VERSIT+ LD: DEVELOPMENT OF A NEW EMISSION FACTOR MODEL FOR PASSENGER CARS LINKING REAL-WORLD EMISSIONS TO DRIVING CYCLE CHARACTERISTICS

Robin Smit, Richard Smokers, Eric Schoen

TNO Science & Industry, business unit Automotive,

P.O. Box 6033, 2600 JA Delft, the Netherlands, robin.smit@tno.nl

ABSTRACT

VERSIT+ LD is a new generation emission factor model which offers the possibility to calculate emission factors for passenger cars as a function of driving cycle characteristics. The model uses empirical relations between emissions and cycle variables derived from the analysis of a large database of laboratory emission test results obtained using real-world driving cycles.

VERSIT+ LD can be applied at various levels of aggregation by using real-world driving cycles which are representative for these situations. Moreover VERSIT+ LD enables the user to estimate the effects of e.g. traffic flow conditions, driving dynamics and driving behaviour on vehicle emissions.

This paper presents the structure of the model, the statistical methodology using weighted least-squares multiple regression, as well as a discussion of the first results.

Keywords: *road traffic emissions, emission factors, emission modelling, passenger cars*

1 INTRODUCTION

Despite great successes in reducing exhaust emissions from passenger cars over the past decades, the emissions from road transport activities still make a significant contribution to air quality problems at a range of scales. Issues of concern include localised exposure to air pollutants in heavy traffic areas (both urban and near highways), photochemical smog formation on a regional scale, total emissions of e.g. NO_x and particulates at a national level, and global warming. To a large extent the size of these problems as well as the effectiveness of possible solutions are assessed on the basis of modelling rather than measurements.

The European Directive on National Emission Ceilings sets challenging goals for emissions of SO_2, NO_x, NMHC, NH_3 at the national level. Monitoring progress in achieving these goals requires accurate average emission factors at the level of the national vehicle fleet, properly reflecting the fact that the relative reduction of real-world emissions in Euro 3 and Euro 4 cars is smaller than the reduction in emission limits and emission values measured in the type approval testing.

Similarly, the strict concentration limits set for NO_2 and PM by the EU air quality directives require accurate predictions of local air quality as a function of the local traffic situation and the effects of improvement measures such as smoothing of traffic flow. Local air quality modelling is based on dispersion models, which require emission factors as input.

On the one hand the difficulties in meeting the above targets presently experienced by many countries lead to a growing need for accurate and more versatile emission estimation tools for road traffic. These tools are becoming ever more crucial for the evaluation of road projects or transport networks and for the design and evaluation of the most cost-effective emission control policies. On the other hand, simulation or other types of modelling of vehicle emissions has become increasingly more difficult due to higher variability in emissions (e.g. De Haan & Keller, 2000). This increasing variability is the result of more complex engine and emission control technology which is used to meet stricter emission standards.

Various traffic emission models exist around the world and they can be roughly classified in terms of their treatment of "kinematic" effects, i.e. the effect of different driving patterns on emissions. Table 1 presents an overview of four types of models. The least complex model (type I) estimates emissions for a few discrete predefined traffic situations (e.g. "urban driving"), whereas the most complex model (type IV) uses several continuous variables (e.g. "root-mean-square-acceleration", "mean idle time") in the prediction process.

Although more complex models may be more accurate and/or more versatile than less complex models, they also require more detailed input data. Therefore, availability of model input data may restrict application of complex models in certain situations. It is noted that there is no "best" type of model. The most appropriate model to use is determined by the application and the required level of detail and accuracy (Smit *et al*., 2002).

