ELSEVIER

Transport Policy 14 (2007) 514–522

TRANSPORT POLICY

www.elsevier.com/locate/tranpol

Multiple classification analysis in trip production models

Cristian Angelo Guevara[a,*], Alan Thomas[b,1]

[a]Universidad de Los Andes, San Carlos de Apoquindo 2200, Las Condes, Santiago, Chile
[b]Chilean Secretariat for Transport Planning (SECTRA), Teatinos 950 Piso 16, Santiago, Chile

Abstract

We analyse various Multiple Classification Analysis (MCA) methods to model trip production (generation). We first show that the MCA version most widely used in transport engineering implies a rarely feasible assumption, the transgression of which may drive a significant overestimation of the future number of trips and a systematic bias in its socio-economic composition. To illustrate this effect, we use Monte Carlo simulation and real data from Santiago, Chile to compare the various MCA approaches, concluding that the aforementioned form should be discarded. Our analysis also shows that the MCA method which is more robust to the structure of the underlying model, is the simple calculation of trip rates as averages for each category. Finally, we hint at the need to use more sophisticated formulations than MCA to model trip production.

Keywords: Multiple classification analysis; MCA; Trip generation; Trip production; Cross-classification

1. Introduction

In 1983, Stopher and McDonald made a pioneering application of multiple classification analysis (MCA), a method regularly used in the Social Sciences, to model trip production (generation) rates within the classic four stages urban transportation model. Afterwards, the method was widely replicated in numerous consulting studies, reported in books and explained in transportation modelling handbooks and reviews.

However, after several years of applications worldwide, we found empirical evidence that the MCA method, as described by Stopher and McDonald (1983), may overestimate the future number of trips in an urban transportation system, a problem that makes a revision of its fundaments a must. With this aim, in this paper we first study the basics of various MCA methods described in the literature. In Section 3 we compare these methods by means of two Monte Carlo experiments and in Section 4 we do the same using real data from Santiago, Chile. Finally, we present the main conclusions and recommendations derived from our research.

2. MCA methods described in the literature

2.1. First MCA method described by Stopher and Mcdonald (MCA_SM1)

We begin analysing the first of two MCA methods described by Stopher and McDonald (1983), which we will consequently denominate MCA_SM1. This is the MCA method most widely used in transport engineering worldwide and is also described in Ortúzar and Willumsen (1994), Clark (1996), ODT (1995), SECTRA (1998) and TMIP (2004), among many others.

MCA_SM1 can be applied to multiple clusters but, to simplify the analysis and the comparison with other methods, we will consider a special case where Outbound Household-Based (OHB) trips are modelled as a function of income and motorization (car ownership) stratum only. Under this setting, MCA_SM1 trip rates for a given income-motorization category (*im*) are estimated as the

*Corresponding author. Tel.: +56 2 4129477; fax: +56 2 2149551.
E-mail addresses: aguevara@uandes.cl (C.A. Guevara), athomas@sectra.cl (A. Thomas).
[1]Tel.: +56 2 6710935; fax: +56 2 696 6477.

0967-070X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tranpol.2007.08.001

summation of: (a) the average number trips by household in the whole sample ($\hat{t}$); (b) the difference between the average number of trips by households of income i and the total average ($\hat{t}_{i.} - \hat{t}$); and (c) the difference between the average number of trips by households of motorization stratum m and the total average ($\hat{t}_{.m} - \hat{t}$)

$$\hat{t}_{im} = \hat{t} + (\hat{t}_{i.} - \hat{t}) + (\hat{t}_{.m} - \hat{t}). \quad (1)$$

After some algebra, (1) can be rewritten as expression (1′), where v^h corresponds to the observed number of trips produced by household h; $1^h_{im} = 1$ if household h belongs to category im and zero otherwise; M, I and H correspond to the total number of motorization clusters, income clusters and households, respectively; and where H_{im} corresponds to the number of households which belong to category im

$$\hat{t}_{im} = \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{M} 1^h_{ik} v^h \Big/ H_i}_{\hat{t}_{i.}} + \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{I} 1^h_{km} v^h \Big/ H_m}_{\hat{t}_{.m}} - \underbrace{\sum_{h=1}^{H} v^h \Big/ H}_{\hat{t}}. \quad (1')$$

Stopher and McDonald (1983) state that MCA_SM1 estimates correspond to the application of the ANOVA method. As it can be checked in any econometric handbook (see for example Greene, 2003) ANOVA corresponds to an Ordinary Least-Squares (OLS) model which is linear in a set of dummy variables indicating the membership to certain socio-economic stratum.

