CALIBRATION AND VALIDATION OF

MICROSCOPIC TRAFFIC FLOW MODELS

Elmar Brockfeld *
Institute of Transport Research, German Aerospace Center
Rutherfordstrasse 2, 12489 Berlin, Germany
phone: +49 30 67055 231, fax: +49 30 67055 202
email: elmar.brockfeld@dlr.de

Reinhart D. Kühne
Institute of Transport Research, German Aerospace Center
Rutherfordstrasse 2, 12489 Berlin, Germany
phone: +49 30 67055 200, fax: +49 30 67055 202
email: reinhart.kuehne@dlr.de

Peter Wagner
Institute of Transport Research, German Aerospace Center
Rutherfordstrasse 2, 12489 Berlin, Germany
phone: +49 30 67055 237, fax: +49 30 67055 202
email: peter.wagner@dlr.de

For Presentation and Publication
83rd Annual Meeting
Transportation Research Board
January 2004
Washington, D.C.

Submission date: November 15th, 2003

WORDS: 4414
Plus 3 Figures (750)
Plus 3 Tables (750)
TOTAL: 5917

**Corresponding author*

Abstract. Microscopic simulation models are becoming increasingly important tools in modeling transport systems. There is a large number of models used in many countries. The most difficult stage in the development and use of such models is the calibration and validation of the microscopic sub-models such as the car following and gap acceptance models. This difficulty is due to the lack of suitable methods for adapting models to empirical data. The aim of this paper is to present recent progress in calibrating a number of microscopic traffic flow models. Ten very different models have been tested using data collected via DGPS-equipped cars (Differential Global Positioning System) on a test track in Japan. To calibrate the models, the data of the leading car are fed into the model under consideration and the model is used to compute the headway time series of the following car. The deviations between the measured and the simulated headways are then used to calibrate and validate the models. The calibration results agree with earlier studies as there are errors of 12 % to 17% for all models and no model can be denoted to be the best. The differences between individual drivers are larger than the differences between different models. The validation process gives acceptable errors from 17 % to 22%. But for special data sets with validation errors up to 60% the calibration process has reached what is known as "overfitting": because of the adaptation to a particular situation, the models are not capable of generalizing to other situations.

INTRODUCTION

For the simulation of traffic flow various macroscopic and microscopic models exist (see (*1*) and (*2*) for an overview) and there is a large number of available models used in many countries. Nowadays, in the time of computer processors increasing rapidly in speed, especially microscopic models become very important tools in modeling transport systems. In the development of these models it is important to check the models against reality, namely to calibrate and validate them. Usually the developers of the models do this on their own using some data sets they have access to and publish the results obtained (see (*3*) for an overview on model calibration). This way every model is calibrated and sometimes validated with other data sets, but if a user has to decide which model to take for some special application, he is not really able to compare them and to choose the best one. So there seems to be a lack in the field of benchmarking these models.

Previous studies in testing models have been performed using single car data recorded on a rural road in the USA (*4, 5, 6*) analyzing the travel times of vehicles between observers/detectors. The aim of this paper is to test some models from a more microscopic point of view by analyzing the car following behavior of the models in detail. The analyses are done using data from experiments conducted in Japan (*7*). By calibrating and validating all models using the same data sets, the models are directly comparable to each other which is a first step of developing a transparent benchmark for these models.

THE DATA SET AND THE SIMULATION SET-UP

The data set

The data used for the calibration and validation of the models have been recorded on a test track in Japan in October 2001 (*7*). Eight experiments have been conducted, where nine cars drove on a 3 km test track (2 x 1.2 km straight segments and 2 x 0.3 km curves) for about 15-30 minutes in each experiment following a lead car, which performed some driving patterns. These are for example driving with constant speeds of 20, 40, 60 and 80 km/h for some time, varying speeds (regularly increasing/decreasing speed) and emulating many accelerations/decelerations as they are typical at intersections. The regularly increasing/decreasing of speed is done performing half, single, double and triple waves on the two straight segments of length 1.2 km on the test track. That means for example a half wave is starting with 40 km/h, accelerating to 60 km/h on the middle of a segment and decelerating to 40 km/h at the end of it. A single wave is accelerating from 40 to 60 at the first quarter of the segment, decreasing to 40 km/h in the second quarter and 20 km/h in the third quarter, and accelerating to 40 km/h in the fourth quarter.

To minimize driver-dependent correlations between the data sets, the drivers were exchanged between the cars after each experiment. Having all cars equipped with the differential global positioning system DGPS, the position of each car is stored in 0.1 second intervals throughout each experiment. From these data other important variables like the speed, the acceleration and the headway between the cars were extracted for simulation purposes. The accuracy of the DGPS is about 1 cm and the appointment of the speeds has got an error of less than 0.2 km/h as described in (*7*). Thus, the data sets have got such a high resolution that they are adequate for the analysis of car-following behavior and calibration of car-following models.

