ADVANCED MODELLING TO OVERCOME DATA LIMITATIONS IN THE NORWEGIAN TRANSPORT MODEL

Andrew Daly, James Fox, Charlene Rohr
RAND Europe

1. INTRODUCTION

The Norwegian National Model was created in 1990 (TØI and HCG, 1990) based on the then available National Travel Survey data, which was from 1984/5. The model was originally developed to predict emissions of carbon dioxide and nitrogen oxides from Norwegian private transport as part of a study of global pollution. It also contributed car ownership and fuel consumption information to a general model of the Norwegian economy. Details of the location of traffic within Norway were not needed.

In phases 2, 3 and 4 of the model development the model was updated to use the 1991/2 National Travel Survey data and to give more detailed information including forecasts of road and rail traffic on specific network links.

The model contains sub-models predicting licence holding, car ownership, travel frequency and mode and destination choice for all travel by persons in Norway. There are separate models for long and short-distance travel, the split being made at a (one-way) length of 100 km. as is common in Scandinavian modelling; a separate model for travel in the Oslo area (Åkershus) is used. Zoning in these versions of the models is based on the 435 municipalities of Norway, although there is more detail in Oslo.

The Stage 5 development of the model was intended to improve the locational detail and the treatment of long-distance modelling. To support the long-distance modelling, further surveys of long-distance travel were undertaken in 1997/8 to supplement the trip diary information of the National Travel Survey. However, because these surveys were undertaken at different times and because requirements had moved on since the National Travel Survey was undertaken in 1991/2, two specific problems arose in the Stage 5 development.

- The National Travel Survey data is coded at the level of municipalities and this level of coding was also used for the destination coding of the long-distance survey. However, to improve locational detail it was decided to move the model to operate on a system of 1428 zones. The issue was then how best to exploit the information available to obtain a model operating at the zonal level.

- After preliminary analyses of the long-distance survey data, there was a strong suspicion of a response bias in the data that was collected. In particular this related to the requirement to report trips longer than 100 km., which some travellers may not be able to judge accurately. The issue was how best to proceed in the circumstances.

In these circumstances, TØI asked RAND Europe, now incorporating HCG, the original developers with TØI of the model system, for advice and assistance. The paper describes how modelling approaches were adapted to deal with the problems.

2. LIMITATIONS IN ORIGIN-DESTINATION INFORMATION

The current version (phase 4) of the Norwegian National Transport Model was estimated from travel survey data from 1991/92. In this survey origin and destination information was collected at the municipality level (Norway is divided into 435 municipalities). Consequently, the phase 4 mode and destination models were developed using a zoning system defined at the 435 municipality level.

In 2000, the fifth phase of model development commenced. In the phase 5 modelling work it was decided to use more detailed networks, to allow a more accurate representation of both road and public transport modes. The new zoning system adopted to define these networks uses 1428 zones in place of the 435 municipalities. Note however that the zones aggregate up to define municipalities, with between 1 and 42 zones defined per municipality. Level-of-service (LOS) network measures were defined at the new detailed zoning level for use in the modelling.

During 1997/98, new survey data was collected for use in the phase 5 long-distance mode and destination choice modelling. However, while trip origins were recorded at the detailed zone level, the trip destinations were only recorded at the less detailed municipality level. As a result, it was not possible to use level-of-service matrices defined at the detailed zonal level directly in the modelling. One option to overcome this problem would have been to simply use LOS measures defined at the coarser municipality level. However, this approach would effectively lose all the improvements in accuracy which are gained from the use of more detailed networks. Instead, three different modelling techniques were tested so that LOS measures defined at the zonal level could be used in the modelling. The three techniques are listed below.

- Using a weighted average of the LOS to each zone within the destination municipality;
- Sampling a random zone within each destination municipality and using the LOS for the sampled zone;
- Extending the mode and destination tree structure used in the modelling so that the zonal destination was represented below the municipality, while the choice was specified at the municipality level.

These three approaches are discussed in more detail in the following sections, and are compared to the simpler model based on municipality-level LOS.