Table 1 – Model Classification

Model Type	**I**	**II**	**III**	**IV**
	Discrete		**Continuous**	
Description	Few Traffic Situations	Several Traffic Situations	Univariate Regression Model	Multiple Regression Model
Kinematic Model Variable(s)	Urban, Rural, Motorway	Stop-and-Go, Uneven Bends, Frequent Bottlenecks, etc.	Average Speed	Various Cycle Parameters
Example	MODEM	HBEFA	COPERT MOBILE	IM-CNR VERSIT+ LD

Despite the numerous emission models that exist today, there is a continuous need for further model improvement and understanding of vehicle emissions. This paper reports on the recent development of a new traffic emission model at TNO called VERSIT+ LD. This new modelling approach is designed to predict accurate emission factors for passenger cars, including the latest Euro classes, as a function of driving characteristics. In terms of model type VERSIT+ LD is classified as a type IV model.

The original VERSIT model was first developed at TNO in 1987. This model was based solely on Eurotest emissions data and emissions were basically modelled as a function of propulsion energy. With the advent of new vehicle technology, it was found that the original modelling approach led to unsatisfactory results, particularly with respect to regulated air pollutants.

2 OVERALL MODEL STRUCTURE

The main objective of VERSIT+ LD is to predict accurate mean emission factors, expressed in grams per kilometre, for a particular vehicle category as a function of specific traffic conditions that are characterized in terms of a limited number of driving cycle variables. The focus is on passenger cars, with a possible extension to vans. Vehicle categories in VERSIT+ LD, which are referred to as "model classes", are defined in terms of a combination of emission standard (Euro 1, 2, 3, 4), fuel type (petrol, diesel, LPG), pollutant (CO, HC, NO_x, PM, CO_2) and if appropriate vehicle subclass (direct/indirect fuel injection, with or without diesel particulate filter) or vehicle weight category. A driving cycle is a unique series of idling, acceleration, cruising and deceleration sequences and can be presented in the form of a speed-time profile. Figure 1 presents the VERSIT+ LD model structure.

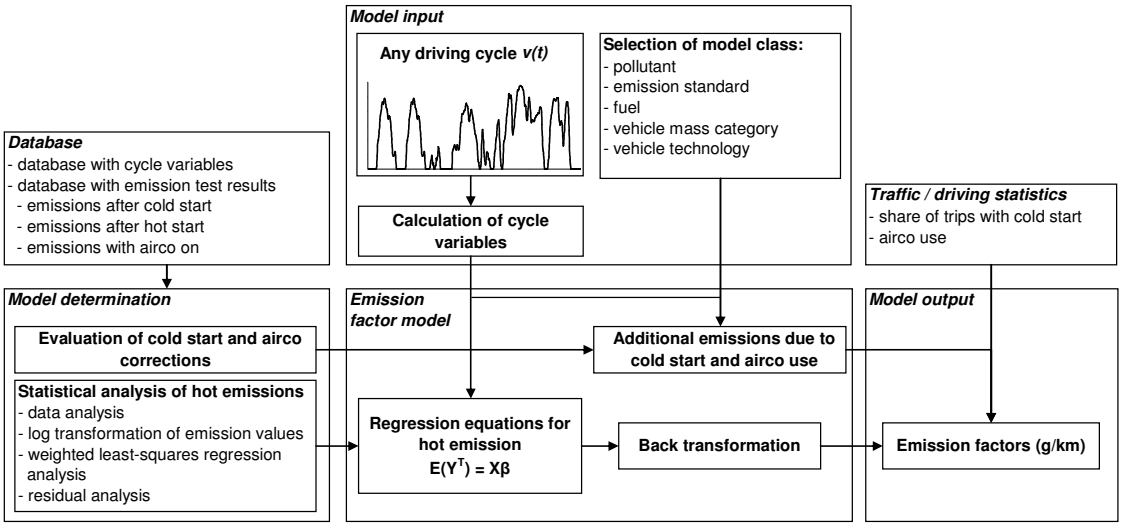

Figure 1 – VERSIT+ LD Model Structure

The core of the model is a set of regression equations presenting an empirical model of the emissions of vehicles in real-world driving with a hot engine. Emissions related to cold start are modelled separately and added to the 'hot' emissions to generate emissions factors including cold start.