Simulation setup

In this paper we present analyses concerning four of the eight experiments, namely the patterns with intervals of constant speeds and driving patterns with wave-performing. The data which have been used for the simulations had time series of about 26 minutes in the first experiment, 25 minutes in the second, 18 in the third and 14 in the fourth experiment. For each of the four experiments one gets the ten trajectories of the cars in form of the DGPS-positions and speeds. From these the accelerations and distances/gaps between the cars have been calculated, which are used for the simulation runs. While the distances could be directly calculated from the GPS positions, the calculation of the acceleration was done deriving it from the speeds by a Savitzky-Golay smoothing filter (*8*), which smoothes the values in intervals of one second with a second-order polynomial. This procedure was necessary because some models need the acceleration of the car driving ahead and of course, these values should not fluctuate too much.
The study was done by analyzing the time-development of the gaps between the cars. For the simulation setup only two cars are considered at a time. The leading car is updated as the speeds in the recorded data sets tell and the following car is updated as defined by the equations and rules of the used model, respectively. Typically, an equation like the following was used:

$$\begin{aligned} \dot{v} &= f(g, v, V, \{p\}) \\ \dot{g} &= V - v \end{aligned} \qquad (1)$$

where v and V is the speed of the following and the leading car, respectively, and g is the headway between the cars. The symbol $\{p\}$ denotes a set of parameters of the model under consideration.

ERROR MEASUREMENT AND OPTIMAL PARAMETER FINDING

The absolute error a model produces with a particular parameter set for a special vehicle pair is calculated via the simple distance between the recorded gaps and the simulated gaps between each vehicle-pair. To get a percentage error it is additionally related to the average gap in each data set:

$$e = \frac{\frac{1}{T}\sum_{t=0}^{T}\left|g^{(sim)}(t) - g^{(obs)}(t)\right|}{\frac{1}{T}\sum_{t=0}^{T} g^{(obs)}(t)}, \text{ with } t \in \{0; 0.1; ...; T\}, \qquad (2)$$

where $g^{(sim)}$ and $g^{(obs)}$ are the simulated and the observed gaps between the cars. The error is calculated for one particular vehicle pair in one particular experiment, thus T is the total time of each experiment (26, 25, 18 and 14 minutes).

Altogether 4*9=36 vehicle pairs (four experiments, each with nine vehicle pairs) were used as data sets for the analyses of the car following behavior. Each model has been calibrated with each of the 36 different constellations separately gaining optimal parameter sets for each “model-data set” combination. To find the optimal parameter constellations in the calibration procedure a gradient-free optimization method known as the “downhill simplex method” (*8*) was used and started many times with different initialization values for each “model-data set” pair. The variation in initialization is done to avoid sticking with a local minimum, which of course can occur because getting a global minimum can not be guaranteed by those optimization algorithms. Subsequently, the validation was performed in two ways. First, a mainly driver-independent procedure: the errors are determined simulating the second, third and fourth experiment with the calibrated parameter values resulting from calibration of the data sets of the first experiment. Second, a driver-dependent validation was done focusing on a special pair of drivers, which drove one after another in each of the experiments.

THE MODELS

By now, ten microscopic models of very different kind with 4 to 15 parameters have been tested (see tab1e 1 for some details about the parameters):

- CA0.1 (cellular automaton model by K. Nagel, M. Schreckenberg) (*9*),
- SK_STAR (model based on the SK-model by S. Krauss) (*10*),
- OVM (“Optimal Velocity Model”, Bando, Hasebe) (*11*),
- IDM (“Intelligent Driver Model”, Helbing) (*12*),
- IDMM (“Intelligent Driver Model with Memory”, Helbing, Treiber) (*13*),

- Newell (model by G. Newell (*14*, *15*), can be understood as the continuous CA with more variable acceleration and deceleration),
- GIPPSLIKE (basic model by P.G. Gipps) (*16*),
- Aerde (model used in the simulation package INTEGRATION) (*17*),
- FRITZSCHE (model used in the british software PARAMICS; it is similar, but not identical to what is used in the german software VISSIM by PTV) (*18*),
- MitSim (model by Yang and Koutsopulus, used in the software MitSim) (*19*).

As the time step for the models should be 0.1 seconds according to the recorded data, some models with a traditional time step of 1 second – as for example used for simple cellular automatons - have been modified to adopt for an arbitrarily small time-step. Thus, every model is simulated with a time step of 0.1 seconds.

The most basic parameters used by the models are the car length, the maximum speed, an acceleration rate (except for the CA0.1-model) and a deceleration rate (for most models). The acceleration and deceleration rates are specified in more detail in some models depending on the current speed or the current headway to the leading vehicle. Furthermore, some models (CA0.1, SK_STAR and MitSim) use a parameter for random braking or another kind of stochastic parameter describing individual driver behavior. Most models use something like a reaction time of the drivers to the behavior of the leading car.

With these kinds of parameters seven of the ten models are covered completely, except for the IDMM, MitSim and FRITZSCHE. The IDMM has as a special feature a memory effect. Depending on the density ahead, the cars try to hold their speeds according to a rolling horizon. The MitSim model defines two thresholds concerning the headway, which cause a switching between three different driving modes. Especially if a driver is very close to the leader the calculations become very sophisticated, depending on the headway, own speed, speed-difference and the current density. In addition to the basic simulation update equation (1) the model needs the speed of the leader one time step before as a special feature. The FRITZSCHE model provides switching to various driving modes, too. For this model the switching depends not only on the headway (g), but also on the speed-difference (dV) between the follower and the leader. Thus, a (dV,g)-car following plane is divided into different regions of free driving, approaching, emergency brake and two other driving behaviors. As a specific, differing to equation (1), the model needs the acceleration of the follower and the leader one time step before and uses some kind of “brake light” of the leader by reacting on its deceleration.