2.1 Average Level-of-Service

In this approach, the average LOS to each of the destination zones comprising the destination municipality is used in the modelling. Using a simple (unweighted) average would not account for the relative attractiveness of each of the zones which comprise a municipality. Therefore a weighted average was used, using a measure of the attractiveness of each zone relevant to the journey purpose in question. Typical measures of zonal attractiveness (size variables) are population or employment by relevant sector.

A further consideration when calculating the average LOS is that for some trips, not all the destination zones within the destination municipality will be available. For example, consider a municipality consisting of three zones, one of which defines a small town, with the other two zones defining the surrounding rural area. A bus service may only exist for trips originating or arriving in the small town. Consequently, bus LOS will be defined only to the zone which describes the small town, and so it would not be appropriate to average the bus LOS over all three destination zones. To account for this availability issue, the average LOS was calculated from *available* destination zones only.

For a given LOS measure L, the average LOS from origin zone i to municipality m for mode k was calculated as follows:

$$L_{ikm} = \frac{\sum_Z L_{ikz} A_z \delta_{ikz}}{\sum_Z A_z \delta_{ikz}} \quad (1)$$

where: L is the LOS measure for the mode in question;
m is the destination municipality, which is comprised of Z destination zones;
A_z is the destination attraction measure for zone z;
δ_{ikz} defines the availability for destination zone z for a trip originating in zone i by mode k.

Different modes have different origin-destination availabilities, and so the LOS averaging calculations had to be performed separately for each mode. The output from the LOS averaging procedure was a set of rectangular LOS matrices, with origin defined at the detailed 1428 zonal level, and destination defined at the 435 municipality level.

A further complication of this approach is that the utility functions used to describe the attractiveness of each mode-destination alternative define destination at the municipality level. As a result the size variables, used to measure the attractiveness of each destination alternative, are also defined at the municipality level. This results in a potential inconsistency, because the size variable will describe some characteristic (e.g. total population) of the entire destination municipality, but it may be that only some of the destination zones which comprise the destination municipality are treated as available in

the modelling. In such cases, using the attraction variable for the entire municipality exaggerates the true attractiveness of the mode-destination alternative in question. To correct for this inconsistency, a correction term c was added to the mode-destination utility functions, using the formula defined in equation (2).

$$c_{ikm} = \ln\left(\frac{\sum_Z A_z \delta_{ikz}}{\sum_Z A_z}\right) \quad (2)$$

where: c_{ikm} is the correction term added to the utilities of mode k;
m is the destination municipality, which is composed of Z destination zones;
A_z is the destination attraction measure for zone z;
δ_{ikz} defines the availability for destination zone z for a trip originating in zone i by mode k.

It should be noted that the correction c_{ikm} varies according to the origin zone of the trip, because the availability term δ_{ikz} depends on both the origin and destination.

If all of the zones comprising the destination municipality are available, then the term inside the brackets is one, and as ln(1) = 0, the correction is zero as we would expect.

2.2 Sampled Level-of-Service

The sampled level-of-service (LOS) approach is similar to the average LOS approach, described in the previous section. In the sampled approach, instead of averaging the LOS a destination zone from the chosen municipality is sampled. Rather than sample all zones with an equal probability, destination zones are sampled according to their relative attractiveness, as shown by equation (3).

$$P_{ikz} = \frac{A_z \delta_{ikz}}{\sum_Z A_z \delta_{ikz}} \quad (3)$$

where: P_{ikz} is the probability of destination zone z being sampled from the Z destination zones which comprise the destination municipality for a trip originating in zone i by mode k;
A_z is the destination attraction measure for zone z;
δ_{ikz} defines the availability for destination zone z for a trip originating in zone i by mode k.

As can be seen from equation (3), only zones which are available for the trip in question are sampled.

This approach shares the main disadvantage of the LOS averaging approach, namely that the utility functions used to describe the attractiveness of the

mode-destination alternatives define destination at the municipality level, and so the attraction variables used to represent the attractiveness of each destination alternative may exaggerate the true attractiveness of the destination. To correct for this problem, the correction term defined in equation (2) was also used in the sampled LOS models.