The model accepts as input any driving pattern and requires a selection of model class by the user. In the model, the driving pattern data are then used to compute values of all relevant driving cycle variables. These variables are discussed in more detail further on. The computed cycle values are subsequently used by the model algorithms to calculate as model output a mean emission factor.

3 METHODOLOGY

3.1 The underlying database

Driving cycles and associated emissions test data are the fundamental building blocks of the model. VERSIT+ LD is based on a large database that contains almost 12000 emission tests performed on about 700 Euro 1 to 4 vehicles using a total number of 126 different driving cycles. The driving cycles are mostly so-called real-world cycles including widely used test cycles such as the urban, rural and motorway parts of the Modem cycle and the Artemis Common Driving Cycle (CADC) as well as a large number of driving cycles developed by TNO representing e.g. different levels of congestion in highway driving and different driving styles. Straightforward averaging of the emissions measured on different cars on the same real-world driving cycle would constitute a class I or II model according to Table 1, depending on the aggregation level of the driving cycle.

This database is continuously growing as a result of the ongoing Dutch In-Use Compliance programme (e.g. Gense *et al.*, 2000) and new dedicated measurement programmes carried out at TNO. Hence, VERSIT+ LD is regularly upgraded with the latest emissions information.

3.2 Data analysis and data transformation

One of the assumptions of least-squares regression analysis is that the error terms have constant variance (homoscedasticity). As a consequence, the variance of the individual vehicle emissions data (σ^2) must have the same constant value and thus must be stable when plotted against the mean emission factors Y_i for all driving cycles i. A model for the relationship between Y_i and σ^2 is formulated as (Logothetis, 1990):

$$\sigma^2 = \phi\left(Y_i\right)^k, \text{ and hence: } \log(\sigma^2) = \log(\phi) + k \log(Y_i) \quad (1)$$

This model is used to examine whether a stabilisation of σ^2 is required. The variance is stable when k is approximately zero. Analysis of the VERSIT+ LD emissions data showed that for each model class a statistically significant relationship (significance level is less than 5%.) exists between Y_i and σ^2. This relationship is positive, i.e. variance increases with the mean emission value. To illustrate this, Figure 2 displays a scatter plot for one model category.

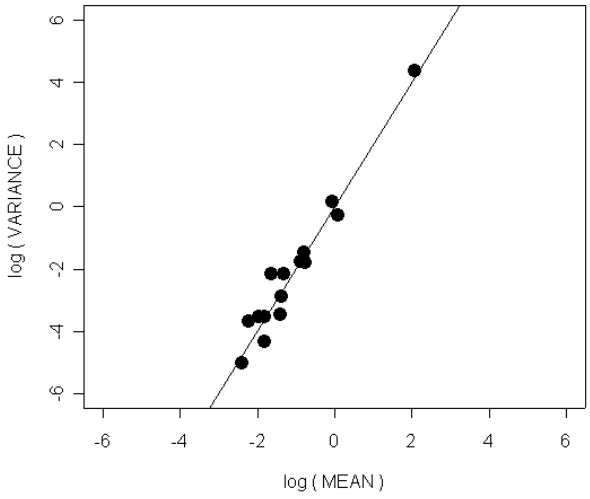

Figure 2 - Relationship between Y_i and σ^2

The k values (including 95% confidence limits) were determined for each model class by simple linear regression of $\log(Y_i)$ on $\log(\sigma^2)$. In Figure 2, the solid line shows the linear regression line where the angle of this line represents the k-value. It was found that the majority of model classes have a k-value of approximately 2. This indicates that a log-transformation (Wu & Hamada, 2000, p. 67) of the raw emissions data would stabilize the variance of these emissions. The log-transformed emission rates Y_i^T are subsequently used in the linear regression and variable selection procedure described below.