SIMULATION RESULTS

Calibration Results

The ten models have been calibrated independently for each of the four experiments (denoted as “11”, “12”, “13”, “21”) with all nine vehicle pairs separately (data sets named “11_1 D1-D2”, “11_2 D2-D3”,…, “21_1 D9-D10”, where D1, … ,D10 are the drivers). Thus, for each model 36 optimal parameter sets and calibration errors were obtained.

The detailed explanation of the best parameter sets neglecting, an overview on all errors for all models and data sets is shown in figure 1. The most important thing to remark is that the differences between the models are very small. For some data set one model is the best, for another data set another, thus no model seems to outperform the others regularly. As can be seen in table 2, the differences between the models (indicated by the amplitude, which is the difference between the best and the worst models error for a particular driver pair in an experiment) are less than 3 percentage points for the most data sets. Only some data sets – as for example 12_2 with 5.028 percentage points and 13_1 with 4.769 percentage points - seem to be very special, so that some models perform well but others have big problems to calibrate them. The average amplitude of 2.567 % can be understood as a measurement for the diversity of the models.

Focusing on the level of the errors for the various data sets it can be seen in figure 1 that for most data sets errors of 12-17 % occur frequently. In special cases – as for example data set 11_8 or 13_2 – the errors are reduced approximately to 10 %, which is surprisingly good. On the other hand there are some data sets letting the models produce errors of about 20 % up to 23 % (data sets 12_2 and 21_1). By calculating the amplitude of the errors each model produces with all 36 data sets, it can be seen in table 3, that they are very big with values of 10.11 % (SK_STAR) up to 12.78 (GIPPS_LIKE). Interpreting these values as a measurement for the diversity of the data sets – and thus of the drivers, too - it can be stated that the diversity in the behavior of the drivers is much bigger

than the diversity of the models under investigation. Of course, the amplitudes of the models results are hardly influenced by extreme values, which are rare in these results. But taking only the obviously interval of 12-17%, where most of the errors are (see figure 1), the diversity of the drivers would be 5 percentage points which is still much higher than the diversity of the models.

Looking further at the mean error values of 15.14 % (GIPPS_LIKE) to 16.20 % (IDM) the different models produce (see table 3), again no model can be denoted to be the best. Especially the two models with a big number of parameters (FRITZSCHE and MitSim) do not provide better results in general than the simpler models.

Validation Results 1 (driver-independent)

For the validation purpose for each model the nine optimal parameter sets obtained by the calibration of the data sets of the first experiment "11" were taken to simulate each model with the other three data sets "12", "13" and "21". In more detail, for each model the optimal parameters obtained by calibrating 11_1 were taken to simulate the data sets 12_1, 13_1 and 21_1. Then for each model the parameters of 11_2 were taken for 12_2, 13_2 and 21_2 and so on. Thus, for each of the calibration results of experiment "11" three validations were conducted as shown in figure 2.

Most of the errors are between 17 and 22 %, which means an additional error of about 5 percentage points in comparison to the error of 12-17 % after the calibration. Except for some special cases, again, there seem to be no general differences between the models. As a very special case the results of the fourth driver pair (data set 11_4) seem to be very abnormal because of very high errors of about 32% for 12_4, 40-55% for 13_4 and 40 % for 21_4 for most models. This is a case known as "overfitting". Obviously the reason is that all models (except for the OVM) adopted very special parameter sets for data set 11_4 during the calibration and thus the other data sets 12_4, 13_4 and 21_4 could not be simulated appropriately with these "overfitted" parameter sets. Likewise abnormal appears data set 21_9 with errors of 35 % to 41 %. In relation to this the data sets 12_9 and 13_9 with errors of 18-20% and 16-18% are well reproduced by the parameters obtained during the calibration of 11_9, thus the parameter set of 11_9 seems to be realistic. But this parameter sets is not able to describe the behavior of the driver pair in data set 21_9.

One more thing to mention is that especially the Aerde model (high errors in data set 13_3 and 21_1 for example) and the OVM model (high errors in data set 13_6 and 21_6 for example) sometimes differ to the "level of errors" produced by the other models. The mean errors of the models after the validation (see table 3) are very similar with most models having errors between 19.25 % (SK_STAR) and 20.72 % (IDM). Only the Aerde model and the OVM model show slightly more problems during the validation with mean errors of 23.13 % and 22.82 %. (Note, that for the representative calculation of the mean values the errors of the "overfitted" data sets x_4 and x_9 are excluded!)

Comparing the validation results with the calibration results ("validation mean" subtracted from "calibration mean" in table 3), the validation produces about 3.2 % to 5.5 % additional error for most of the models, which is quite good. According to the problems mentioned above, the Aerde model and the OVM model have a bit worse values of 7.63 and 6.63 %. It seems to be interesting, that the MitSim model has one of the best values with 3.75% although the number of parameters is very big. This is remarkable because of course the calibration of the many parameters is quite sophisticated, but in this case the results and parameter sets obtained seem to be valid.

Validation Results 2 (driver-special)

The idea of the driver-special validation is to analyze the results obtained from a special driver pair which drove subsequently in each of the experiments. The basic question is, whether the models are able to describe this driver pair better than in the driver-independent validation. Because the order of the nine drivers following the leading car has been changed during the experiments, there is only one driver pair driving subsequently in any experiment. This is the case for the drivers D9 and D10 with the data sets 11_9, 12_9, 13_9 and 21_9. For these four data sets the parameter sets obtained by the calibration are taken and validated with each of the three other data sets. So every data set 11_9, 12_9, 13_9 and 21_9 is validated three times.