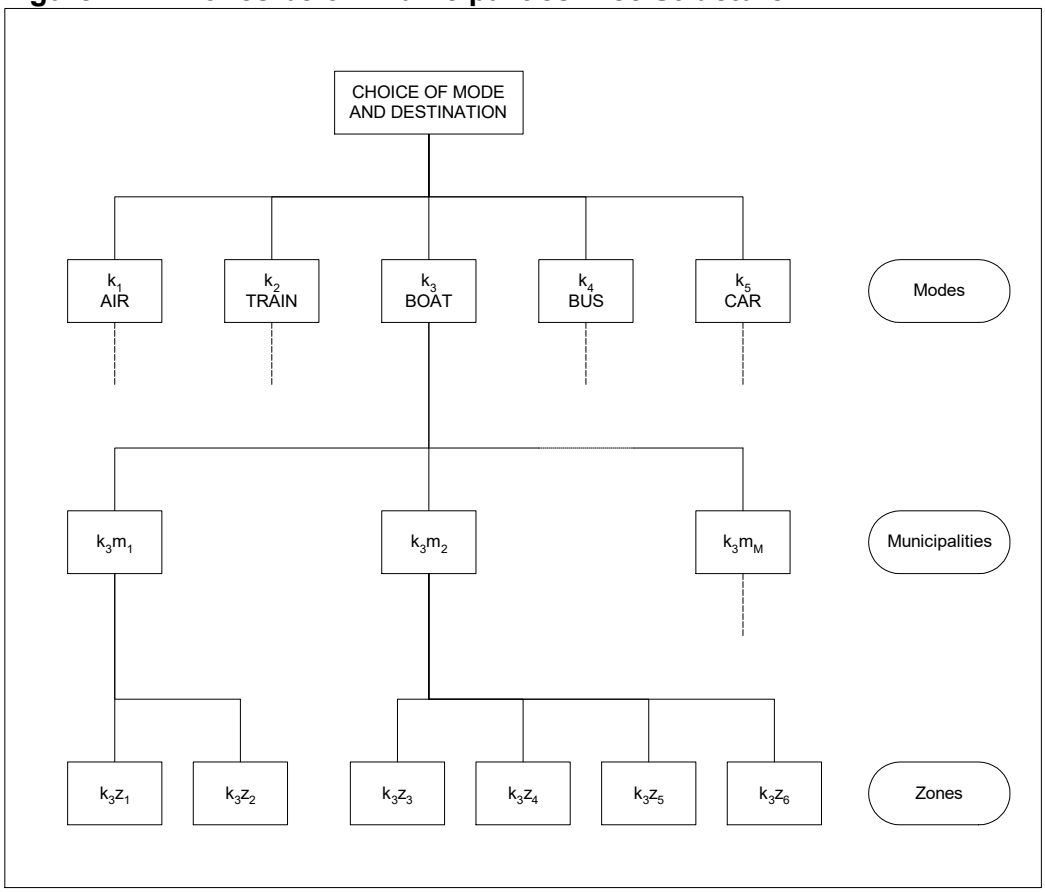
2.3 Extended Tree Structure

The third approach tested retains the detailed destination information in the modelling by using an extended tree structure in the modelling. In this 'zones below municipalities' tree structure, destination choice is represented both at the municipality *and* zonal levels. This approach allows the choice of mode to be defined at the municipality level, while defining the utility functions at the detailed zonal level so that detailed network level-of-service (LOS) measures can be used. Figure 1 overleaf illustrates the approach used.

It can be seen from Figure 1 that mode choice is represented higher in the tree structure than destination choice. The implication of this is that destination choice is more elastic than mode choice. The decision to model mode choice above destination choice was made on the basis of results obtained from simpler models, where the destination was represented at the municipality level only. This assumption was retested using the revised structure described here. For different travel purposes, model tests may reveal a different ordering of the choice decisions give the best model results. Another point to note from Figure 1 is that the number of zones comprising with each municipality varies; in this case zones 1 and 2 comprise municipality 1, and zones 3 to 6 comprise municipality 2. Mapping arrays were used to define the relationship between zones and municipalities.

