A SIMPLE MODEL VERSUS A COMPLEX MODEL; WHAT DO WE GAIN?

Barry Zondag
Significance
Frank Hofman
Rijkswaterstaat
Jasper Willigers
Significance

The Dutch strategic transport model is a large-scale complex model system. It is often questioned whether a simple model would not do the 'job' with enough quality or just as well. This paper explores the gains from adding this aspect to the search for an optimal model specification.

1. INTRODUCTION

The national transport model of the Netherlands, current version LMS GM4, has been developed and improved in numerous projects over the last three to four decades. In these various model update and improvement studies the model specification of the previous model has often been used as a starting point for each new specification round. As a result of this process the complexity has been built up over different GM (and LMS) versions and the models for the choice of mode of transport, destination, time of the day and access and egress transport are complex in number of explanatory variables and these choices are integrated in an extensive nest structure.

This complexity has made the model less transparent for non-model experts and it is requested by policy makers whether a simpler model would not do the 'job' with enough quality as well. In general, the quality of the current version is good on most aspects. This has been recently confirmed by a back cast project (Significance, 2021) for the period 2004-2018 and an independent academic review.

The purpose of this research work and paper is to provide insight into the added value of the complexity, consisting of number of variables and nesting, of the LMS GM4 model. Second research question is whether and to what extent a different model specification arises if a model specification is built up from scratch. The test work has been performed for two motives: commuting and shopping.

In this paper we will describe the research set-up, the findings for commuting and shopping purposes and conclude with a section on observations and discussion.

2. RESEARCH SET-UP

In this research we try to check on the necessity of the current complexity of the model. In the view of its users the complexity consists because of the large number of variables, especially for different socio-economic segments, and its current nesting structure to address the differences in correlation between alternatives. As it is outside the scope of this research to test and evaluate each variable we have grouped variables by topic and we have set-up a stepwise specification and testing process.

In this research five steps have been identified between a basic model and the current version of the model specification, as following:

1. **Time**: Simple model, multi-nominal logit structure (MNL) and explanatory variables for travel times by mode, alternative specific constants by mode and size variables by purpose;
2. **Cost:** Adding travel costs by mode, including existing travel costs reduction factors and reimbursements, and parking costs. This step also includes estimated costs coefficients by income class;
3. **Spatial variables**: adding spatial variables like intrazonal constants, train variables for short distance, urbanization factors both at origin and destination zones by mode and match between education level of employees and jobs for commuters;
4. **Segment variables**: adding segments variables, mainly dummy variables, to address heterogeneity in mode and destination choices. Variables tested and included are among others age, gender, type of participation, education level, car availability in household and ownership of student card for public transport;
5. **Structure of model**: in this last step a nested logit model structure, instead of MNL structure in step 1 to 4, is freely estimated and tested for significance.

The stepwise testing of the model specification has been executed for the purposes commuting and shopping. Both purpose models have been estimated for each step in Alogit estimation software and applied in so-called apply-models; running and comparing the models with the estimation data (weighted and expanded). The different steps are compared on the 'standard' evaluation criteria, as usually applied, including model fit, average travel distances and trip length distribution by mode, time- and cost elasticities and spatial aspects of particular interest like traffic flows to the 4 main cities in the Netherlands by mode.

3. SIMPLE TO COMPLEX MODEL FINDINGS FOR COMMUTING

The stepwise development of the model specification has been applied to the mode/destination/time-of-day/access and egress transport nested models for commuting. The estimated models are evaluated by step (see section 2) on the 'standard' criteria as presented in the paragraphs below.

3.1 Estimated coefficients

It is unfortunately outside the scope of this paper to present the estimation results for each step and all coefficients. In total the commuting model incorporates 102 coefficients of which 47 are added in step 1, 9 in step 2, 21 in step 3, 20 in step 4 and 5 in step 5. The estimated model after step 5, the 'final' model, is here compared with the estimated model during the last LMS model specification (GM4). This model is not exactly similar as the specification study in steps 3 and 4 has not been exhaustive in this research and we have followed a different sequence of steps.