Under this setting, some Social Science statisticians derived simplified expressions for those OLS estimates, which are applicable to specific cases. The MCA_SM1 method is one of those particular cases. It corresponds to the OLS estimates of a model in which the number of observations by category is exactly the same (see for example Glass and Stanley, 1986). Unfortunately, Stopher and McDonald failed to declare this important assumption associated with the MCA_SM1 method.

The problem is that the number of observations by category could hardly be the same if surveyed households are, as usual, randomly sampled. In our example, because income and motorization are positively correlated, necessarily more observations will be sampled for extreme categories (high income and high motorization; low income and low motorization), and less from the cross-extreme ones (high income and low motorization; low income and high motorization). Remarkably, even in the example used by Stopher and McDonald (1983) to describe the MCA_SM1 method, the number of observations differs substantially by category.

The effect of the transgression of this assumption is important. It may lead to a significant overestimation of the future number of trips and a systematic bias in its socio-economic composition. To illustrate why this occurs, we could first rewrite expression (1) assuming that the conditions under which it is valid are holding, in other words, assuming that number of observations by category is the same. This leads to expression (1″) were $\bar{t}_{im}$ corresponds to the average number of trips by household for each category im

$$\hat{t}_{im} = \frac{1}{M}\sum_{k=1}^{M} \bar{t}_{ik} + \frac{1}{I}\sum_{k=1}^{I} \bar{t}_{km} - \hat{t}. \quad (1'')$$

Expression (1″) means that the OLS estimator of a model, which is linear in income and motorization stratum can be written as the summation of the simple average of $\bar{t}_{im}$ by motorization stratum and by income stratum, minus the total average.

On the other hand, if the number of observations by category is uneven, (1) can be rewritten as

$$\hat{t}_{im} = \sum_{k=1}^{M} \frac{H_{ik}}{H_i}\bar{t}_{ik} + \sum_{k=1}^{I} \frac{H_{km}}{H_m}\bar{t}_{km} - \hat{t}. \quad (1''')$$

When the number of observations by category is not the same, the error we incurred in when using expression (1) to estimate the trip rates, will approximately correspond to the difference between (1‴) and (1″), since the average is consistent estimator of the true trip rate. To see how this bias behave, lasts to recall that, for small i, H_{im} decreases with motorization and, for big i, H_{im} increases with motorization. On the other hand, $\bar{t}_{im}$ increases both with motorization and income in general.

Thus, for example, within a low income stratum, the first term in (1″) is going to be underestimated by the respective expression in (1‴) because smaller $\bar{t}_{im}$ (those of households with low motorization) are going to be weighed by a bigger factor ($H_{im}/H > 1/M$) and the bigger $\bar{t}_{im}$ (those of households with high motorization) are going to be weighed by a smaller factor. An equivalent thing occurs with other stratums and terms. In other words, because extreme categories are "over-represented" and cross-extreme categories are "underrepresented", $\hat{t}_{i.}$ and $\hat{t}_{.m}$ are smaller than they "should be" for low income and motorization strata and bigger than they "should be" for high income and motorization strata, generating an important bias in the trip rates estimated with MCA_SM1.

The practical impact of this specification problem is dual. For the point in time when the model is estimated, this bias will not affect the modelled total number of trips, but its socio-economic composition is going to be spurious. Because wealthiest households' trip rates are upward biased and poorest households' trip rates are downward biased, the composition of the estimated trips will be affected accordingly. For future scenarios the problem is even worst. The bias will additionally imply the overestimation of the impact in the total number of trips resulting from the household's sliding from low socio-economic categories to high socio-economic categories, as economy grows. This effect is illustrated and discussed in Section 3 using Monte Carlo simulation, and in Section 4 we show its impact using real data from Santiago, Chile.

2.2. Second MCA method described by Stopher and McDonald (MCA_SM2)

MCA_SM2 method is presented by the authors as a correction of MCA_SM1 for cases in which "interaction" among variables is present. In practice, this method corresponds to a numerical correction that tries to consider the fact that the number of observations by category is not equal. In this sense, when the authors talk about "interaction" it has to be understood that this really means correlation among explanatory variables. This method appears also described in Ortúzar and Willumsen (1994) and Clark (1996), but not in the other references cited before.