As can be seen in figure 3, the data set 21_9 is a very special one, confirming the results for it during the calibration. Simulating it with the parameters obtained by the calibration of the other data sets, the errors for it are always much bigger than the other results (bars at the right side of the first three diagrams). Simulating the

other data sets with the calibration parameters of data set 21_9 (bottom diagram of figure 3), the errors are much higher, too.

The other results obtained look quite good as the errors are only slightly higher than in the calibration cases. This can be seen, too, in table 3, comparing the mean calibration error of each model to the mean errors of this validation 2 (the difference between "validation2 mean" and "calibration mean" is shown in the last row). Interestingly, the OVM model is the best in this case with an additional error of only 0.41 percentage points. But also the other models produce only 1.41 percentage points (MitSim) up to 2.43 percentage points (Aerde) additional error. Apart from the fact that the used data for this analysis of driver-dependent validation is not so comprehensive, it can be stated that the validation using data sets from the same drivers is much better than using those of different drivers. Of course, this result is of special interest, because the obtained calibration results will not help to get parameter sets which can easier be generalized. But it gives an insight, how good data sets of special drivers can be validated. This probably sets a lower limit for the errors well generalized parameter sets are able to produce.

CONCLUSIONS

The error rates of the models in comparison to the data sets during the calibration for each model reach from 9 % to 24 %. Surprisingly, no model appears to be significantly better than any other model and the average error rates of the models are very close to each other between 15.1% and 16.2%. All models share the same problems with certain data sets while other data sets can be reproduced quite well with each model. Interestingly, it can be stated that models with more parameters do not necessarily reproduce the real data better. The results of the validation process give a similar picture. The additional errors in comparison to the calibration are – apart from singular cases of "overfitting" - mainly in the area of 3 to 5 percentage points. Using different data sets from the same drivers for calibration and validation, the additional validation errors mainly reduce to 1.5 to 2.5 percentage points.

The results after the calibration and the validation agree with results that have been obtained before with a completely different data set taking the travel times on road segments instead of headways for the error measurement (*3*). (In these studies about 15 to 27 % were found to be the minimum calibration error and additional validation-errors were found to be about 2 to 5 percentage points. It was found, too, that out of about ten models the differences are not as big as could be expected.) However, the results of the validation show, that when calibrating and validating with special data sets, the parameters of a model can be "overfitted" and thus the results can be very unsatisfactory with surprisingly high errors. The calibration tends to optimize the model for a given data-set, thereby sacrificing generality.

Concerning the generalization of data recorded on test tracks to real life it is known that on test tracks the driving behavior is much more careful. As a typical indicator the time headway of consecutive cars is about 1.0 to 1.5 seconds on highways (*20*) and about 2.5 seconds in experimental situations like on test tracks and in driving simulators. The same values of about 2.5 seconds are found for this data set, but - most important – the distribution of time headways shows a qualitative agreement to real data (*21*). Thus, for the analyses performed in this study it seems to be possible to generalize the driving behavior in the experiment to real situations.

There are two conclusions that can be drawn. First, one should call for the development of better models. Additionally, one should think about a different calibration technique which avoids "overfitting" and could produce results which stay more general. The other way to interpret the results is that – from this microscopic point of view – errors of about 15-25 % can probably not be suppressed no matter what model is used. These are due to a really stochastic component in the driver's behavior.

Finally, if one would centralize these results and take them as given reality, the recommendation would be to take the simplest model for a particular application, because complex models likely will not produce better results. The only reason a complex model could be preferred would be, if the user is very familiar with the model and knows the consequences for its behavior whatever a parameter (or set of parameters) is changed. But the results obtained should be confirmed by testing microscopic models with much more different data sets than in this contribution to get a more precise insight what the models are able to describe and which error rates probably have to be accepted.

ACKNOWLEDGEMENTS

Thanks are to G. S. Gurusinghe, T. Nakatsuji, Y. Azuta, P. Ranjitkar and Y. Tanaboriboon for producing these very interesting data sets and giving us access to these.

