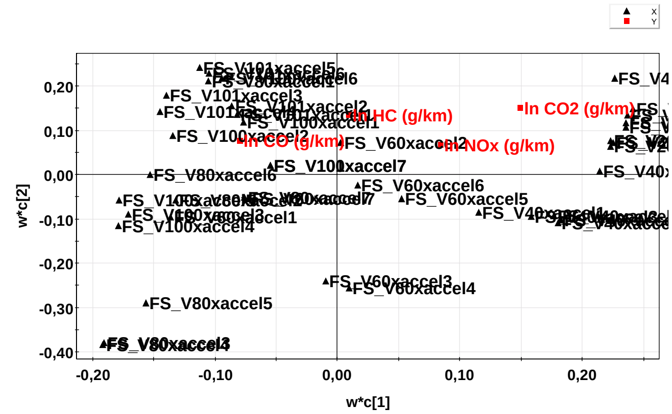


Figure 8:Superimposed PLS loadings of Y and X in MVA

The plot of first two principal components (t1,t2) estimated by MVA is shown in figure 9.

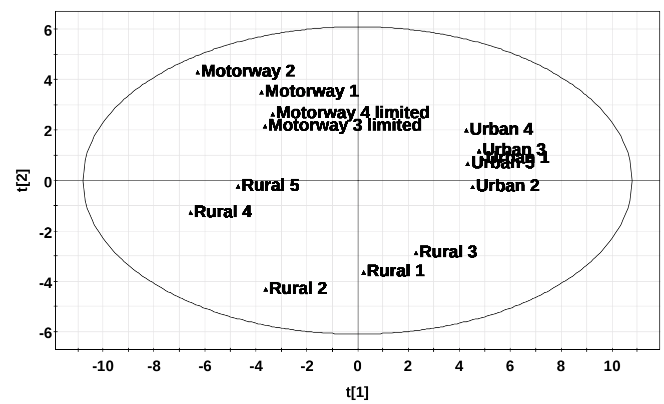


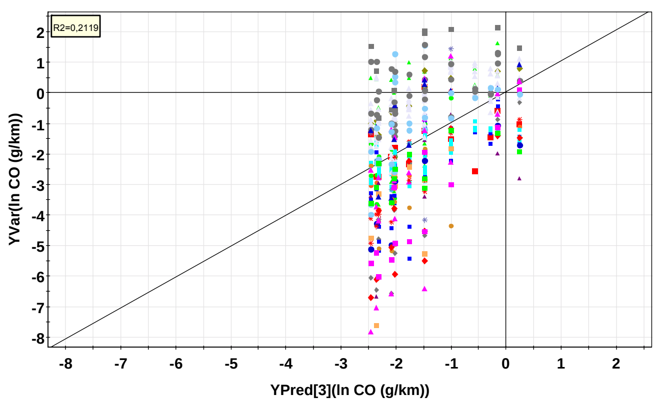
Figure 9: Score plot of t1 vs t2 - MAV.

Considering both figures, 8 and 9 we note a homogeneous distribution of driving cycles respect to speed/acceleration class. CO2 and NOX emissions result positively correlated (higher values) with speed/acc. classes identifying urban DC's, negatively (lower values) with variables characterizing motorway DC's. CO has an opposite trend, HC results positively correlated to the second PC indicating higher values for urban and motorway, lower values for rural DC's. Results are coherent with those obtained by model MG.

VEHICLE EFFECT

The analysis of results has shown that for CO, HC and NOX emission variability explained by regression model is very low. This is well represented in the figure 10, where lnCO observed data are reported respect to values predicted by MG. The variance explained by the model is only about 21% of total variance. In the figure it is evident the large variation of emission data corresponding to the value predicted by model for each driving cycle. The spread out of data can be mostly explained by the peculiar effect of each individual vehicle on emission. This is the summation of the effects of physical parameters like vehicle weight, engine displacement and max power, of (sometime hidden) technological characteristics like electronic control strategy of engine and exhausts gas treatment system, typical for a vehicle model, plus obviously the effect of the status of specific individual related to production variability and mileage.

Thus, a question arises in the prediction of emission factors about which sample of vehicles and emission data should be considered as representative of the specific class, in this case of EUROIII 1.4-2,0 l gasoline passenger cars. The answer is a specific matter and is above the limits of present paper. However, a useful information to achieve in this analysis is the weight that each vehicle's effect has respect to the overall mean.