MCA_SM2 differs from MCA_SM1 just in that $\hat{t}_{i.}$ and $\hat{t}_{.m}$ are now calculated as weighed averages as follows

$$\tilde{t}'_{im} = \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{M}\left(\frac{H_k}{H}\right)1_{ik}^{h}v^{h}\Big/H_i}_{\tilde{t}'_{i.}} + \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{I}\left(\frac{H_k}{H}\right)1_{km}^{h}v^{h}\Big/H_m}_{\tilde{t}'_{.m}} - \underbrace{\sum_{h=1}^{H}v^{h}\Big/H}_{\hat{t}}. \quad (2)$$

It can be noted that the weights considered in MCA_SM2 will tend to smooth the factors in (1‴), improving by this the estimated coefficients, but hardly turning them into the OLS estimates. The net effect of this method should then be a partial improvement in the estimates. This will be verified afterwards by the numerical examples in Sections 3 and 4.

2.3. MCA method of linear ordinary least squares (MCA_LOLS)

The Social Science statistician Nagpaul (2001) describe, what we will call MCA_LOLS, as a type of model where (for our example) the number of trips by household is a function of: (a) the sample average ($\hat{t}$); (b) a "deviation coefficient" $\theta_{i.}$ for the households which belong to a specific income stratum; (c) another "deviation coefficient" $\theta_{.m}$ for the households which belong to a specific motorization stratum; and (d) an error term (ε)

$$v^{h} = \hat{t} + \sum_{i=1}^{I}\theta_{i.}1_{i.}^{h} + \sum_{m=1}^{M}\theta_{.m}1_{.m}^{h} + \varepsilon^{h} \quad \forall h. \quad (3)$$

Nagpaul (2001) indicates that the vector of coefficients θ has to be estimated as the one that minimizes the sum of the square errors (ε). If the number of observations by income-motorization category is the same, the components of θ will correspond to the following coefficients in expression (1) $\theta_{i.} = (\hat{t}_{i.} - \hat{t})$ and $\theta_{.m} = (\hat{t}_{.m} - \hat{t})$. If the number of observations by category is not the same, Nagpaul (2001) proposes an ad-hoc iterative method to find the least-squares errors estimates of the vector θ. Nevertheless, the utilization of this rather complex procedure is unnecessary because the algebraic solution of this problem is well known as the OLS estimates (Greene, 2003).

However, before using the OLS procedure it is necessary to note that the coefficients of model (3) are not identifiable, infinite combinations of coefficients minimize the square error. This occurs because both vectors of dummy variables add up to 1, and thus, perfect colinearity exists. Additionally, if the average trip rate ($\hat{t}$) that appears in (3) has to be estimated instead of been fixed, it would also be perfectly colinear with the dummy variables. Model (3′) below is one of the possible models fully equivalent to (3) and estimable by OLS. In it, $i = 1$ and $m = 1$ were considered as a base and we added a constant coefficient that has to be estimated

$$v^{h} = \beta_0 + \sum_{i\neq 1}\beta_{i.}1_{i.}^{h} + \sum_{m\neq 1}\beta_{.m}1_{.m}^{h} + \varepsilon'^{h} \quad \forall h. \quad (3')$$

Just for pedagogic interest, once the OLS estimates in (3′) are obtained, it would be possible to re-write the model in such a way as to recover the coefficients of problem (3). In any case, the trip rates estimated with models (3′) and (3) are going to be numerically equal. Thus, (3)—and equivalently (1)—should be seen just as a convenient way to present the results obtained by an OLS model like (3′), where the coefficients are re-ordered to recover an interpretation of the marginal impact from a household belonging to a given income or motorization stratum.

Understanding the MCA_LOLS method as an application of OLS, it is possible to use all the computational and statistical tools available for it in the literature. Particularly, if some distribution of the error (ε') is assumed, for example Normal (μ,σ), it would be possible to use statistical tests (F, R^2, etc.) to identify variables for stratification or the size of each stratum.

Summarizing, MCA_LOLS estimates corresponds to ANOVA (or OLS) estimates correctly calculated when MCA_SM1 is not applicable because the number of observations by category is not the same. On the other hand MCA_SM2 method can be seen as a numerical approximation of MCA_LOLS in cases where MCA_SM1 is no applicable. Thus, undoubtedly, if MCA_LOLS is always straightforward to compute, it makes no sense to use an inappropriate method (MCA_SM1) or just an approximation (MCA_SM2) of the correct method to apply in that case (MCA_LOLS).