REFERENCES

1. D. Chowdhury, L. Santen, and A. Schadschneider, *Statistical Physics of Vehicular Traffic and Some Related Systems*, Physics Reports **329**, 199 (2000). [See also at arxiv.org/abs/cond-mat/0070053]
2. D. Helbing, *Traffic and Related Self-Driven Many-Particle Systems*, Reviews of Modern Physics **73**(4), 1067-1141 (2001). [See also at http://arxiv.org/abs/cond-mat/0012229]
3. M. Brackstone and M. McDonald, *Car-following: a historical review*, Transportation Research F **2**, 181-196 (1999).
4. E. Brockfeld, R. D. Kühne, A. Skabardonis, P. Wagner, *Towards a benchmarking of Microscopic Traffic Flow Models* TRB 2003 Annual Meeting (CD-ROM), TRB2003-001164. To be published in TRR 2003.
5. See http://www.ce.berkeley.edu/~daganzo/.
6. Karen R. Smilowitz & Carlos F. Daganzo, *Experimental Verification of Time-Dependent Accumulation Predictions in Congested Traffic*, Transportation Research Record **1710**, 85 – 95 (2000)
7. G. S. Gurusinghe, T. Nakatsuji, Y. Azuta, P. Ranjitkar, Y. Tanaboriboon, *Multiple Car Following Data Using Real Time Kinematic Global Positioning System*, TRB2003-004137.
8. W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C*, chapter 10, Cambridge University Press and references therein (2002).
9. K. Nagel and M. Schreckenberg, J. Physique I, 2221, **2**, 1992.
10. S. Krauss, P. Wagner, and C. Gawron, *Metastable states in a microscopic model of traffic flow*, Phys. Rev. E **55**, 5597 – 5605 (1997).
11. M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, *Dynamical model of traffic congestion and numerical simulation*, Physical Review E, 239-253, **51**, 1999.
12. M Treiber, A Hennecke and D Helbing, *Congested traffic states in empirical observation und numerical simulations*, Physical Review E, 1805-1824, **62**, 2000.
13. M. Treiber, D. Helbing, *Memory effects in microscopic traffic flow models and wide scattering in flow-density data* , see at http://arxiv.org/abs/cond-mat/0304337.
14. G. Newell, *Theories of instabilities in dense highway traffic*, J. Op. Soc. Japan **5**, 9 – 54 (1962).
15. G. Newell, *A simplified car-following theory – a lower order model*, Transportation Research B **36**, 195 – 205 (2002).
16. P.G. Gipps, *A behavioral car following model for computer simulation*, Transp. Res. B **15**, 105 – 111 (1981).
17. Brent C. Crowther, *A Comparison of CORSIM and INTEGRATION for the Modeling of Stationary Bottlenecks*, Master thesis, Virginia Polytechnic Institute and State University (2001).
18. H.-T. Fritzsche, *A model for traffic simulation*, Transp. Engin. Contr. **5**, 317 – 321 (1994).
19. Kazi Iftekhar Ahmed, Modeling Drivers' Acceleration and Lane-Changing Behavior, Ph.D. thesis, MIT (1999).
20. M. Brackstone, Beshr Sultan and Mike McDonald, *Motorway driver behaviour: studies on car following*, Transportation Research F **5**, 329-344 (2002).
21. P. Wagner and I. Lubashevsky, *Empirical basis for car-following theory development*, available at http://arxiv.org/abs/cond-mat/0311192.

