A new approach for modal study of instantaneous real world emissions by three-way contingency table analysis with ordered categories

Ida Camminatiello and Luigi D'Ambra
Department of "Matematica e Statistica"
University of Neaples "Federico II"
camminat@unina.it; dambra@unina.it

Giovanni Meccariello and Mario Rapone
"Istituto Motori" - National Research Council (IM-CNR), Naples
g.meccariello@im.cnr.it; m.rapone@im.cnr.it

Abstract: *The aim of this paper is to evaluate the NOx emissions trough an accurate analysis of vehicle driving behaviour. For this purpose, a three-way contingency table will be carried out, crossing the NOx emissions, the speed and the acceleration. This contingency table will be analysed by the partition of Marcotorchino index. To complement the survey Ordered Non-Symmetric Correspondence Analysis (ONSCA) will be applied.*

Keywords: Marcotorchino index, kinematic parameters, car emissions

1. Introduction

The physical models, developed to predict exhaust emissions, are extremely complex and require hours of powerful computer running time, so the models mostly used are based on regression analysis of emission data, collected either in laboratory or on the road. Such model estimate real world emission by average or instantaneous emissions data.
The statistical approach proposed in this paper is based on a large emission data base, built within the European project ARTEMIS. In particular urban driving cycle are considered. This data base is a collection of emissions measured in laboratory with vehicles of different technology (diesel, petrol, LPG) and homologation class (EURO 1 to 4).
Instantaneous values of NOx emissions of vehicles are the response, instantaneous values of speed and acceleration are the explicative variables in the proposed approach. A three-way contingency table for each driving cycle is built, where the value in a cell is the frequency of instantaneous values of vehicle speed (v(t)), acceleration (a(t)) and NOx emissions detected in the specific class of speed, acceleration and NOx emissions. The table 1 describes the categories.

	Speed (km/h)	Accelaration (m/s2)	NOx (g/s)
Categories	$0 \leq speed1 \leq 10$	$-\infty \leq acc1 \leq -1.4]$	$nox1 < -0.625$
	$10 < speed2 \leq 20$	$-1.4 \leq acc2 \leq -0.6$	$-0.625 < nox2 < -0.225$
	$20 < speed3 \leq 30$	$-0.6 \leq acc3 \leq -0.2$	$-0.225 < nox3 < 0.475$
	$30 < speed2 \leq 40$	$-0.2 \leq acc4 \leq 0.2$	$0.475 < nox4 < 1.175$
	$40 < speed5 \leq 50$	$0.2 \leq acc5 \leq 0.6$	$nox5 \geq 1.175$
	$Speed6 > 50$	$0.6 \leq acc6 \leq 1.4$	
		$1.4 \leq acc7 \leq +\infty$	

Table 1. The description of categories.