2.4. MCA method of simple average by category (MCA_SAC)

The last method that we analyse corresponds to MCA_SAC, named in this form because the trip rates in this case are simply calculated as the average number of trips by household for each category. This method, also known as Category Analysis (Ortúzar and Willumsen,

1994), is equivalent to the estimation of an OLS model with dummy variables representing each category (see, for example, Goodman, 1973). Nevertheless, to achieve identification, it is necessary to set the coefficient of one category to zero, just as it is shown (for our example) in

$$v^h = \varphi_0 + \sum_{i\neq 1;m\neq 1} \varphi_{im} 1^h_{im} + \varepsilon''^h \quad \forall h. \tag{4}$$

MCA_SAC (4) can be rewritten as MCA_LOLS (3′) plus a set of interaction (non-linear) coefficients δ as it is shown in

$$v^h = \beta_0 + \sum_{i\neq 1} \beta_{i.} 1^h_{i.} + \sum_{m\neq 1} \beta_{.m} 1^h_{.m} + \sum_{i\neq 1;m\neq 1} \delta_{im} 1^h_{im} + \varepsilon'''^h \quad \forall h. \tag{4′}$$

Therefore, MCA_LOLS could be considered as a restricted model of MCA_SAC. Thus, if it is assumed, for example, that errors are distributed Normal $(0,\sigma^2)$, it would be possible to perform an F-test to check for the statistical difference between both estimated trip rates.

If the null hypothesis is accepted, that is if MCA_LOLS and MCA_SAC are statistically equal, this would be an indication that the underlying model is linear. In that case MCA_LOLS and MCA_SAC would both be consistent, but the first would be more efficient because it entails the estimation of fewer coefficients with the same information. Thus, MCA_LOLS should be chosen.

If the null hypothesis is rejected, this would be an indication that the underlying model is non-linear. In this case MCA_SAC would be consistent but MCA_LOLS will not, because the omitted attributes would be correlated with the observed linear attributes, causing endogeneity (Guevara and Ben-Akiva, 2006).

In this sense, MCA_SAC is more robust to the unknown specification of the underlying model and then should always be preferred in the case that a statistical test to compare it with MCA_LOLS is not available.

If the number of observations by category is too small the trip rates estimated with MCA_SAC will be biased because of the small sample size. Moreover, if the number of observations in a category is equal to zero, that trip rate could not even be estimated. That was one of the main arguments used by Stopher and McDonald (1983) against taking simple averages and favouring the usage of MCA_SM1.

The fact is that, if MCA trip rates by household have to be estimated when few or none observations are available for a category, the first thing to do would be to redefine the stratum boundaries (maybe by grouping categories) and (or) to estimate a model like (4′) where a set of δ coefficients is considered only for the categories with enough observations. Which model should be chosen will depend upon the data used in each case and could be checked using straightforward statistical tests.

3. Comparison of MCA methods using Monte Carlo experiments

To illustrate the impact of the misspecification of the diverse MCA methods, in this section we analyse their level of accuracy through two Monte Carlo experiments. The underlying (or real) model used in the first experiment considers that the trip production rates are linear in income and in motorization. For the second experiment the underlying model considered is non-linear in the same variables. For both experiments we considered a sample of 1000 households distributed by income and motorization. These variables were built as positively correlated and thus, extreme categories became more populated than others, as can be seen in Table 1.

3.1. Linear underlying model

In this experiment, we built the number of simulated trips for each household in the sample, as the summation of the deterministic component (shown in Table 2) and an error term independent and identically distributed (*iid*) Uniform (−0.5, 0.5). The deterministic component or "real trip rates" in this case are linear in household income (1, 2 and 3) and motorization stratum (0, 1 and 2 or more cars). Indeed, it can be seen that these rates raise 0.2 trips for each increase in motorization stratum (independent of the income stratum) and 0.6 trips for each increase in income stratum (independent of the motorization stratum).

It has to be remarked that the consistency, misspecification, or relative efficiency of each of the analysed methods will not change, independently of the error structure considered. This occurs because the four MCA methods

Table 1
Households by category considered for all, Monte Carlo simulations

Income	Motorization			Total
	0	1	2	
1	352	132	3	487
2	109	213	99	421
3	2	24	66	92
Total	463	369	168	1000

Table 2
"Real" rates, linear model, Monte Carlo simulations

Income	Motorization		
	0	1	2
1	1.00	1.20	1.40
2	1.60	1.80	2.00
3	2.20	2.40	2.60

analysed try to minimize the square error of the model. Thus, the utilization of more sophisticated error structures would just obscure the results we try to illustrate though this experiments.

Under this setting, we calculated the modelled trip rates using the methods MCA_SM1, MCA_SM2, MCA_LOLS and MCA_SAC, which results are shown in Tables 3–6, respectively.