Therefore, the loglikelihood of the model after step 5 is not yet at the level of the GM4 level but rather comparable. The model also has 10 estimated coefficients less, making the difference of 86.2 points loglikelihood more than bridgeable. Coefficients that appear in both model estimates have the same sign and a similar size. This means that even by taking another route we end up close to the same destination.

3.2 Likelihood ratio tests

The likelihood ratio tests evaluate the significance of adjustments between two model versions. In Table 1 the successive models per step, and for steps 3 and 4 in more detail per sub-step were compared. According to the likelihood ratio, adding 1 coefficient (or D.o.f.) requires a gain of at least 2.0 points loglikelihood (LL) to be significant at 95% level. When adding multiple coefficients, the required gain in loglikelihood increases less than linearly.

The table shows that all adjustments are very significant (p-value is less than 0.05). The biggest jump in loglikelihood occurs by adding the effect of car ownership and SOV card. This means that the representation of the choices made in travel survey are better represented in the model with each additional step. The improvements are substantial and leaving out steps would seriously limit the explanatory power of the model, either influencing the quality of forecasting spatial differences in mode usage and destination choices or simulating the differences in transport behavior by population segment.

Table 1 – Likelihood ratio tests Commuting

Commute	Modelspecification	LL	D.o.f.	dLL	dLL/dD.o.f.
Step 1	Time, ASC's, Size	-128197.4	47		
Step 2	+ Costs	-127049.5	56	1147.9	127.5
Step 3	+ Intrazonal	-126815.4	63	234.1	33.4
	+ Train short distances	-126504.4	64	311.0	311.0
	+ Urbanity	-126352.2	73	152.2	16.9
	+ Education level jobs	-126119.0	77	233.2	58.3
Step 4	+ Car ownership & SOV	-124749.9	84	1369.1	195.6
	+ Level of education	-124625.5	86	124.4	62.2
	+ Age	-124519.5	89	106.0	35.3
	+ Distance segments	-123987.6	94	531.9	106.4
	+ Own E-bike	-123898.1	95	89.5	89.5
	+ Gender	-123864.9	97	33.2	16.6
Step 5	+ Nesting	-123642.7	102	222.2	44.4
GM4	Current model	-123556.5	112		

3.3 Tour length

Comparing the modal split in tours between the survey data and model is less informative as an evaluation indicator as the Alternative Specific Constants in the model will ensure a good fit at least at the National level. And as this modal split in tours is matching well the differences in modal split in kilometers stem entirely from differences in average tour length per mode of transport. In this paragraph we therefore focus on the tour lengths. The average tour length per mode of transport is given for each step in Table 2. The reference values from the NTS (OviN) and the result of the most recent operational model GM4 have been added for comparison.

For most modes of transport, the differences between the successive steps are quite small. Only for the train there are big differences to see. In step 1 and step 2, the average tour length for the train is too small, due to too many short-distance tours. In step 3, short-distance train travel is made less attractive, leading to an increase in the average train tour length. Steps 4 and 5 improve the average tour length of the train even further.

Table 2 – Average tour leg length per mode of transport for commuting

Commute	Train	Car driver	Car passenger	Tram/metro	Bus	E-bike	Bicycle	Walk
NTS	42.9	23.4	21.0	12.2		3.8		1.9
GM4	40.4	24.4	21.4	10.5	13.0	5.3	3.8	1.4
Step 1	32.6	24.3	21.1	11.4	13.4	5.3	3.9	1.4
Step 2	30.6	24.3	21.1	11.2	13.2	5.4	3.9	1.4
Step 3	38.2	24.5	21.1	11.2	13.3	5.3	3.9	1.4
Step 4	39.6	24.7	21.2	11.1	13.0	5.4	3.8	1.4
Step 5	42.5	24.7	21.3	11.0	12.9	5.3	3.8	1.4

The figures below show the trip length distributions for different specification steps for the car (Figure 1) and for the train (Figure 2). To limit the number of lines in the figures, steps 2 and 4 have been omitted from the figures (the impact of these steps is small). For the car you can see that the differences between the other steps are also small. The result of step 5 is almost identical to that of GM4, which is therefore almost impossible to see in the figure. For the train, the difference between step 1 and step 3 is clearly visible. The improvement is also visible between step 3 and step 5. The line of step 5 is again above the line of GM4.