To implement the tree structure shown in Figure 1 using the ALOGIT estimation software, the *choice* is defined for the *composite alternatives* at the municipality level, whereas the utility functions used to describe the attractiveness of the mode-destination alternatives are defined at the lowest level in the tree, the zonal level. This approach allows both LOS and attraction (size) variables defined at the detailed zonal level to be used directly in the modelling. The model coefficients estimated at the zonal level allow the utilities of each mode-zone alternative to be determined, which in turn allows the probability of mode-municipality choice to be determined. The ALOGIT estimation software therefore estimates the vector of model coefficients that maximises the likelihood for the chosen mode-municipalities.

Figure 1: Zones below Municipalities Tree Structure



where: k = 1, …5 modes
m = 1, …435 municipalities
z = 1, …1428 zones

It should be noted that the tree coefficient linking zones and municipalities in Figure 1 is constrained to one; this constraint is theoretically necessary, as we assume that the definition of the zone boundaries used does not affect the coefficient estimates obtained.

2.4 Comparison of Approaches

The three different approaches were applied to the same data set to investigate which approach gave the best model results. The data set chosen for the investigations was the long-distance tour model specified for the visiting friends journey purpose. The model specification for this purpose had already been determined from earlier model development tests, which were made using LOS defined at the municipality level. This model specification was used *without modification* for each of the three approaches, enabling a consistent comparison of the approaches to be made.

In Table 1, summary model results are presented for four different models:

- VFR: original model, using LOS defined at the municipality level;
- VFR_AVE: used the LOS averaging approach described in Section 2.1;
- VFR_SAMP: used the LOS sampling approach described in Section 2.2;
- VFR_ZBM: used the extended tree structure approach described in Section 2.3.

Note that the first model uses LOS defined at municipality for *both* origin and destination. The three subsequent models use LOS defined at the more detailed zonal level at the origin end with the alternative approximations at the destination end as given above.

Table 1: Comparison of Model Results - Key Model Statistics

	VFR	VFR_AVE	VFR_SAMP	VFR_ZBM
Observations	1134	1155	1155	1155
Final log likelihood	-5773.7	-5956.0	-5971.1	-5936.4
$Rho^2(0)$	0.279	0.272	0.269	0.377

It should be noted that the number of observations in the original VFR model is lower than in the other three models, and therefore the results of this model are not directly comparable. This difference arises because trips with a length of less than 90 km are excluded from the modelling. It is possible for a trip to be made to a destination *zone* which is just over 90 km from the origin, but where the average LOS to the destination *municipality* is just under 90 km. In the original model, the use of municipality LOS means such observations are rejected, and this leads to a lower number of observations.

Comparing the averaging approach to the sampling approach, the averaging approach results in a significantly better log likelihood. However, the best log likelihood and model fit (shown by the $Rho^2(0)$ measure) is obtained using the extended tree structure approach.

Table 2 presents the actual coefficient estimates obtained in the four models.

Table 2: Comparison of Model Results – Coefficient Estimates

Coeff.	Definition	VFR	VFR_AVE	VFR_SAMP	VFR_ZBM
CWDummy	Winter dummy on car	-0.629 (-3.1)	-0.631 (-3.0)	-0.611 (-3.0)	-0.656 (-3.1)
CCDummy	Company car dummy on car	1.361 (2.2)	1.358 (2.1)	1.315 (2.1)	1.426 (2.1)
G_cost3	Cost term for high income households	-0.221 (-1.9)	-0.341 (-3.1)	-0.273 (-2.5)	-0.332 (-3.0)
G_cost2	Cost term for low income households	-0.421 (-3.6)	-0.578 (-5.2)	-0.514 (-4.7)	-0.569 (-5.1)
CA_drlic	Driving licence dummy on car	1.016 (3.9)	1.003 (3.7)	0.980(3.8)	1.053 (3.8)