We can see in Table 3 that MCA_SM1 trip rates have a "reasonable behaviour" in the sense of be growing with income and motorization. This is one the main arguments stated in the literature in favour of MCA_SM1 against MCA_SAC. However, it can also be noted that, compared with the real rates (Table 2), the MCA_SM1 method is, as expected, upward biased for high income and high motorization categories, and downward biased for low income and low motorization categories.

Table 3
MCA_SM1 estimated trip rates, linear model, Monte Carlo simulation

Income	Motorization (#cars)		
	0	1	2
1	0.697	1.19	1.75
2	1.43	1.92	2.49
3	2.13	2.62	3.19

Table 4
MCA_SM2 estimated trip rates, linear model, Monte Carlo simulation

Income	Motorization (#cars)		
	0	1	2
1	0.982	1.19	1.34
2	1.59	1.80	1.94
3	2.18	2.39	2.54

Table 5
MCA_LOLS estimated trip rates, linear model, Monte Carlo simulation

Income	Motorization (#cars)		
	0	1	2
1	0.994	1.21	1.39
2	1.59	1.80	1.98
3	2.15	2.37	2.55

Table 6
MCA_SAC estimated trip rates, linear model, Monte Carlo simulation

Income	Motorization (#cars)		
	0	1	2
1	1.00	1.20	1.31
2	1.57	1.81	1.98
3	2.23	2.35	2.55

As it was discussed before, this bias will produce the miscalculation of the socio-economic composition of the modelled trips and the overestimation of the impact in the total number of trips resulting from household's sliding from low to high categories as economy grows. For this experiment, this last effect goes up to 183%. Consider for example the case in which the unique change corresponds to the sliding of 100 households from $i=2$ and $m=0$ to $i=2$ and $m=1$. If the real rates were considered (Table 2), the impact of this sliding in the total number of trips would be of $(1.8-1.6)\times 100=20$ trips. On the other hand, if MCA_SM1 estimated rates (Table 3) are considered, the impact in the total number of trips would be of $(1.92–1.43)\times 100=48$ trips. That is, a 183% of overestimation of the change in the total number of trips caused by the sliding of households.

On the other hand, comparing the MCA_SM2 trip rates (Table 4) with those in Table 3 and the "real trip rates" reported in Table 2, it could be noted that, as expected, the application of the MCA_SM2 method significantly reduces the bias problem. Nevertheless, it could be checked that the trips calculated using MCA_SM2 estimates does not equal the observed number of trips, neither by stratum or category. The effect and importance of this last problem will be discussed further on in Section 4.

At last, it can be noted that MCA_LOS and MCA_SAC trip rates (shown in Tables 5 and 6, respectively) are very similar among each other and to "real trip rates" reported in Table 2, and that the bias over extreme categories is no longer present. It can actually be checked (not reported here for the sake of space) that these trip rates are statistically equal.

Afterwards, we repeated the Monte Carlo experiment described in Tables 1 and 2, by generating several error vectors. The objective was to compare the precision attained by the methods examined, independently of the error component. We observed that from the 10th simulation, the ranking of methods became stable. However, to guarantee robustness, 20 Monte Carlo simulations were considered to make the comparison. Subsequently, for each simulation we calculated the average of the absolute difference among "real" and estimated trip rates by category. The results of these 20 simulations were then averaged and are presented in the first row of Table 7.

For this experiment both MCA_LOLS and MCA_SAC are consistent and, accordingly, their average errors shown in the first row of Table 7 are relatively small (below 5%) and similar. MCA_LOLS error is smaller than MCA_SAC because the first method is more efficient as it entails the estimation of only five coefficients instead if nine with the same data. Additionally, the errors for the categories with fewer observations (not reported for the sake of space) are bigger for MCA_SAC because of the small sample bias effect mentioned before.

Table 7
Average error, 20 Monte Carlo simulations

	MCA_SM1 (%)	MCA_SM2 (%)	MCA_LOLS (%)	MCA_SAC (%)
Linear underlying model	27	5.3	2.4	4.8
No linear underlying model	39	24	18	4.8

Table 8
"Real" rates, nonlinear models, Monte Carlo simulations

Income	Motorization (#cars)		
	0	1	2
1	0.800	1.20	2.10
2	1.60	1.80	2.00
3	2.70	2.40	2.80

On the other hand, MCA_SM2 average error stays, as expected, someway between MCA_SM1 and MCA_LOLS (for this example, near but behind MCA_SAC). Finally, MCA_SM1 estimated trip rates are the worst ones by far, with an average error that exceeds the 25% mark. This illustrates the significant impact of the misspecifications that can be faced when MCA_SM1 is used and the number of observations by category is not the same (Table 1).