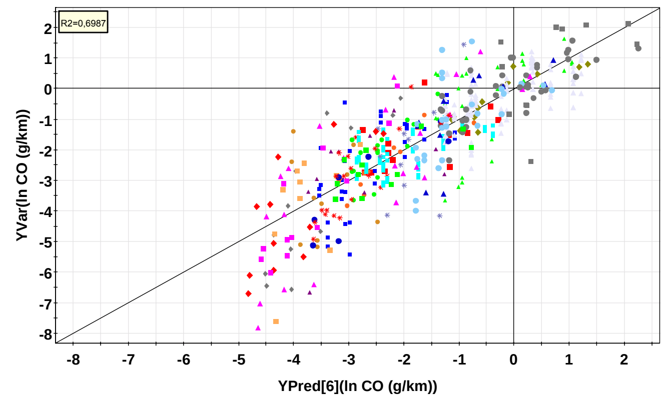


RMSEE = 1,56904

Figure 10: MC-Observed vs predicted values of lnCO.

To this end, a new model (MGD) was built introducing in the model MG a dummy variable D_j (0,1) j=1,N_v (N_v is the number of vehicles) for each vehicle, which estimates the effect of vehicle respect to the general mean, besides the effect of driving cycle estimated by continuous variables. The regression equation of MGD model takes the form of:

$$\ln Y = a_0 + a_1 MV + a_2 MV2 + a_3 MV3 + a_4 MVA\_POS + $$
$$+ a_5 Trunning + a_6 Tidle + a_7 INVDIST + \sum_J \beta_j D_j + \varepsilon$$



RMSEE = 0,972909

Figure 11: MGD Observed vs predicted values of lnCO.

Ln Emission	M_{CD}.R2VY[6](cum)	M_{CD}.Q2VY[6](cum)
ln CO (g-km)	0,69863	0,630737
ln CO2 (g-km)	0,888078	0,835882
ln HC (g-km)	0,594939	0,528908
ln NOx (g-km)	0,573228	0,501927

Table 6: MGD Summary Y overview

Considering vehicle effect in the model by the introduction of vehicle dummy variables increases the variance explained by model, significantly for CO, HC and NOX, as it results from table 6 and from figure 10 compared to figure 11, which refer to same observed data.

It is possible to calculate the percentage effect of each vehicle model on emission factors expressed in original scale (g/km) on the basis of coefficients and prediction errors estimated by PLS model for dummy variables, as it is shown in table 7 and fig. 12 [13]. It results that there are few vehicles which present values significantly different from others; for CO two vehicle models are extremely critical.

Vehicle Model	$\hat{p}$ *CO*	$\hat{p}$ *CO2*	$\hat{p}$ *HC*	$\hat{p}$ *NOx*
156 J TS 2000	-70,435	8,627	242,848	-7,843
147 1.6 TNO	-25,646	4,909	-51,376	26,934
147 1.6 TUG	-50,736	15,072	-67,784	0,513
147 2.0 TS 16V	-89,380	14,930	74,140	2,124
147 1.6 4D	-64,636	4,568	77,194	-81,081
156	129,920	-5,286	239,626	74,448
206 XS16S	-54,486	-15,959	-58,308	33,674
306 1.8 16V	-79,735	-8,372	18,971	-56,003
316I	-31,231	-5,177	-15,203	-57,068
SAAB 95	-86,589	8,112	-47,183	-86,864
ACCORD 2.0I TEC	230,624	-2,812	-53,146	-62,302
ASTRA CARAVAN	-17,724	-25,608	-93,755	-65,406
COROLLA TS	244,629	14,872	44,270	90,606
DEMIO	648,147	-11,269	68,951	-58,241
FOCUS	195,582	-1,303	50,092	245,987
FOCUS 1.6 16V	15,569	-10,748	-52,728	-58,410
GOLF VAR. 1.6	-86,283	-28,201	-15,991	-35,877
LAGUNA II 1.616V	-73,572	-4,280	-18,403	143,392
MEGANE 1.6 16V	-91,009	-21,236	-72,323	-38,000
MEGANE SCE. 2.0	-71,811	-21,183	-63,369	77,434
MONDEO 2.0	89,295	-0,440	-24,214	59,438
PRIMERA 2.0 CVT	-92,754	-3,358	-80,898	-3,169
PT CRUISER	79,252	7,250	-24,593	-11,333
PUNTO 1.8 HGT	-91,623	1,162	-27,089	-28,369
SCENIC 1.6 16S	-56,207	-4,508	-32,954	107,525

ZAFIRA 1.8 16V	851,653	-33,931	95,852	-55,776

Table 7: Estimate vehicle percentage effect.