LIST OF FIGURES

LIST OF TABLES

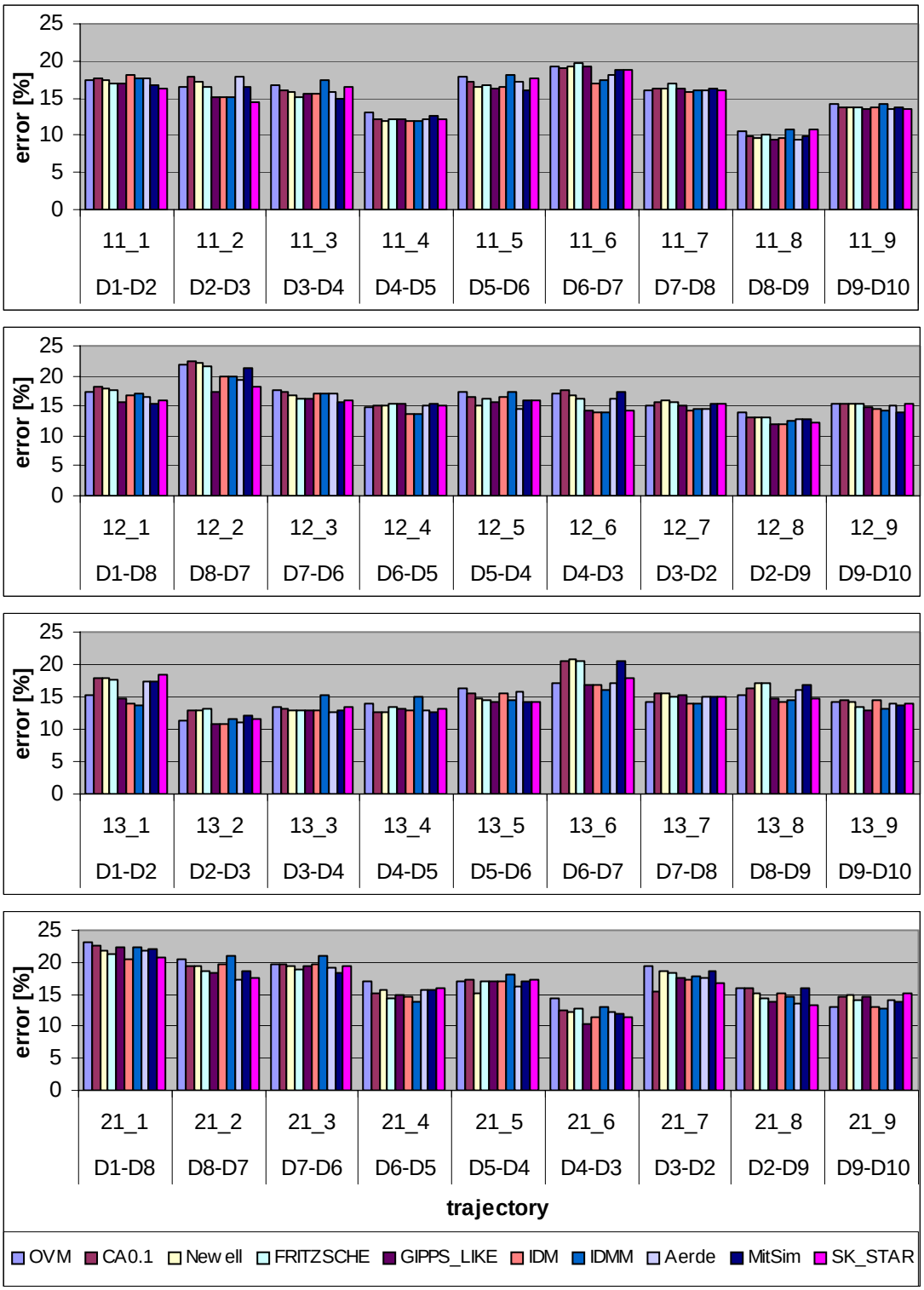

FIGURE 1 Errors of the models after the calibration (data sets of four experiments (11, 12, 13, 21) with 9 driver pairs; D1…D10: drivers).

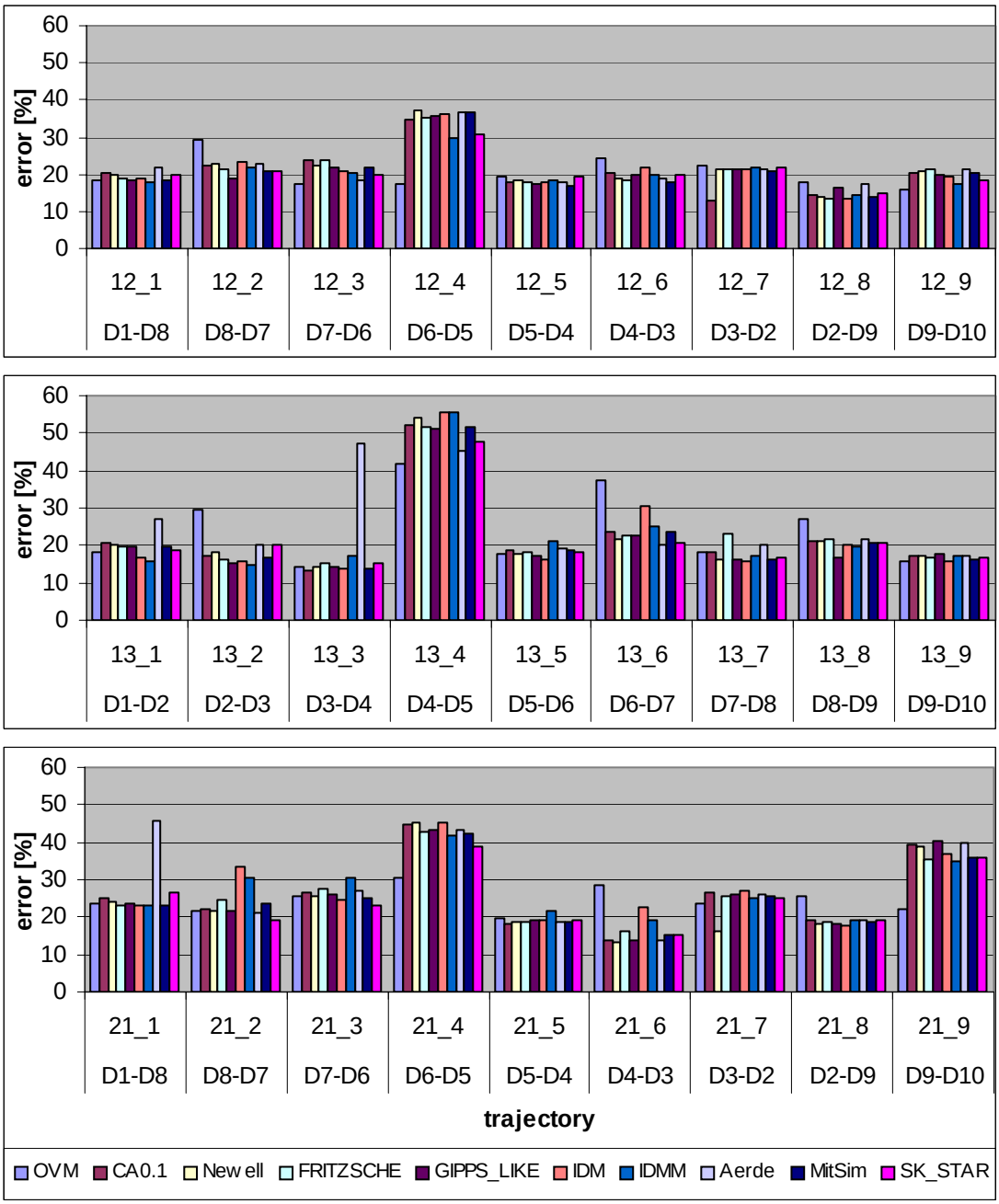

FIGURE 2 Errors of the models after the driver-independent validation1 using the best parameter sets of experiment 11 (D1…D10: drivers).