3.2. Non-linear underlying model

In this section we consider a case in which the "real" trip production rates are non-linear in income and motorization and thus, the effect of a change in income stratum depends on the motorization stratum of the household and conversely. As with the linear case, simulated trip rates were built as the sum of the "real" rate for each category (Table 8) and an error term distributed $U(-0.5, 0.5)$.

In the second row of Table 7 is presented an analysis of the average precision of the various methods examined by means of 20 Monte Carlo simulations. It can be noted that, just as occurred with the linear model, MCA_LOLS is better than the MCA_SM2 and this one is better than MCA_SM1. Nevertheless, in this case the associated errors of those methods are several times higher than those of MCA_SAC. This is because, as we discussed before, the MCA_SAC method is the only one that is consistent when the underlying model is non-linear.

It can also be noted that the errors of MCA_SAC are equal to 4.8%, both for the linear and the non-linear underlying models. This is because its error level is associated only with the number of observations by category and the simulation errors of each of the 20 Monte Carlo experiments, aspects that were not changed when passing from the linear to the non-linear model.

Table 9
Household based trip production rates, diverse methods, ODS 2001

Income	Motorization			Income	Motorization		
	0	1	2		0	1	2
MCA SM1				*MCA SM2*			
1	0.085	0.22	0.41	1	0.23	0.15	0.26
2	0.32	0.46	0.65	2	0.45	0.37	0.48
3	0.60	0.73	0.93	3	0.71	0.64	0.75
4	0.79	0.93	1.1	4	0.95	0.87	0.98
5	0.76	0.89	1.1	5	0.99	0.91	1.0
MCA LOLS				*MCA SAC*			
1	0.16	0.074	0.12	1	0.15	0.20	0.40
2	0.41	0.32	0.37	2	0.39	0.37	0.60
3	0.70	0.61	0.66	3	0.71	0.61	0.59
4	0.91	0.83	0.87	4	0.96	0.81	0.84
5	0.88	0.79	0.84	5	1.1	0.73	0.86

4. Comparison of MCA methods using real data from Santiago, Chile

The final exercise corresponded to the comparison of the MCA methods using real data obtained from the Origin and Destination Survey (ODS) carried out in the city of Santiago, Chile in 2001. Details regarding ODS 2001 can be found in SECTRA (2003).

As opposed to what occurred with the Monte Carlo simulations, in this case the "real" rates are obviously not known. The observed rates were calculated as the number of OHB trips by household observed in the survey sample between 7:15 and 9:15 AM of a weekday in a non-holiday season. Using this data, the estimated rates were calculated using MCA_SM1, MCA_SM2, MCA_LOLS and MCA_SAC, and are reported in Table 9.

In Table 9 it can be noted that, despite the change from income stratum 4 to 5, the rates obtained with MCA_SM1 have a "reasonable" behaviour, in the sense of be rising with income and motorization. However, it can also be noted that, compared with MCA_SAC and MCA_LOLS, MCA_SM1 tends to reduce the rates for low income and motorization stratum and to amplify them for the highest ones.

On the other hand, the MCA_SM2 method does rectify the MCA_SM1 method's trip rates in the right direction, by flatting them and thus making them more similar to the ones obtained with the MCA_LOLS and MCA_SAC methods. However, this improvement still hides a problem

for the estimation of future trips, which will become apparent later when we discuss future land use scenarios.

The trip rates estimated with MCA_LOLS and MCA_SAC are very similar. We performed an *F-test* to verify the null hypothesis that both estimated coefficients were statistically equal. The inputs necessary to carry out this test are:

- The number of observations ($n = 9038$).
- The number of coefficients of the unrestricted model, MCA_SAC in this case ($K = 15$).
- The number of constraints imposed to arrive to the restricted model MCA_LOLS in this case ($J = 8$).
- The R^2 of both models (0.1145 for MCA_LOLS and 0.1166 for MCA_SAC).

The value of the *F-test* in this case is 2.74, which surpasses the critical value of $F[J;\ n-K] = F[8;\ 9{,}023] = 1.94$ at the 95% level of confidence. This implies that the MCA_SAC model is statistically different to MCA_LOLS and, according to what was discussed previously; the MCA_SAC method should be selected to estimate the trip production rates.