Coeff.	Definition	VFR	VFR_AVE	VFR_SAMP	VFR_ZBM
CA_avail	Car availability dummy	1.530 (6.1)	1.654 (6.4)	1.605 (6.4)	1.733 (5.9)
G_trfer	Number of interchanges	-0.480 (-3.7)	-0.421 (-3.2)	-0.419 (-3.3)	-0.376 (-2.8)
G_freq	Service frequency	-0.0015 (-2.3)	-0.0013 (-2.1)	-0.0011 (-1.9)	-0.0013 (-2.0)
CA_time	Car time	-0.0058 (-12.6)	-0.0056 (-12.7)	-0.0057 (-13.1)	-0.0055 (-12.7)
G_time	In-vehicle time	-0.0024 (-4.5)	-0.0021 (-3.9)	-0.0023 (-4.2)	-0.0022 (-4.0)
BU_accdis	Bus access and egress distance	-0.063 (-2.0)	-0.052 (-2.1)	-0.028 (-1.8)	-0.049 (-1.9)
BO_accdis	Boat access and egress distance	-0.038 (-4.3)	-0.046 (-4.3)	-0.040 (-4.4)	-0.053 (-3.7)
TR_accdis	Train access and egress distance	-0.016 (-5.1)	-0.018 (-5.5)	-0.015 (-5.2)	-0.018 (-5.3)
AI_accdis	Air access and egress distance	-0.021 (-7.7)	-0.022 (-7.8)	-0.017 (-6.8)	-0.022 (-7.5)
BU_const	Bus mode constant	-0.946 (-2.1)	-1.011 (-2.3)	-1.071 (-2.5)	-1.099 (-2.2)
BO_const	Boat mode constant	0.265 (0.5)	0.326 (0.6)	0.530 (0.9)	0.440 (0.7)
TR_const	Train mode constant	0.238 (0.8)	0.366 (1.1)	0.293 (1.0)	0.381 (1.2)
AI_const	Air mode constant	0.165 (0.4)	0.489 (1.2)	0.083 (0.2)	0.494 (1.2)
SizeMult	Size term – population in destination zone	1.000 (*)	1.000 (*)	1.000 (*)	1.000 (*)
Tr_zn_Mun	Tree coefficient (zones to munic.)	1.000 (*)	1.000 (*)	1.000 (*)	1.000 (*)
nc	Tree coefficient (munic to modes)	0.732 (14.0)	0.690 (15.3)	0.742 (14.1)	0.722 (10.2)

Notes: t-ratios are shown in brackets after the coefficient value
'(*)' denotes that a coefficient is constrained to a fixed value

The magnitude and significance of the coefficient estimates are in general similar between the four models. Interestingly the cost terms are more significant in the three models using zonal LOS, which is likely to result from the greater accuracy with which the true cost of journeys can be approximated using a more accurate zoning system. The averaging and extended tree structure models result in more significant cost terms than the sampling model, this is consistent with the patterns in the overall model fit observed in Table 1. Differences in the policy coefficients of time and cost of up to 25 % are found, illustrating the importance of obtaining the best model to avoid policy biases.