3.3 Regression Analysis

VERSIT+ LD consists of a set of statistical models that have been constructed using least-squares multiple regression analysis. The aim of this statistical approach is to find an empirical relationship between emission rates and a limited number of variables, which are selected from a pool of driving cycle variables. For most driving cycles the database contains a substantial number of test results for the same vehicle class. Given that the scatter in individual vehicle emissions is generally quite large and the fact that we are interested in modelling the average emission behaviour of a class of vehicles, a statistical fit can be more efficiently obtained by first averaging for each driving cycle all emission results obtained on that cycle. Subsequently a weighted least-squares multiple regression approach (denoted as WMR) is used which accounts for the fact that different numbers of cars have been tested on the different cycles.

In the WMR procedure a first-order regression model is fit to the experimental data according to the following equations (Neter *et al.*, 1996):

$$\mathbf{E}\left(\mathbf{Y}^{\mathbf{T}}\right) = \mathbf{X}\,\boldsymbol{\beta} \quad (2)$$

$$\boldsymbol{\beta} = \left(\mathbf{X}'\,\mathbf{W}\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathbf{W}\,\mathbf{Y}^{\mathbf{T}} \quad (3)$$

As explained in the previous section the response variable Y_i^T is the mean of the log transformed hot running emissions measured in g/km for a particular driving cycle *i*. In Equation 2 $\mathbf{E(Y^T)}$ represents the vector of expected Y_i^T values or mean responses for all $i = 1$ to n with n the total number of cycles for which measurement data are available. $\mathbf{X}$ represents an $n \times p$ matrix which contains the predictor variable observations (values of driving cycle variables per cycle) for a selected set of p regression variables. $\boldsymbol{\beta}$ represents a $p \times 1$ column vector of regression coefficients, and is determined according to Equation 3, in which $\mathbf{W}$ is a $n \times n$ diagonal matrix containing the weights. These weights are equivalent to the number of emission tests that were conducted on each specific driving cycle. Weights represent the amount of information that is contained in a particular observation. Observations with small variances due to a large number of underlying measurements (and hence a large weight) provide more reliable information about the regression relationship than those with large variances. In this way WMR approach effectively produces a multidimensional hyperplane of best fit through the log-transformed emission test data averaged per cycle.

3.4 Variable Selection Procedure

A pool of in total 49 predictor variables is considered for inclusion in the regression analysis. All these variables quantify certain aspects of driving cycles and they can be calculated once a particular speed-time profile is known. The choice of variables has been based on theoretical considerations (e.g. van de Weijer, 1997), a review of relevant international literature and additional work conducted at TNO.

A classification framework of all 49 cycle variables is provided in Table 2 (next page). Driving cycle variables are categorized according basic type of statistic (location, dispersion) and "aspect". Two abbreviations are used in Table 1, namely σ (standard deviation) and cov (coefficient of variation).

It would take up too much space to describe the computation of each variable in detail, but it is expected that the majority of variables are self-explanatory. This may not be the case for many of the "transient" variables, some of which have been developed at TNO. Transient variables incorporate the rate of change in acceleration per unit distance or unit time. As an example, the computation of RPSI (m/s^3) is given by:

$$RPSI = \frac{\int_0^T v_t \left(\frac{da^+}{dt} \right) dt}{x} \qquad (4)$$

where v_t is the instantaneous speed (m/s), a^+ is the positive instantaneous acceleration (m/s^2), T is the cycle time (s) and x the cycle distance (m).

Given the large number of possible combinations of predictor variables, a procedure is needed to define an optimal and reduced subset of predictor variables for the final regression model. This final model should have a large maximized predictive power, should be unbiased and should preferably include a limited number of variables.

Table 2 – Categorisation of Cycle Parameters Used in the Development of VERSIT+ LD

Aspect	Location Statistic	Dispersion Statistic
Stop	Number of Stops per Kilometre, Mean Stopped Time	-
Driving Mode Proportion	Percentage of Time Stopped, in Acceleration or in Deceleration	-
Speed	Speed (Mean, Maximum, Running, Log), Unit Travel Time	σ (Speed, Running Speed), cov (Speed, Running Speed), TAD
Acceleration	Acceleration (Mean, Maximum) Deceleration (Mean, Maximum)	σ (Acceleration, Deceleration) cov (Acceleration, Deceleration) Root-Mean-Square Acceleration
Power	Power, Acceleration Power, Deceleration Power, RPA, PKE	σ (Power, Acceleration Power, Deceleration Power)
Rolling/Drive Train Resistance	RPS, RSS, RPSS	-
Aerodynamic Drag	RCS, RPCS	-
Transient Factors	Δa_1, Δd_1, Δa_2, Δd_2, D_{ad}, D_{dd}, RAI, RDI, RPSI, RNSI, RPSAI, RNSAI	-