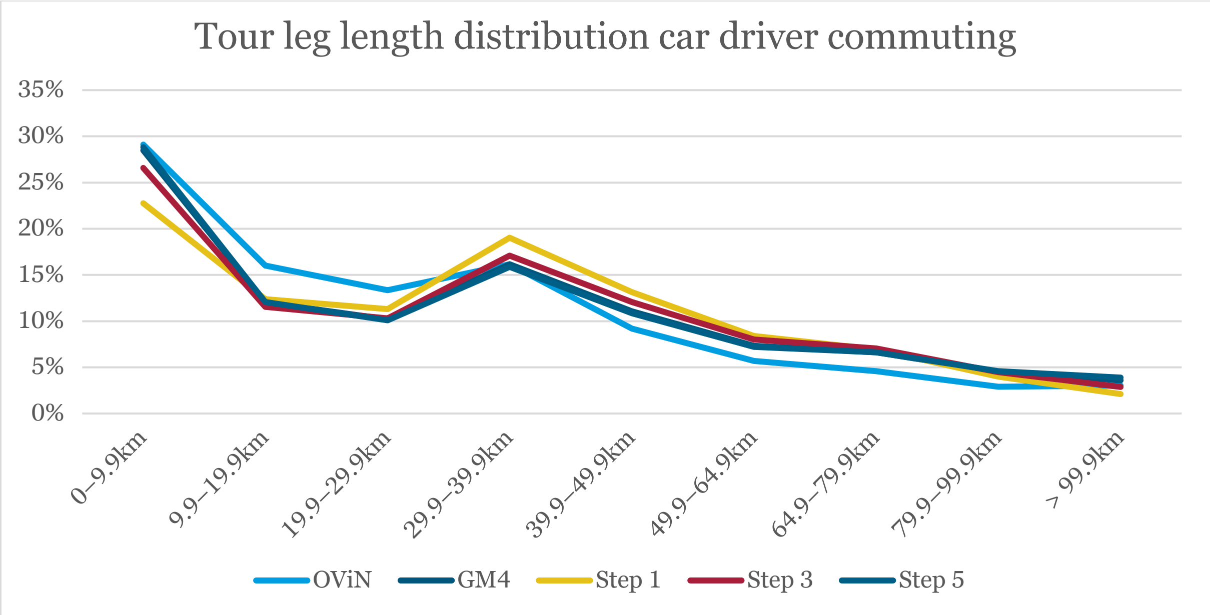

Figure 1 – Distance distribution (tour leg length) for car driver from apply-runs vs. NTS (OviN),