Beyond the statistical recommendation of using MCA_SAC for this data, it should be noted that the trip rates of MCA_LOLS and MCA_SAC are very similar for all categories, except the ones with low income and high motorization; and high income and low motorization. Because those categories are the ones with fewer observations, this would be an indication that the difference between MCA_LOLS and MCA_SAC may be caused by a sample size bias. As it was discussed before, in real applications this problem should be rectified by grouping categories with few observations together and (or) by estimating models similar to (4′). However, we will not follow that procedure, because we are more interested in maintaining a complete compatibility among the analysed methods we want to compare.

Therefore, considering MCA_SAC as the selected method, we can now illustrate other important effect of the bias caused by the usage of alternative MCA methods, by comparing their forecasting characteristics. We did this by calculating the trips forecasted with each of the four MCA methods for two Land Use scenarios developed by SECTRA for the years 2005 and 2010. Those Land Use scenarios considered a combination of economic and housing development assumptions, which were specific for the city of Santiago, Chile, and are not reported here for the sake of space.

In the last column of Tables 10 and 11 we present the number of OHB trips by motorization stratum, estimated with the MCA_SAC method for the two land use scenarios, respectively. This number is then used as a reference to calculate the difference between the total number of modelled trips obtained with other methods, amount which is reported on the other columns of the same tables.

It can be seen that, for both time horizons, the total number of trips predicted by MCA_SAC is surpassed by those predicted by MCA_SM1; at the same time the MCA_SM1 estimate is surpassed by that of MCA_SM2. On the other hand, MCA_LOLS method forecasts a total number of trips that is very similar to that of MCA_SAC.

Going now within strata, it can be noted that for categories with zero car (transit captives), the number of trips estimated with MCA_SM1 is smaller than that of MCA_SAC. On the contrary, the number of trips estimated with MCA_SM1 is greater, for households with two or more cars, than the one with MCA_SAC.

These results imply that, all other things being equal, if MCA_SM1 estimates are used in a classic four stage urban transportation model, the number (and the relative percentage) of modelled trips by car will be overestimated

Table 10
Number of OHB trips and percentage referred to MCA_SAC, diverse methods, 2005 scenario

Motorization	MCA_SM1 (%)	MCA_SM2 (%)	MCA_LOLS (%)	MCA_SAC
0	−19	8	−0.53	419,022
1	19	9	0.92	329,156
2+	30	18	−0.39	172,105
Total	3.8	10	0.015	920,283

Table 11
Number of OHB trips and percentage referred to MCA_SAC, diverse methods, 2010 scenario

Motorization	MCA_SM1 (%)	MCA_SM2 (%)	MCA_LOLS (%)	MCA_SAC
0	−18	4,6	−1.8	459,660
1	18	9,9	2.2	448,738
2+	30	18	−0.44	272,236
Total	6.8	9.7	0.055	1,180,635

and the number of transit trips will be underestimated. This effect will be caused by the underestimation (between 19% and 18%) of transit captives users; the overestimation of travellers with a car available (between 19% and 30%); and the negative effect in the modal share of transit that has associated the unrealistic increase in congestion produced by the overestimation of the increase in the number of car trips. Furthermore, this overestimation of trips, vehicles and congestion in the system, would imply an overestimation of social and private benefits of transportation infrastructure projects, which would drive the modeller to erroneous social and private project evaluations.

On the other hand, in Tables 10 and 11 we see that MCA_SM2 still overestimates the number of trips, but that this overestimation is more flat among motorization strata. However, despite the fact that the relative percentage of trips by mode would be better than that derived from the application of MCA_SM1, the overestimation of car trips would also cause an overestimation of congestion and thus, on the shift from transit to car.

Finally, it can be seen that the MCA_LOLS method, which we discarded using the *F-test*, delivers very similar results compared to those of MCA_SAC. This is because the potential inconsistency due to omitted attributes, which characterise ACM_LOLS, is much less important than the serious specification problems of ACM_SM1 and ACM_SM2. In other words, this means that the statistical difference between MCA_LOLS and MCA_SAC is far below the structural misleading picture that can be obtained by using MCA_SM1 or MCA_SM2.

5. Conclusions and recommendations

The first conclusion of this research is that MCA_SM1, the MCA method most widely used to estimate OHB trips worldwide, should be discarded because it is supported by an assumption with very low probability of occurrence in real world, the transgression of which may imply a severe bias in transportation systems modelling.