In order to achieve this, it was first investigated whether certain variables can be expressed as linear combinations of remaining variables in the pool. These variables contain no additional information and are therefore superfluous. Statistical analysis resulted in a reduction of the original pool of 49 variables to a pool of 34 variables.

To find the combination of variables that provides the best fit an automatic variable selection procedure ("all-possible subsets") is employed using Mallow's C_p criterion. This criterion is used to find an unbiased regression model with small total mean squared error (bias plus random), and is computed as (Daniel & Wood, 1971):

$$C_p = \frac{SSE}{\sigma^2_{pooled}} - (n - 2p) \qquad (5)$$

where *SSE* represents the "error sum of squares", σ^2_{pooled} represents the pooled variance of the individual log-transformed emission rates over each driving cycle, n the number of driving cycles on which the model is based, and p the number of variables (incl. the intercept). To illustrate the procedure, Figure 3 (next page) graphically shows the results of this procedure for one model class (Euro 2 diesel, NO_x).

Every dot represents a possible combination (subset) of model variables. This chart can be used to identify a combination with:

- a minimized C_p value (i.e. minimized total mean squared error);
- no bias (when the C_p value is equal to or less than p, i.e. below the solid $C_p = p$ line); and
- a minimized number of variables (low p value).

The selected combination for this case is presented by the triangle in figure 3. This combination suggests (not shown) the use of eight model variables in the regression equation for NO_x emission

rates from Euro 2 diesel cars, namely average speed, logarithm of average speed, maximum speed, number of stops per kilometre, mean deceleration, mean deceleration power, RDI and RPSI.

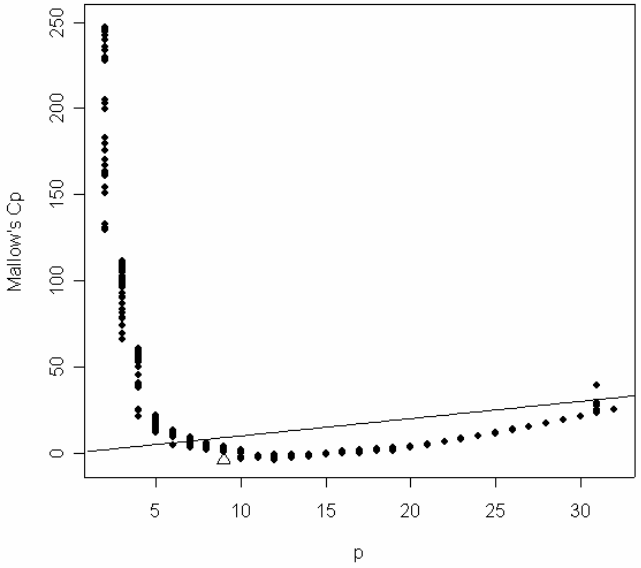

Figure 3 – Automatic Variable Selection Procedure (Example)

3.5 Residual Analysis

Once the model variables have been selected, the WMR procedure is used to estimate regression coefficient values. A next step is to verify that the assumptions of WMR are not violated. This is done by means of residual analysis (Neter *et al*., 1996). This way the regression model is checked for normality of error terms, constant error variance and presence and effect of outlying observations. Once the conditions of valid application of WMR are satisfied, the last step of the modelling process involves back-transformation.