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

2. A combined approach for studying the instantaneous emissions of car fleet.

When the variables are collected in a contingency table, classical statistical tools as correspondence analysis and log linear models are applied.
The aim of correspondence analysis, as well as many multivariate data analytic techniques is to determine scores which describe how different two categories are. To determine the scoring of the rows and columns and the strength of the association between them, the Pearson ratio is partitioned using the method of singular value decomposition.
The log linear analysis focuses on detecting interactions in a multiway contingency table. The basic strategy in loglinear modeling involves fitting models to the observed frequencies in the cross-tabulation of categorical variables. The models can then be represented by a set of expected frequencies that may or may not resemble the observed frequencies. Once the model has been fitted, it is necessary to decide which model provides the best fit. The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. The Pearson Chi-squared statistic or the likelihood ratio (L^2) can be used to test a model fit.
In our analysis, both the methodologies can not be used because there is a directional relationship between the variables (one response variable and two predictors). Moreover the chi squared test requires that the expected cell frequencies are not too small (preferably at least five). Instead our contingency table show several cells equal to zero.
The most proper statistical methodology to analyse our data is the partitioning the Marcotorchino index τ_M for a three-way contingency table with three ordered categorical variables using orthogonal polynomials (Beh E.J., Simonetti B., D'Ambra L., 2007). It allows us to study the dependency relationship between the emissions and kinematic variables respecting the asymmetric and ordinal structure of the data and picking up the nonlinear relationship within the data.
Finally, to graphically describe the dependence structure between the variables, Ordered Non-Symmetric Correspondence Analysis (ONSCA) proposed by Lombardo, Beh, and D'Ambra (2007) will be carried out.
Consider a three-way contingency table *N* that cross-classifies *n* units according to *I* row, *J* column and *K* tube categories. Suppose that the relationship between these three variables is such that the *J* column and *K* tube categories are predictor variables and are used to predict the outcome of the *I* row response categories. A measure of predictability can be made by calculating the Marcotochino index τ_M. It may be interpreted as a measure of deviation from complete independence given the marginal information provided by the predictor (column and tube) variables. If the three variables are completely independent, then the τ_M is zero. If the variation in the row categories are fully accounted for by the column and tube categories then the τ_M is 1.
To determine where possible sources of association exist between the three categorical variables, one may consider the numerator of τ_M. The partition of the Marcotorchino numerator N_{τ_M} involves the generation of orthogonal polynomials (Emerson, 1968) for each of the categorical variables involved in the partition. Considering these polynomials, it is possible identifying sources of variation in terms of the location, dispersion and higher order moments within and between variables.
For the sake of simplicity, the partition of the Marcotorchino numerator N_{τ_M} can be expressed as

$$N_{\tau_M} = \tau_{IJ} + \tau_{IK} + \tau_{JK} + \tau_{IJK} \qquad (1)$$

where τ_{IJ} is the numerator of Goodman-Kruskal (1954) index between the *I* row response categories and the *J* column predictor categories, τ_{IK} is the numerator of Goodman-Kruskal index between the *I* row response categories and the *K* column predictor categories, τ_{JK} is the Chi-

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

squared statistic between the two predictors, τ_{IJK} is the trivariate association between the response and two predictor variables.
To formally test whether there exists (or not) an association between two or more of the variables we may consider C statistic (Light and Margolin; 1971)

$$C = C_{IJ} + C_{IK} + C_{JK} + C_{IJK} \quad (2)$$

The term, C_{IJ}, can be compared with the statistic obtained from the chi-squared distribution for determining if there is a significant asymmetric association between the row and column categories. The other terms can be treated in the same manner. Therefore, the Marcotorchino index, τ_M can be used to determine a global association between the three variables by comparing against a chi-squared statistic.
The proposed approach is illustrated by results relative to a EURO 3 car fleet in the range of 1200-1400 cc.
For our table $N_m = 0{,}297$, it has an associated C-statistic of $C = 3743.451$ (p-value =0,000). Therefore we can conclude that the speed and acceleration influence the NOx emissions.
If we take into account the ordinal nature of the three variables, we can partition the Marcotorchino index and test whether there exists (or not) an association between the variables. The results are shown in table 2.

	IJ	***IK***	***JK***	***IJK***	**Marcotorchino num.**
τ	0,015353	0,120243	0,112082	0,050185	0,297863
%Cont.	5,154303	40,36856	37,628703	16,84843	100
C	192,9517894	1511,1771	1408,6122	630,70967	3743,451
p-value	0,000	0,000	0,000	0,000	0,000

Table 2. The Partitions of Marcotorchino numerator and the C-statistics.

By partitioning C we find that:

- both of predictor variables are statistically significant in influencing the NOx emissions, however the speed is more influential predictor than the acceleration
- there is a statistically significant association between the two predictor variables;
- there is an interaction between all three variables.

Since all three variables are statistically related to one another, we can identify whether there is a location, dispersion, or higher order interaction between at least two of the variables. The decomposition of C_{IJ}, C_{IK}, C_{JK}, C_{IJK} into these components show that the location is the dominant component of interaction between each pair of variables and among three variables.
To provide graphical summary of the relationship between the response and explanatory variables one may consider the coordinates obtained from the ONSCA.
The figure 1 shows the projection of the emission and acceleration categories on subspace spanned by linear and quadratic component. The graph shows a positive relationship between modalities of the two variables. In fact constant or deceleration levels affect low NOx emissions. Similarly, acceleration levels tend to lead to high NOx emissions. In the plot 95% confidence circles have been included. The confidence circles highlight that the levels 4 and 7 of variable acceleration are not statistically significant because their circles involve origin.
The figure 2 shows the projection of the emission and speed categories on subspace spanned by linear and quadratic component Also in this case the plot shows a positive relationship between the categories of two variables. That is low speeds affect low NOx emissions. Similarly, high speeds tend to lead to high NOx emissions. In this case all the categories are statistically significant

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

because no confidence circle involves the origin. Moreover the categories 5 and 6 of variable speed could be unified.