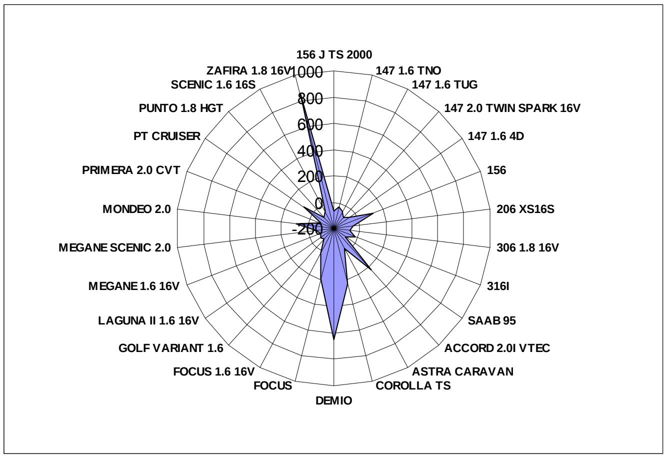


Figure 12: Radar plot of estimate vehicle percentage effect on CO emission.

CONCLUDING REMARKS

A novel average emission model has been developed which can consider the effect of transient modes on emissions. To improve predictability and analysis of results a hierarchical multi block approach was utilized. The application of model to a case study has illustrated the potentiality of the approach. In this case study the vehicle effect is more relevant than driving cycle effect for CO, HC and NOX. Thus a not good fit to data results, both for PLS and average speed model, whereas for CO2 fit is good. This result is explained in the paper by the individual vehicle's effect, whose percentage impact on different emissions general mean is estimated.. This kind of analysis should be considered when emission factors are calculated by large data bases, considering the fact that vehicle composition of data bases generally do not reflect the on road fleet composition, which obviously is varying for different geographic areas.

ACKNOWLEDGMENTS

Experimental data reported in the paper were obtained in the research work carried out within the EU Artemis project. This research was partially funded by EU Artemis project.

REFERENCES

1. Ntziachristos, L., Samaras, Z., COPERT III, Computer programme to calculate emissions from road transport, Methodology and emission factors (Version 2.1), ETC/AEM, November 2000, European Environment Agency Technical report No 49
2. EPA United States Environmental Protection Agency Air and Radiation EPA420-R-03-010 August 2003 User's Guide to MOBILE6.1 and MOBILE6.2 Mobile Source Emission Factor Model
3. Emfac2001/Emfac2002 Calculating emission inventories fro vehicles in California User's Guide (www.arb.ca.gov)
4. Joumard, R., Jost, P., Hickman, J., (1995) Influence of Instantaneous Speed and acceleration on Hot passenger Car Emissions and Fuel Consumption SAE paper 950928
5. Hickman A J and McCrae I S (editor) (2003). Revised technical annex (February 2003) ARTEMIS. Assessment and reliability of transport emission models and inventory systems. Project funded by the European Commission within the 5th Framework Research Programme. DG TREN Contract No. 1999-RD.10429.ARTEMIS website - http://www.trl.co.uk/artemis/
6. Rapone, M., (2005) .A Multiblock Regression Model For The Determination Of Emission Factors. Application To A Case Study, CNR IM report n. 2005 P 1570, proposed for publication
7. M. Rapone, L. Della Ragione, G. Meccariello, M. V. Prati "The effect of different kinematic aspects on emission factors of diesel cars", Proceedings 6th international conference on internal combustion engines – ICE 2003 – Capri (NA) Settembre 2003, SAE_NA 2003-01-062. (n. 2003p1468).
8. Hinkle, J., Rayens, W., 1994, Partial Least Squares and Compositional Data: Problems and Alternatives, Depart,ment of Statistics, University of Kentucky, (1994) http://www.ms.uky.edu/̃jhkinkle/JEHdoc.html
9. Wold, S., Sjostrom, M., Eriksson, L., (2001), "PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems" 58, 2001, pp.109-130.
10. Tenenhaus M, (1998) , La regression PLS Theorie et Pratique. Editions Technip Paris 1998.
11. Wold, S., Kettaneh, N., Tjessem, K., 1996, Hierarchical Multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometrics 1996; 10: pp. 463–482.
12. André M. (2004): Real-world driving cycles for measuring cars pollutant emissions - Part A : The Artemis European driving cycles. INRETS report, Bron, France, n°LTE 0411, 97 p.
13. Van Garderen K., J, Shah, C. (2002), The interpretation of dummy variables in semilogarithmic equations in the presence of estimation uncertainty, UVA Econometrics, Discussion Paper 2002/10, www.fee.uva.nl/ke/UvA-Econometrics

CONTACT

Mario Rapone: m.rapone@im.cnr.it.