3.6 Back-transformation

The required output from VERSIT+ LD are arithmetic mean emission factors for a pre-specified driving cycle. The arithmetic mean is considered to be the best measure of central tendency since it reflects the disproportionate effect of high-emitting vehicles on total emissions. The regression functions discussed in previous sections, however, predict the mean log-transformed emission factor. Simply taking the antilog of the $E(Y_i^T)$ would result in the geometric mean, which would underestimate $E(Y_i)$. Hence a correction term is required. The expected untransformed arithmetic mean emission factors can be computed by (Johnson *et al*., 1994):

$$E(Y_i) = EXP\left(E(Y_i^T) + 0.5\,\sigma^2_{pooled}\right) \qquad (6)$$

4 MODEL VERIFICATION & VALIDATION

Model verification is the process of making sure that the model is doing what it is intended to do. This is done by comparing model predictions to the observations on which the model is based. As an example, Figure 4 graphically depicts the results for one model class (NO_x of Euro 2 diesel vehicles).

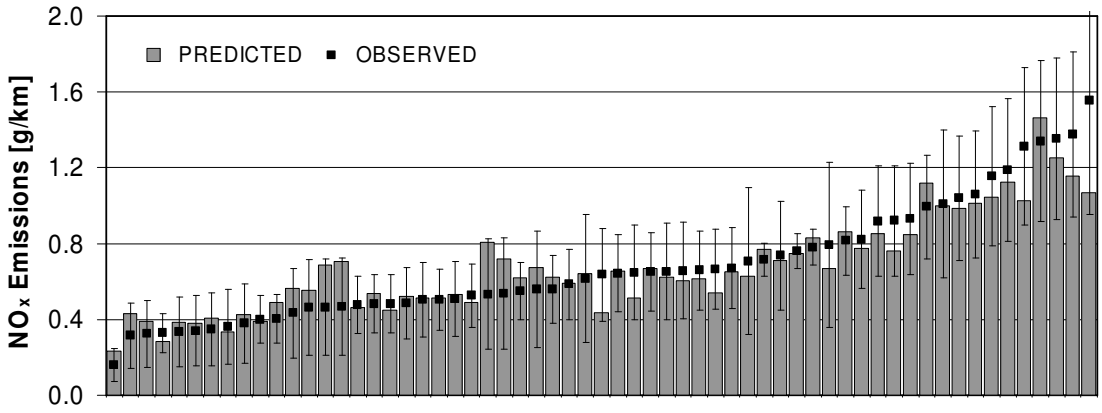

Figure 4 – Model Verification (Euro 2 Diesel Car)

Every bar in Figure 4 represents the predicted emission factor for one particular driving cycle (ordered by size of the average measured emissions). For this model class, the regression equation is based on 61 driving cycles in total. The mean observed values are presented as black squares. The error bars denote the 95 percent confidence interval ($2\Delta Y_{i,95\%}$) calculated on the basis of the pooled variance of all underlying vehicle data:

$$\Delta Y_{i,95\%} = 1.96 \cdot Y_i \cdot \sqrt{\sigma^2_{pooled} / n} \qquad (7)$$

Figure 4 shows that for this model the predicted values are sometimes below, sometimes above the measured test data but they always fall within the 95% confidence interval.

It can be argued that the predicted means are better estimates of the "true" mean (which is unknown) than the measured data points (average emissions over a single cycle) since the predicted values use information from all data points on which the model is based, whereas observed values do not include this additional information.