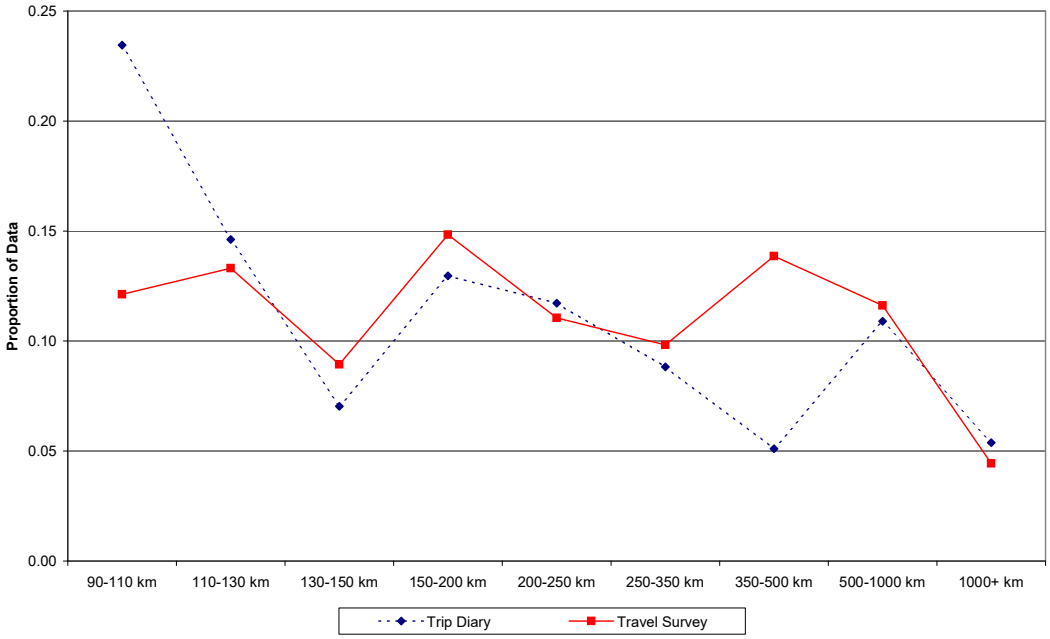
In conclusion, it has proved possible to incorporate detailed zonal destination information in the modelling. While the LOS averaging approach gave better results than the LOS sampling approach, the best model results were obtained using the extended tree structure approach. This structure was also the most convenient for application.

The approach described here was then applied by TØI to estimate the models for all the long-distance travel purposes.

3. DIFFERENTIAL REPORTING RATES

Prior to the model estimation phase, a concern was raised regarding the main data source to be used for estimation, the 1997/98 Travel Survey data. The concern was that differential reporting rates for different distances meant that the distribution of the observed data across different distance bands was not representative of the 'true' trip length distribution. To illustrate this issue, the trip length distribution observed in the 1997/98 Travel Survey data was compared to an independent data source, one-day Trip Diary data collected during 1998 and 2001. The resulting plot is shown in Figure 2.

Figure 2: Travel Survey and Trip Diary Trip Length Distributions



Notes: the x-axis is not linear
the distributions are only based on trips greater than 90 km in length

It appears that the Travel Survey data under-reports short trips in the 90-110 km band, and over-reports longer trips in the 350–500 km band. Both of these patterns are clearly evident from Figure 2. In order to use the travel survey data in the models without bias, it was necessary to adopt a methodology which corrected for these biases. The methodology used was inspired by the WESML correction, developed by Manski and Lerman (1977). For each observation, a weight was defined according to the distance band in which the observation fell, using the formula defined in equation (4).

$$weight = \frac{Q_b N}{N_b} \tag{4}$$

where: b is the trip distance band
Q_b is the 'true' proportion of data in distance band b
N is the total volume of Travel Survey data

N_b is the volume of Travel Survey data in distance band b

The effect of this weight term is to add a lower weight for over-reported distance bands, and a higher weight to under-reported distance bands, to correct for deviations from the true distance band distribution. However, defining the 'true' distance band distribution was a problem. The problem with the Trip Diary data is that the volume of data is low, with a total of 725 observations over 90 km in length, compared to 4733 in the Travel Survey data. Therefore, rather than simply accepting the Trip Diary data as the true distribution, different assumptions were tested by expressing the true distribution as a linear combination of the two data sources using the formula defined in equation (5).

$$Q_b = \mu Q_b^{TD} + (1-\mu) Q_b^{TS} \quad (5)$$

where: μ is varied
Q_b^{TD} is the proportion of Trip Diary data in distance band b
Q_b^{TS} is the proportion of Travel Survey data in distance band b

Three values of μ were tested in the modelling:

- $\mu = 0$: assumes the Travel Survey data is the true distribution
- $\mu = N^{TD}/(N^{TD}+N^{TS})$: a weighted average of the two data sets
- $\mu = 1$: assumes the Trip Diary data is the true distribution

As in Section 2, the tests were made on the VFR model. Table 3 compares the results obtained using the three approaches using a model with level-of-service specified at the municipality level for both origins and destinations. The coefficients listed in Table 3 are defined in Table 2.