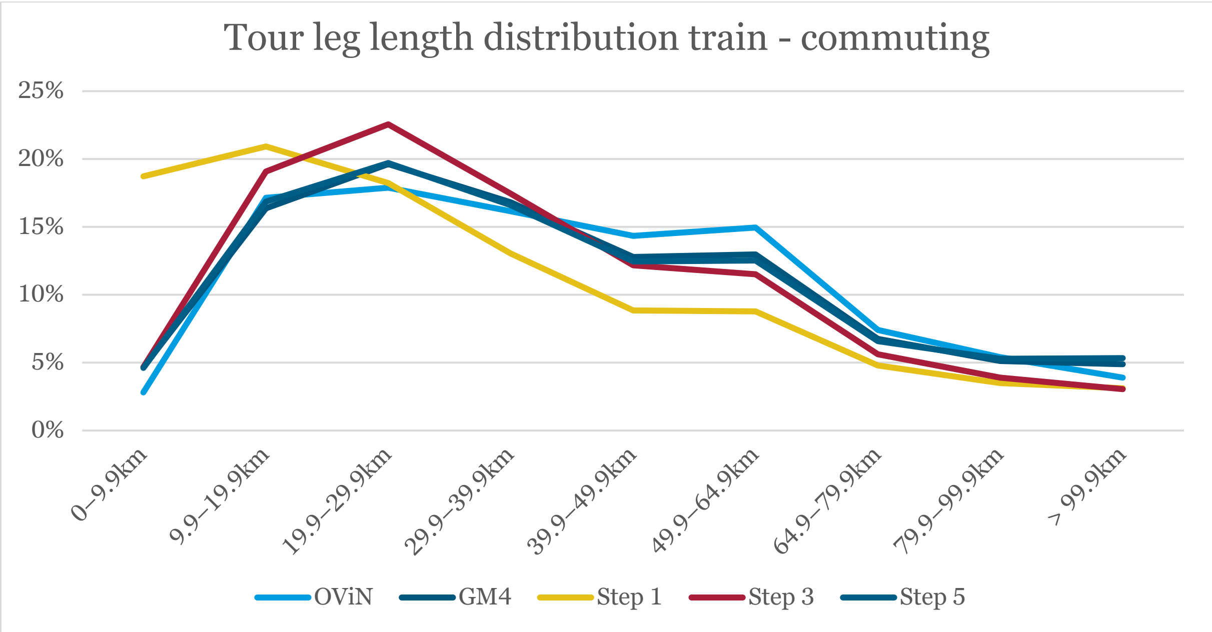

Figure 2 – Distance distribution (tour laying length) for train from apply runs vs. NTS (OviN)

For most modes of transport, except the train, the average lenght of a tour leg can be modelled rather adequatley with a rather simple model including travel times, alternative specific constants and zonal attractors.

3.4 Elasticities

Table 3 4 and Table 5 give respectively the tour elasticities and kilometer elasticities for cost and time per mode of transport[1]. The elasticities can vary greatly between versions, with the earlier steps generally having cost elasticities lower and time elasticities higher than in the comparable estimation of the existing full-scale model. The results for the model of step 5 come close to the elasticities from the GM4 estimate.

Table 3 – Tour elasticity Commuting

Commute	Car fuel costs	Car time		Train costs	BTM cost		BTM IV time	
	Car driver	**Car driver**	**Car passenger**	**Train**	**Tram/metro**	**Bus**	**Tram/metro**	**Bus**
GM4	-0.09	-0.21	-0.38	-0.37	-0.29	-0.34	-0.30	-0.46
Step 1	0.00	-0.69	-0.75	0.00	0.00	0.00	-0.73	-0.90
Step 2	-0.09	-0.52	-0.68	-0.24	-0.22	-0.24	-0.60	-0.78
Step 3	-0.09	-0.50	-0.62	-0.26	-0.22	-0.24	-0.53	-0.76
Step 4	-0.14	-0.36	-0.62	-0.49	-0.40	-0.45	-0.43	-0.66
Step 5	-0.10	-0.22	-0.41	-0.43	-0.33	-0.39	-0.29	-0.49

Table 4 - Kilometre elasticities Commuting

Commute	Car fuel costs	Car time		Train costs	BTM cost		BTM IV tijd	
	Car driver	**Car driver**	**Car passenger**	**Train**	**Tram/metro**	**Bus**	**Tram/metro**	**Bus**
GM4	-0.26	-1.06	-1.09	-0.45	-0.35	-0.41	-0.75	-1.06
Step 1	0.00	-1.66	-1.63	0.00	0.00	0.00	-1.36	-1.62
Step 2	-0.17	-1.43	-1.37	-0.28	-0.24	-0.27	-1.18	-1.47
Step 3	-0.18	-1.43	-1.47	-0.29	-0.25	-0.27	-1.08	-1.44
Step 4	-0.31	-1.29	-1.36	-0.58	-0.46	-0.52	-0.92	-1.32
Step 5	-0.30	-1.10	-1.12	-0.51	-0.40	-0.46	-0.73	-1.10

[1] No train time elasticities can be calculated from the apply runs, because train time is not a separate explanatory variable in this model. Train time is included in the logsums from the station selection model.

In reviewing these elasticities it is always difficult the value these findings as there is no directly observed data on these elasticities. Therefore, at the start of the model development project for LMS GM4 a note has been prepared on expected bandwidths for time and cost elasticities based upon the literature and elasticities from other large scale model systems (de Jong, 2020). This enables us to position our findings in comparison with estimation results from other data sources and findings from the literature.