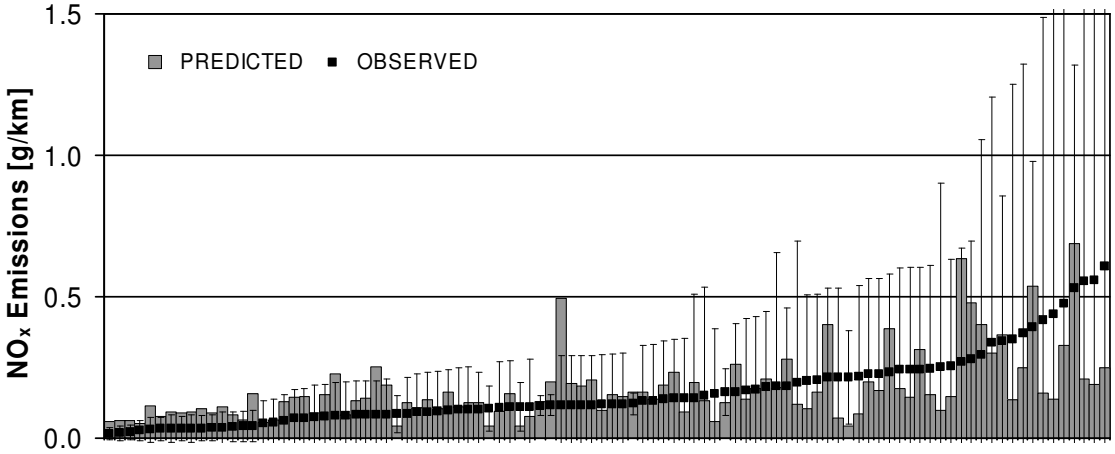

Figure 5 – Model Verification (Euro 2 Petrol Car)

Figure 5 shows the visual model verification for NO_x of Euro 2 petrol cars. This model class is based on a total of 98 driving cycles. In this case, not all predictions lie within the 95%

confidence intervals of the test data. About 5% of the predictions lie outside these intervals, showing a less good fit than for the Euro 2 diesel car model displayed in Figure 4. Nevertheless, the predicted values generally follow the observations quite well. The large confidence intervals on the right side of Figure 5 typically relate to cycles over which only one or a few vehicles have been measured.

Model validation ideally involves comparison of model results to independent observations from the system being modelled. Initial validation of the first version of VERSIT+ LD is conducted by random selection and removal from the database of 25% of the emission data points (cycles on which vehicles have been tested). The regression model is subsequently refitted to this reduced dataset ("reduced model"). This process is carried out repeatedly and the "sample" of model predictions for each cycle is then compared to the observations. This way the robustness of the model is tested. Figure 6 shows the results of this exercise for NO_x emission predictions for Euro 2 petrol cars.

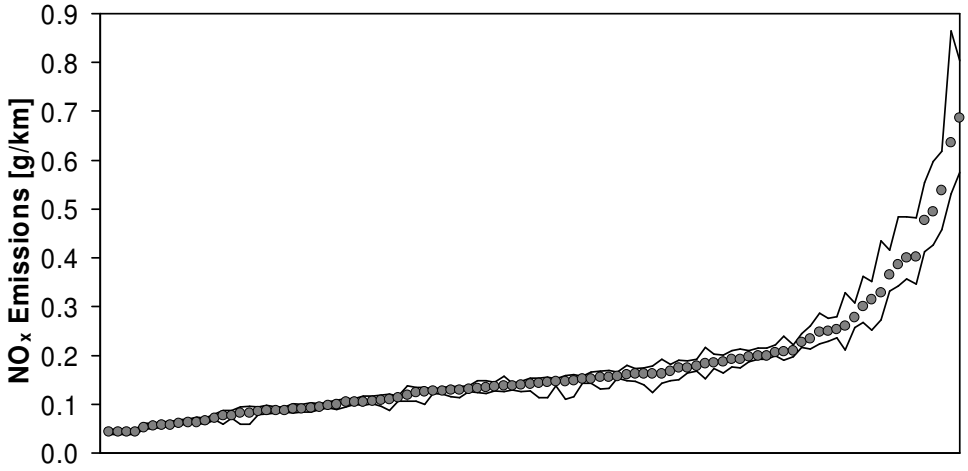

Figure 6 – Model Validation (Euro 2 Petrol Car)

In Figure 6 the grey dots represent the model predictions that are based on all 98 driving cycles (i.e. "full model"). These points are sorted from smallest to largest values. The solid lines in the chart represent the maximum and minimum predicted values that were computed with ten different reduced models. These lines thus show the largest deviations from the full model predictions that were found for a particular cycle in this model validation exercise. Figure 6 shows that deviations are low (generally less than 15%) and that the largest positive deviations (up to 36%) are found when high emission factors are predicted. Overall, the model appears to be quite stable.