Table 3: Comparison of Weighted Model Results

	$\mu = 0$	**$\mu = N^{TD}/(N^{TD}+N^{TS})$**	**$\mu = 1$**
Observations	1134	1134	1134
Log likelihood	-5780.5	-5757.2	-5587.6
Degrees of freedom	19	19	19
$Rho^2(0)$	0.278	0.281	0.302
$Rho^2(c)$	-5.497	-5.515	-5.616
CWDummy	-0.652 (-3.1)	-0.648 (-3.1)	-0.644 (-2.9)
CCDummy	1.267 (2.0)	1.241 (2.0)	1.080 (1.8)
G_cost3	-0.183 (-1.5)	-0.239 (-2.0)	-0.604 (-5.0)
G_cost2	-0.391 (-3.3)	-0.446 (-3.7)	-0.804 (-6.5)
CA_drlic	1.039 (3.9)	1.024 (3.8)	0.968 (3.5)
CA_avail	1.558 (6.1)	1.561 (6.1)	1.648 (6.3)
G_trfer	-0.467 (-3.6)	-0.442 (-3.4)	-0.278 (-2.1)
G_freq	-0.0012 (-2.0)	-0.0013 (-2.0)	-0.0014 (-2.2)
CA_time	-0.0061 (-12.7)	-0.0060 (-12.6)	-0.0059 (-11.6)
G_time	-0.0028 (-5.0)	-0.0028 (-4.9)	-0.0026 (-4.3)
BU_accdis	-0.068 (-2.1)	-0.070 (-2.1)	-0.085 (-2.4)

	$\mu = 0$	$\mu = N^{TD}/(N^{TD}+N^{TS})$	$\mu = 1$
BO_accdis	-0.038 (-4.4)	-0.039 (-4.5)	-0.044 (-5.1)
TR_accdis	-0.017 (-5.3)	-0.017 (-5.4)	-0.021 (-5.8)
AI_accdis	-0.021 (-7.7)	-0.021 (-7.5)	-0.019 (-6.7)
BU_Const	-1.053 (-2.3)	-1.027 (-2.2)	-0.937 (-2.0)
BO_Const	0.234 (0.4)	0.292 (0.5)	0.648 (1.1)
TR_Const	0.205 (0.6)	0.244 (0.8)	0.509 (1.6)
AI_Const	-0.015 (0.0)	0.060 (0.1)	0.523 (1.2)
SizeMult	1.000 (*)	1.000 (*)	1.000 (*)
nc	0.710 (14.6)	0.714 (14.6)	0.690 (15.5)

Notes: t-ratios are shown in brackets after the coefficient value
'(*)' denotes that a coefficient is constrained to a fixed value

Considering first the total model likelihood, the best results are obtained with μ=1, the model run which assumes that the Trip Diary data does indeed represent the true trip length distribution. The improvement in likelihood obtained relative to the μ=0 model is large, at almost 200 likelihood points. Study of the individual coefficient estimates reveals that the main change in the μ=1 model results compared to the other two models is in the cost coefficients G_cost3 and G_cost2. These cost coefficients are both more significant and larger in magnitude (i.e. have more impact on the mode-destination choice decision) in the μ=1 model. Accepting the hypothesis that the Trip Diary data does indeed define the true trip length distribution, then this is a highly plausible result, because travel cost is directly proportional to distance, and by eliminating distance bias, we should be able to get a better estimate of the importance of cost on the mode-destination choice decision.

Figure 3 compares the trip length distributions predicted by the three models to the observed distribution. It should be noted from that the observed Travel Survey distribution in Figure 3 is based on the 1134 visiting friends observations only, whereas the Travel Survey distribution in Figure 2 is based on the 4733 observations made by all purposes.