If we compare the results the elasticities up to step 3 are largely outside the scope of expected bandwidth, the time elasticities are much higher than expected and the cost elasticities are lower than expected. Step 4, adding population segments, and step 5, adding the nesting are essential steps to gather elasticities in line with the expected bandwidth from the literature. In this case the complex model is needed to get 'realistic' elasticities and elasticities from simpler models should be treated more careful especially if those are models are applied to calculate the impacts of travel time or cost changes.

3.5 Main urban areas (G4)

To evaluate how the model specifications perform for the major cities, Table 5 gives the modal split for the highest urban category, for the different steps and for the estimation data. The column on the far right shows the correlation coefficient of the model results relative to the estimation data. Tours to the metropolitan areas are characterized by a relatively low share of car driver and (to a lesser extent) of car passenger and e-bike. The shares of modes of transport of public transport (especially train and tram/metro) and conventional bicycle are relatively high.

The simple step 1 model already gives above-average modal shares for train and tram/metro for the metropolitan areas, due to availability and level-of-service. However, the differences for these modes of transport compared to the national average modes of transport are not yet sufficiently pronounced, and the share of the car is still too high. Adding the cost coefficients significantly improves the modal split. By including parking costs, the share of car drivers decreases, in favour of the other modes of transport.

In step 3, variables for spatial characteristics are added, including dummy coefficients per degree of urbanity based on a specification study. For destinations in central urban areas (urbanity classification 5), there are urban dummies for car driver, train and tram /metro. This further improves the modal split. Steps 4 and 5 do not lead to a major improvement or deterioration across all modes of transport.

Table 5 – Modal split Commuting for destinations in central urban areas

Commute	Train	Car driver	Car passenger	Tram/metro	Bus	E-bike	Bicycle	Walk	Correlation[2]
Data	16.2%	33.2%	1.9%	13.7%	2.9%	1.2%	27.9%	3.0%	
Step 1	11.6%	49.3%	2.5%	10.0%	2.3%	1.4%	20.3%	2.6%	91.1%
Step 2	14.1%	37.5%	3.2%	12.1%	2.9%	1.6%	25.4%	3.2%	98.6%
Step 3	15.9%	32.7%	3.0%	13.5%	3.1%	1.6%	27.8%	2.3%	99.9%
Step 4	16.0%	32.8%	2.9%	13.6%	3.0%	1.5%	28.2%	2.0%	99.9%
Step 5	15.7%	33.4%	3.0%	13.3%	2.6%	1.5%	28.3%	2.3%	99.9%

In summary to improve the modal split for urban destinations the inclusion of travel times and costs (especially parking costs) and urban characteristics and dummies are important aspects of the modelling. The influence of adding population segments and a nested structure on this aspect is much smaller.

4. SIMPLE TO COMPLEX MODEL FINDINGS FOR SHOPPING

The MDToD model for shopping purpose has been re-estimated following the same stepwise process as the model for commuting. In this paragraph we will present the findings for the shopping module in short addressing the same evaluation criteria as before.

4.1 Estimated coefficients

The newly stepwise estimated model result in a small improvement in loglikelihood in comparison with the operational LMS GM module for shopping. The improvement has been realized by a better combination of segment-mode variables. In total the commuting model incorporates 86 coefficients of which 42 are added in step 1, 2 in step 2, 16 in step 3, 22 in step 4 and 4 in step 5.

4.2 Likelihood ratio tests

All adjustments, new steps, are very significant (p-value is less than 0.05) and result in substantial improvements in the loglikelihood. This means that the representation of the choices made in travel survey are better represented in the model with each additional step The biggest jump in loglikelihood occurs in step 2, adding costs coefficients, and in step 4 especially by adding the effect of car ownership and SOV card. These findings are in line with findings for the commute model.

[2] Correlation coefficient of the modal split in the model result relative to the estimation data

4.3 Tour length

The average tour leg length does not differ much for the various estimation steps. Most modes seem reasonable except the train for which the tour leg length is underestimated. For this purpose, including complexity is not improving the average tour leg length for the train (for commuting it does). This probably has to do with the small amount of observations on train travellers for the purpose shopping.