Further validation of VERSIT+ LD should in principle be achieved by comparing model predictions to measurement results for driving cycles on which so many vehicles have been tested that the arithmetic mean of the measured emissions is an accurate estimate of the average emissions within a few percent. The pooled variances σ^2_{pooled} found in our analysis generally between 0.1 and 3.0, with a typical value of 1.5.

Equation 7 then shows that to obtain a mean emission value that is accurate within 10% (i.e 95 percent confidence interval divided by the mean ≤ 0.10) requires testing of at least 600 vehicles over a given cycle. This not only shows that a scientifically sound validation of VERSIT+ LD is not readily achievable, but also that models of type I and II, if they are solely based on mean emission rates over representative driving cycles, are expected to have a limited accuracy given the size of test programmes used to provide the underlying data.

5 ONGOING WORK

In terms of ongoing work, sub-models for additional emissions related to cold start and airco use are currently being developed. Furthermore computation of the accuracy of emission factor predictions (i.e. confidence intervals) is being incorporated in the model. Another important aspect that requires further work is the development of a methodology that will determine to what extent extrapolation is acceptable with respect to the model output. Finally, the constant generation of new test data at TNO will allow for regular model updates and model validation.

6 CONCLUSION

With the first version of VERSIT+ LD nearing completion a new generation emission factor model is presented which offers the possibility to calculate emission factors for passenger cars as a function of driving cycle variables. The model uses empirical relationships derived from the analysis of a large database of measurement data. Using representative real-world driving cycles at a high aggregation level VERSIT+ LD can be used to fill type I and II models as identified in Table 1. At a lower aggregation level VERSIT+ LD can for instance be used to estimate the effects of measures influencing traffic flow or driving dynamics on emissions. At this point in time, cycle variables obtained from recorded driving patterns must be used for this kind of application. To enhance applicability in local traffic and air quality modelling the development of intermediate models linking more readily obtainable input data (e.g. average speed, intensity and traffic flow conditions) to typical values of the cycle variables used in VERSIT+ LD will be necessary.

7 ACKNOWLEGDEMENT

TNO gratefully acknowledges financial support from the Netherlands Ministry of Housing, Spatial Planning and the Environment for the development of the VERSIT+ LD model.

8 REFERENCES

Daniel, C. & Wood, F.S. (1971), Fitting Equation to Data, Wiley, New York.

Gense, N.L.J., Rijkeboer, R. & van de Burgwal, H.C. (2000), In-Use Compliance testing of passenger cars in the Netherlands, Intertech's 6th Int. Clean Transport Conference on Vehicle In-use Compliance Testing, Berlin, Germany, October 2000.

De Haan, P. & Keller, M. (2000), Emission factors for passenger cars: application of instantaneous emission modelling, Atmospheric Environment, Vol. 34, pp. 4629-4638.

Johnson, N.L., Notz, S. & Balakrishnan, N. (1994), Continuous Univariate Distributions, Vol. 1, New York, Wiley.

Logothetis, N. (1990), Box-Cox transformations and the Taguchi method, Appl. Statist., Vol. 39, No. 1, pp. 31-48.

Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996), Applied Linear Statistical Models, 4th Edition, McGraw-Hill, Chigago, IL, ISBN 0 256 11736 5.

Smit, R., Ormerod, R. & Bridge, I. (2002) Vehicle emission models and their application - Emission inventories, Clean Air, CASANZ, Australia, Vol. 36, No. 1, pp. 30-34.

Van de Weijer, C.J.T. (1997), Heavy-Duty Emission Factors – Development of Representative Driving Cycles and Prediction of Emissions in Real-Life, Ph.D. Dissertation, Technical University of Graz, Austria.

Wu, C.F.J & Hamada, M. (2000), Experiments: Planning, Analysis and Parameter Design Optimization, John Wiley & Sons, New York.