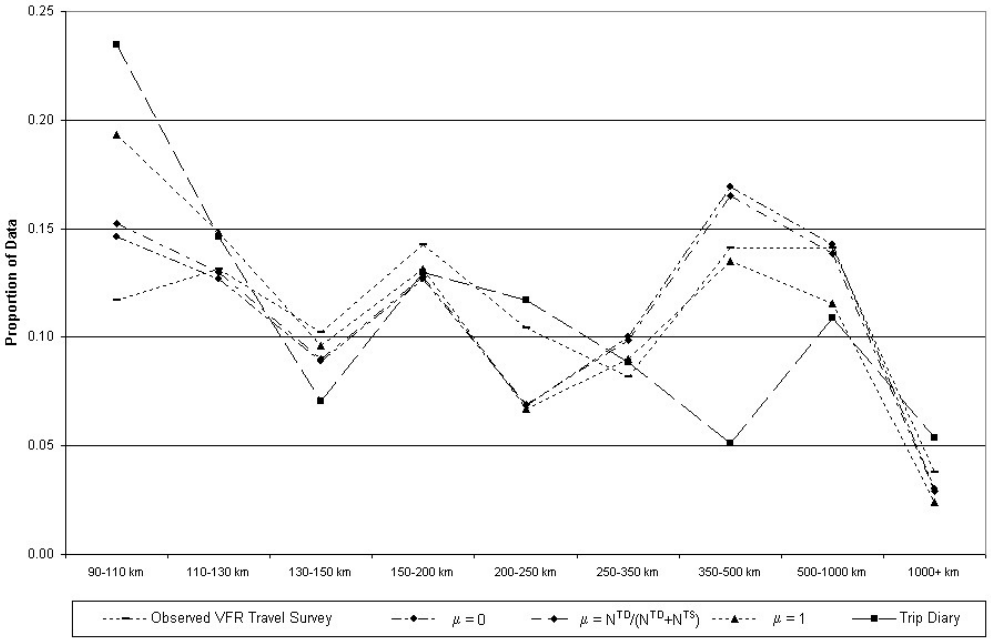


Figure 3: Predicted VFR Trip Length Distributions

Figure 3 demonstrates the predicted trip length distributions moving away from the biased travel survey data, and towards the 'true' trip diary distribution, as the value of μ increases. The effect of the correction is particularly apparent for the 90-110 km band, where the μ=1 distribution (shown in blue) has moved up the y-axis considerably. The plots in Figure 3 demonstrate that the corrected models predict sensible trip length distributions, which overcome the biases in the (uncorrected) Travel Survey data.

In conclusion, the WESML-based weighting correction has been shown to successfully overcome the biases in the observed data. The best model results were obtained when it was assumed that the one-day trip diary data alone best defined the true trip length distribution.

It is important to note that the 'WESML' correction procedure is being used here as an approximate correction and not to give a true maximum likelihood estimation as described in the work of Lerman and Manski. Nevertheless the procedure does appear to give good results and to correct the biases in the survey data.

This procedure was then applied by TØI to the model estimation for all long-distance travel purposes.

4. CONCLUSIONS

The paper has shown how two district advanced modelling techniques could be successfully used to overcome limitations in the Norwegian survey data.

The best approach for overcoming the lack of detail in the destination information was to use an extended tree structure. This approach allowed coefficients to be estimated using more detailed and hence more representative level-of-service, while at the same time allowing choice to be specified at the less detailed municipality level. However, it is interesting to note that using an average of the level-of-service to zones within a municipality gave better model results than sampling a random zone within a municipality.

To investigate suspected biases in the observed trip length distribution, a WESML-type correction was used. A significant improvement in model fit was obtained when the WESML-type correction was used. Furthermore, validation of the trip length distribution predicted by the 'corrected' model demonstrated that WESML-type correction was having the desired impact upon the trip length distribution.

Advanced modelling procedures of these types may well have wider applications in other modelling studies to deal with inconsistent or biased data of various kinds.

REFERENCES

Manski, C. and S. Lerman (1977) The Estimation of Choice Probabilities From Choice-Based Samples. *Econometrica* 45: pp.1977- 1988.

Transport Economic Institute (TØI) and Hague Consulting Group (1990) A model system to predict fuel use and emissions from private travel in Norway from 1985 to 2025, Report to Norwegian Ministry of Transport, TØI and HCG.

ACKNOWLEDGEMENTS

The authors are grateful to Jens Rekdal and Tom Normann Hamre, both formerly of TØI, for their collaboration on this work. Responsibility for any errors, omissions and interpretations is that of the authors alone.