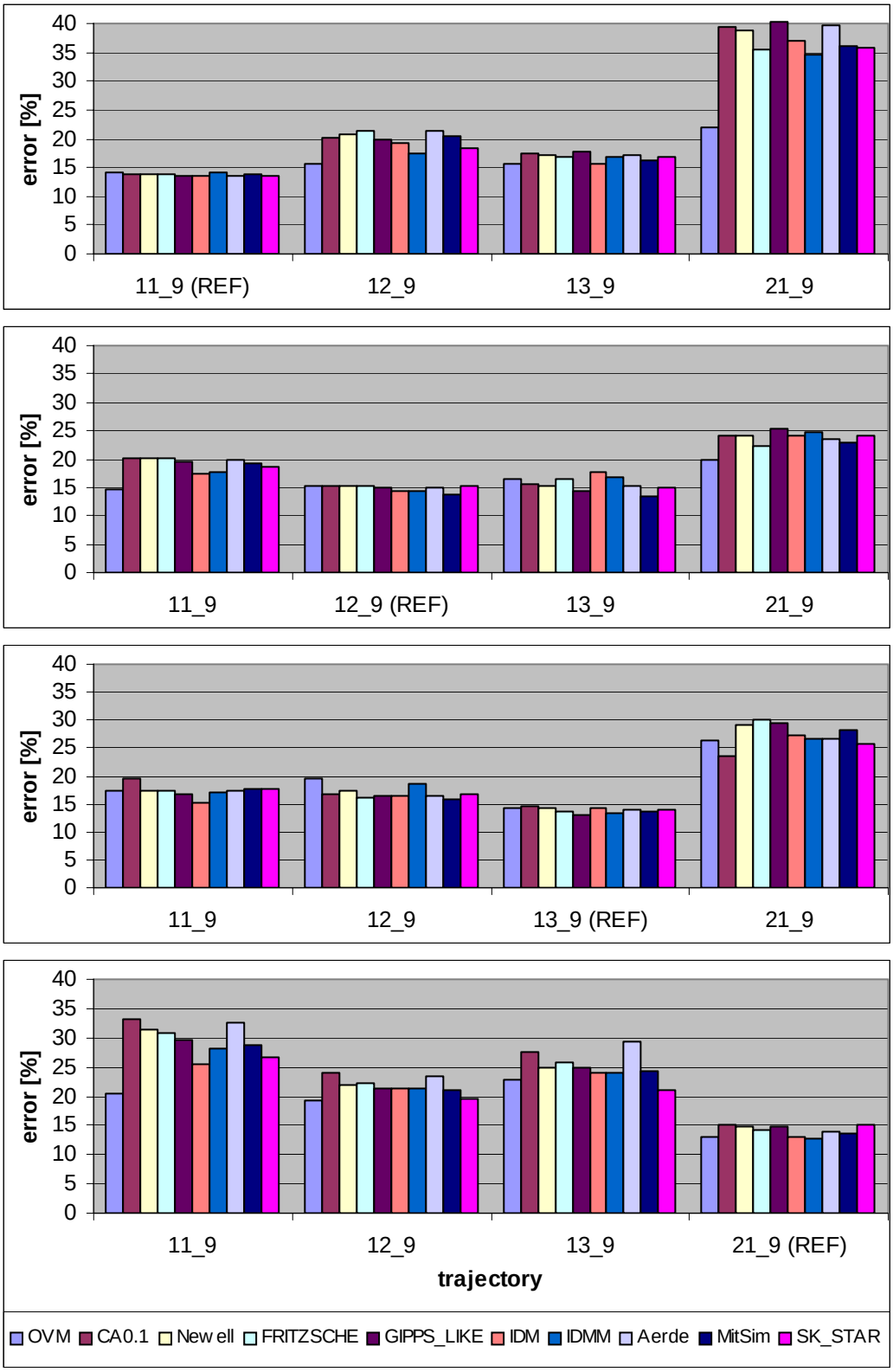

FIGURE 3 Errors of the models after the driver-special validation2 of the driver pair D9-D10 ("REF" denotes the data set from which the optimal parameter constellations after the calibration were taken).

TABLE 1 Short description of the parameters used for the models under investigation (the following parameters are used by all models: maximum speed *Vmax*, vehicle length *l*; used by all models except for CA0.1: acceleration *a*)

Model (acronym)	Main additional parameters	Amount of parameters (+ for internal calculation)
CA0.1	***p*** random braking ***T*** distance keeper to leader	4
Newell	***p*** random braking ***tau*** reaction time ***T*** distance keeper to leader ***b*** deceleration	7
FRITZSCHE	***bMax, bMin*** limits for deceleration ***bNUll*** acceleration/deceleration for a "following mode" ***tD*** for calculating a "safe gap" ***tS, tR, kPlus, kMinus, fX*** thresholds and variables describing different driving modes ***A0*** measurement how near the cars tend to be; threshold for hard braking (**at** acceleration one time step before) (***an*** acceleration of the leader one time step before) (***bn*** kind of "brake light" of the leader [0;1])	13 (+3)
GIPPSLIKE	***B*** deceleration ***invTauT*** strength of acceleration ***tau*** reaction time	6
IDM	***b*** deceleration ***delta*** exponent decreasing the acceleration dependent on the speed ***tau*** reaction time	6
IDMM	***b*** deceleration ***tauAdapt*** adaptation time for holding recent speeds depending on the density (-> memory effect) ***T0*** netto time gap between successive cars ***betaT*** strength of the adaptation	7
Aerde	Cars try to reach a desired gap ***h2*** for calculation of desired gap at maximum speed ***C2, C3*** for calculation of desired gap dependent on the speed	6
MitSim	***h1, h2*** thresholds for gap-measurement, which causes switching to different driving modes ***ap, vp, gp, kp*** parameters determining the acceleration in the case of undercritical gaps and a quicker leading car ***am, gm, km*** parameters determining the deceleration in the case of undercritical gaps and a slower leading car ***dvpm*** measurement how near the cars tend to be. ***bMax*** maximum deceleration ***epsA*** additional random accel./decel. (***an*** acceleration of the leading car one time step before)	15 (+1)
OVM	***S0*** for calculating a preferred gap	4
SK_STAR	***bMax*** maximum deceleration ***invTauT*** strength of acceleration ***tau*** reaction time ***r*** threshold for random braking (***gStar*** auxiliary variable to memorize desired gaps)	7 (+1)