4.4 Elasticities

The estimated time and cost elasticities following a stepwise procedure are after step 5 rather comparable, with exception of BTM (Bus, Tram, Metro), with the existing LMS GM4 model. The elasticities by step show a huge variation in elasticities between step 1 and step 5. The very high time elasticities for tours and kilometres (for km close or above 2) from step 1 are reduced by each step towards more expected elasticities in step 5 (between -0.6 and -0.8 for km in step 5). For the cost elasticities the additional steps are needed to increase the cost elasticities form a rather low level for commuting. The steps in addition to a simple travel time based model seem essential to derive more realistic elasticities, especially step 4 socio-economic segmentation, seem to play an important role.

4.5 Main urban areas

In the first step, a travel time-based model, there is a sharp difference between the observed and modelled modal split for the four main urban areas in the Netherlands. Adding cost information in step 2, especially on parking costs, and spatial variables in step 3 does improve the modal split substantially for the urban areas. The exceptions are car passenger, which remains overestimated, and BTM (Bus, Tram, Metro) which remains underestimated. For these modes probably additional variables are needed to improve their fit. Step 4, socio-economic segmentation, and step 5, nesting, seem to have little impact on the modal shift for urban areas.

5. OBSERVATIONS AND DISCUSSION

Each of the steps results in a very substantial increase of the model fit, which means that the more advanced models give a better estimate at the individual level of the travel choices made. If we compare the model specification in step 5 with the operational LMS GM4 the differences are relatively small. This confirms that, if the same data is used describing transport behaviour, NTS data, and as explanatory variables, the model specification process most likely ends up in a rather similar model.

Comparing the steps with more aggregated reference values, step 1 gives already a reasonable fit for average travel distances and trip length distribution (except the train). The inclusion of step 2, cost, and step 3, spatial variables, are important to improve the match with the observed train trip length distribution and traffic flows by mode to the 4 main urban areas. Without including step 2 and 3 both reference values are poorly simulated by the step 1, travel time only, model. Adding the segment variables, step 4, and nesting structure, step 5, are critical steps to improve the time- and cost elasticities by mode. Without these steps the elasticities for step 1 to 3 are mainly outside the 'acceptable' range, drawn from the literature, for these elasticities. The time elasticity tends to become very high in the simple models in combination with very low-cost elasticities, especially for public transport.

In table 6 we try to illustrate the policy implications of our findings for a few aspects of interest to policy makers.

Table 6 – Illustrative overview of policy interest and required complexity of model

Policy interest	Model aspect	Complexity of model
Transport forecast at national level	model share in tours and km	Rather simple model might be enough, travel time based + limited segment information
Infrastructure policy	Changing travel times	Complex model needed including travel cost, sufficient socio-economic segmentation and nesting to get plausible time elasticities
Travel cost policies	Changing travel costs	Complex model needed including travel costs, sufficient socio-economic segmentation and nesting to get plausible cost elasticities
Region specific studies	Focus on improving accessibility in urban areas	Complex model needed including parking costs and spatial variables – probably extension of current model needed to improve this for all modes of transport

The table above is not complete and serves as an example that simplifying the models does have an impact on the quality of the results and such ambition should be considered carefully. The research framework of this study offers a valuable structure to evaluate the impact of potential changes to the model. Bottom-line might be that the wish for simple/transparent and high-quality models might remain a utopia and that in practice each time a difficult trade-off needs to be made between these aspects.

In this study we did not check whether the estimation of segmentation variables is specific for the data set used. In the future a check of the performance of the model or of the robustness of the model parameters might be realised using a hold –out sample.

BIBLIOGRAPHY

de Jong GC, Bandwidths for elasticities and values of time based on a review of the literature, The Hague 2020 (memo 19012-M04)

Significance, Backcast LMS vergelijking prognose en waargenomen ontwikkeling, The Hague 2021

TNO, Review LMS/NRM, TNO 2022 R10341, The Hague, The Netherlands 2022 (https://puc.overheid.nl/doc/PUC_718907_31)