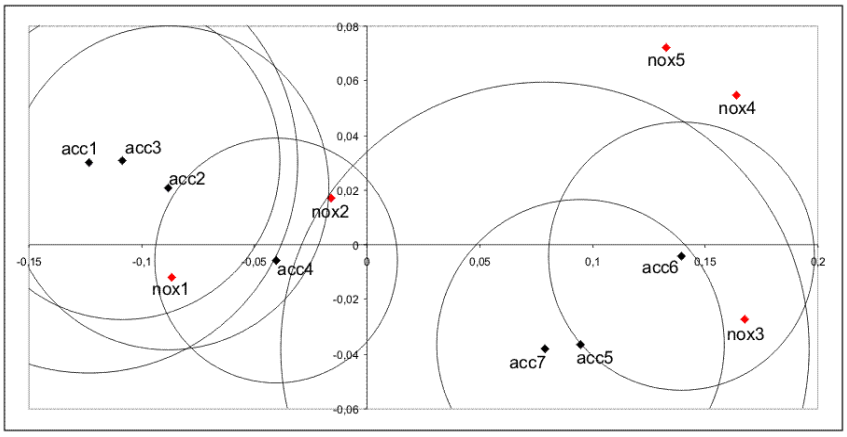

Figure 1: Non-symmetrical correspondence plot: NOx-acceleration

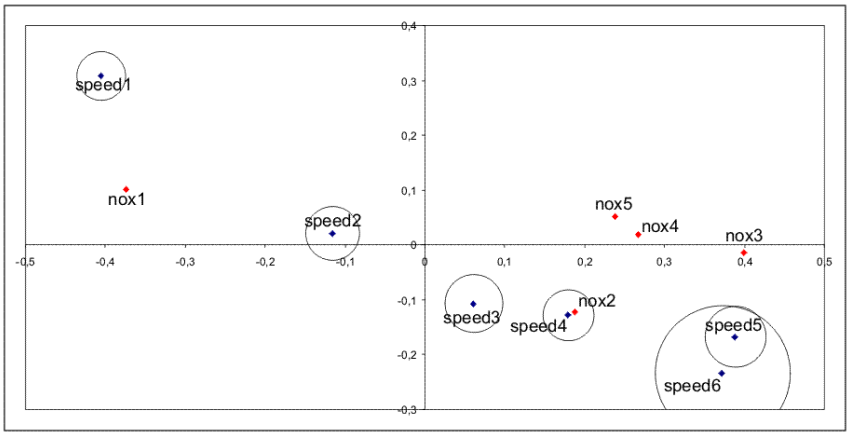

Figure 2: Non-symmetrical correspondence plot: NOx-Speed

Bibliography

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.

Barth M.J., Younglove T., Malcolm T., Scora G. (2002). Mobile source emissions new generation model: using a hybrid database prediction technique, Final report to U.S. EPA under Award 68-C-01-169.

Beh E. J., Simonetti B., D'Ambra L. (2007), Partitioning a Non-Symmetric Measure of Association for Three-way Contingency Tables, Journal of Multivariate Analysis, 98: 1391-1411.

Emerson, P. L. (1968), Numerical construction of orthogonal polynomials from a general recurrence formula, Biometrics Journal, 24: 696-701.

Goodman, L. A., Kruskal, W. H. (1954), Measures of association for cross-classifications, Journal of the American Statistical Association, 49: 732-764.

Hickman A J and McCrae I S (eds.) (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/.

Lombardo R., Beh E.J., D'Ambra L. (1997), Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials, Computational Statistics & Data Analysis, 52: 566-577.