Additionally, the MCA_SM2 method should be seen as a numerical correction of MCA_SM1, which improves to some extent its results but is still weak, especially in modelling future scenarios. Sophisticated efforts to improve this method, such as the one by Rengaraju and Satyakumar (1994), should be abandoned since the method that MCA_SM2 tries to mimic (MCA_LOLS) is straightforward to compute. Thus, MCA_SM2 should also be discarded.

The MCA_LOLS and MCA_SAC methods are clearly superior to the previous ones, in terms of precision and theoretical basis. The selection of one or another will depend on the case investigated, a decision that can be tested statistically. Anyhow, it can be affirmed that the MCA_SAC method is more robust to the structure of the underlying model. Even if the underlying model is not the "appropriate" one for MCA_SAC, the estimated rates will still be consistent, but just less efficient than the ones of MCA_LOLS. On the other hand, if the underlying model is the "incorrect" one for MCA_LOLS, the parameters estimated with it would be inconsistent.

Additionally, we have to remark that this research should not be seen as a general recommendation for the usage of the simple average by category to model trip production. MCA_SAC is the more robust among the methods analysed, but its statistical adjustment was extremely poor in the experiment with real data ($R^2 = 0.1166$). We tried to improve this model by adding household size as a stratification variable (covering with this the majority of the MCA trip production models estimated worldwide), attaining by this a statistically significant improvement, but with a resultant model with still low explanatory power ($R^2 = 0.1332$).

This reflects the need to use more sophisticated formulations than MCA to model trip production. Drafting on this line, we preliminarily explored some non-linear specifications considering continuous and dummy variables, including accessibility and household composition. This exercise allowed substantial improvements ($R^2 = 0.3012$) but, undoubtedly, lots of effort is still to be done on this line. Other frameworks should be studied, such using logit models for trip production as, among many others, in Wardman and Preston (2001).

Acknowledgments

We would first like to thank SECTRA for providing the data used in this research. We would also like to thank the valuable comments carried out by two anonymous referees and by Juan of Dios Ortúzar, Joaquín de Cea, Reinaldo Guerra, Luis Rizzi, Marisol Castro and the participants of the 11th World Conference on Transport Research at Berkeley, California. Remaining errors or omissions are of our exclusive responsibility.

References

Clark, S., 1996. National multi-modal travel forecasts. literature review: aggregate models. Working Paper 465, Institute of Transportation Studies, Univerity of Leeds.

Glass, G.V., Stanley, J.C., 1986. Métodos Estadísticos Aplicados a las Ciencias Sociales. Prentice-Hall, Englewood Cliffs, NJ.

Goodman, P.R., 1973. Trip generation: a review of the category analysis and regression models. Working Paper 9, ITS Leeds.

Greene, W., 2003. Econometric Analysis, fifth ed. Prentice-Hall, Englewood Cliffs, NJ.

Guevara, C.A., Ben-Akiva, M., 2006. Endogeneity in Residential Location Choice Models. Transportation Research Record: Journal of the TRB, No. 1977, Washington, DC, 2006, pp. 60–66.

Nagpaul, P.S., 2001. Guide to Advanced Data Analysis using IDAMS Software. Retrived online from 〈www.unesco.org/webworld/idams/advguide/TOC.htm〉.

ODT, 1995. Travel Demand Model Development and Application Guidelines. Prepared for Oregon Department of Transportation Planning Section, Transportation Planning Analysis Unit, by Parsons and Douglas Inc.

Ortúzar, J. de D., Willumsen, L., 1994. Modelling Transport. Wiley, London.

Rengaraju, V., Satyakumar, M., 1994. Structuring category analysis using statistical technique. Journal of Transportation Engineering 20 (6), 930–939.

SECTRA, 1998. Metodología de Análisis de Sistemas de Transporte Urbano. Obtenido en línea desde 〈http://www.sectra.cl/contenido/metodologia/transporte_urbano/Analisis_sistema_transporte_urbano.htm〉. Developed by Fernández y De Cea Consultants for SECTRA, MIDEPLAN, Chile.

SECTRA, 2003. Actualización de Encuestas Origen y Destino de Viajes, V Etapa. Final Report of Homonymous Study. Prepared by DICTUC for SECTRA, MIDEPLAN, Chile.

Stopher, P.R., McDonald, K., 1983. Trip generation by cross-classification: an alternative methodology. Transportation Research Record 944, 84–91.

TMIP, 2004. Report on Findings of the Second Peer Review Panel for Southern California Association of Governments (SACG). US DOT Travel Model Improvement Program

Wardman, M.R., Preston, J.M., 2001. Developing national multi-modal travel models: a case study of the journey to work. In: Ninth World Conference on Transport Research, Seoul.