TABLE 2 Statistical overview on the errors produced by all models after the calibration (in [%])

driver pair	data set	BEST MODEL	WORST MODEL	Amplitude (WORST - BEST)	MEAN	STDERROR
D1-D2	11_1	16.219	18.016	1.797	17.315	0.527
D2-D3	11_2	14.491	17.888	3.397	16.237	1.207
D3-D4	11_3	15.020	17.408	2.388	15.965	0.739
D4-D5	11_4	11.911	13.001	1.090	12.209	0.332
D5-D6	11_5	16.046	18.066	2.020	17.006	0.689
D6-D7	11_6	16.943	19.624	2.681	18.672	0.889
D7-D8	11_7	15.923	16.891	0.968	16.229	0.293
D8-D9	11_8	9.423	10.814	1.391	10.046	0.499
D9-D10	11_9	13.435	14.246	0.811	13.773	0.252
D1-D8	12_1	15.476	18.049	2.573	16.826	0.917
D8-D7	12_2	17.307	22.335	5.028	20.420	1.758
D7-D6	12_3	15.518	17.482	1.964	16.698	0.667
D6-D5	12_4	13.569	15.326	1.757	14.823	0.658
D5-D4	12_5	14.609	17.362	2.752	16.142	0.873
D4-D3	12_6	13.902	17.628	3.726	15.733	1.493
D3-D2	12_7	14.276	15.793	1.517	15.131	0.542
D2-D9	12_8	11.837	14.019	2.183	12.726	0.633
D9-D10	12_9	13.871	15.374	1.503	14.921	0.533
D1-D2	13_1	13.573	18.369	4.796	16.388	1.817
D2-D3	13_2	10.805	13.192	2.387	11.801	0.913
D3-D4	13_3	12.604	15.162	2.558	13.240	0.713
D4-D5	13_4	12.594	15.127	2.533	13.216	0.780
D5-D6	13_5	14.102	16.189	2.087	14.895	0.742
D6-D7	13_6	16.065	20.715	4.650	18.410	1.900
D7-D8	13_7	13.934	15.464	1.530	14.844	0.579
D8-D9	13_8	14.261	17.056	2.796	15.671	1.100
D9-D10	13_9	13.016	14.563	1.547	13.887	0.513
D1-D8	21_1	20.524	23.082	2.558	21.818	0.785
D8-D7	21_2	17.211	20.897	3.687	19.052	1.186
D7-D6	21_3	18.303	20.884	2.581	19.409	0.673
D6-D5	21_4	13.782	17.023	3.241	15.249	0.932
D5-D4	21_5	15.178	18.202	3.023	16.948	0.776
D4-D3	21_6	10.493	14.381	3.888	12.227	1.063
D3-D2	21_7	15.481	19.432	3.950	17.752	1.100
D2-D9	21_8	13.392	15.982	2.590	14.788	0.966
D9-D10	21_9	12.771	15.249	2.477	14.014	0.869
	all			2.567	15.680	

TABLE 3 Average errors of the models in [%] after completing the calibration and the driver-independent validation process (validation 1 without the "overfitted" data sets x_4 and x_9; validation 2 without the "overfitted" data set 21_9)

		OVM	CA0.1	Newell	FRITZSCHE	GIPPSLIKE	IDM	IDMM	Aerde	MitSim	SK_STAR
Calibration	BEST	10.58	9.95	9.68	10.15	9.49	9.75	10.81	9.42	9.95	10.69
Calibration	WORST	23.08	22.52	22.22	21.72	22.26	20.52	22.23	21.68	21.98	20.80
Calibration	Amplitude (WORST-BEST)	12.50	12.57	12.54	11.58	12.78	10.78	11.41	12.26	12.03	10.11
Calibration	MEAN	16.20	16.18	16.03	15.92	15.14	15.16	15.58	15.50	15.71	15.39
Validation 1	MEAN	22.82	19.72	19.29	20.28	19.25	20.72	20.67	23.13	19.46	19.74
Validation 1	VAL1 MEAN - CAL MEAN	6.63	3.54	3.26	4.36	4.12	5.56	5.09	7.63	3.75	4.35
Validation 2	MEAN	16.60	18.21	17.96	18.03	17.46	16.97	17.44	17.93	17.12	17.16
Validation 2	VAL2 MEAN - CAL MEAN	0.41	2.02	1.93	2.12	2.32	1.81	1.87	2.43	1.41	1.77