Graz Symposium Virtual Vehicle (GSVF) 2019

Towards a Generally Accepted Validation Methodology for Sensor Models - Challenges, Metrics, and First Results

PHILIPP ROSENBERGER[1], JAN TIMO WENDLER[2], MARTIN HOLDER[1],
CLEMENS LINNHOFF[1], MORITZ BERGHÖFER[1],
HERMANN WINNER[1], AND MARKUS MAURER[2]

[1]*Institute of Automotive Engineering (FZD), Technische Universität Darmstadt*
{rosenberger, holder, linnhoff, winner}@fzd.tu-darmstadt.de
moritz.berghoefer@web.de
[2]*Institute of Control Engineering (IfR), Technische Universität Braunschweig*
{wendler, maurer}@ifr.ing.tu-bs.de

Abstract

In order to significantly reduce the testing effort of autonomous vehicles, simulation-based testing in combination with a scenario-based approach is a major part of the overall test concept. But, for sophisticated simulations, all applied models have to be validated beforehand, which is the focus of this paper.

The presented validation methodology for sensor system simulation is based on a state-of-the-art analysis and the derived necessary improvements. The lack of experience in formulating requirements and providing adequate metrics for their usage in sensor model validation, in contrast to e.g. vehicle dynamics simulation, is addressed. Additionally, the importance of valid measurement and reference data is pointed out and especially the challenges of repeatability and reproducibility of trajectories and measurements of perception sensors in dynamic multi-object scenarios are shown. The process to find relevant scenarios and the resulting parameter space to be examined is described. At the example of lidar point clouds, the derivation of metrics with respect to the requirements is explained and exemplary evaluation results are summarized. Based on this, extensions to the state-of-the-art model validation method are provided.

Keywords: Model Validation, Perception Sensor Simulation, Safety Validation, Autonomous Driving

1 Introduction

In the past few years, several research projects have been established to work on safety validation of automated driving [1, 2]. Simulation-based testing in combination with a scenario-based approach [3, 4] is a major part of the overall test concepts, as it promises to reduce the testing effort significantly [5]. For sophisticated simulations, all applied models have to be validated beforehand, as well. Since fidelity criteria [6] as well as requirements [7] are progressing, the final objective is to provide a generally accepted validation methodology. Within the PEGASUS project [1], first steps towards such a methodology have been performed and are presented in this work. The methodology is derived from a state-of-the-art analysis [8] and the derived improvements in the validation method are exemplarily applied to vehicle dynamics models in [9].

It has been clarified that validity cannot be proven generally, but sample validity in a sophisticated way that allows generalization over a broad parameter space is the reachable goal of the validation process [9]. The mentioned sample validity (German: "Stichprobenvalidität" [9]) means that the validation has not been falsified through a systematic and objective process with empirical samples. Now, the lack of experience in formulating requirements and providing adequate metrics for the usage in sensor model validation, in contrast to e.g. vehicle dynamics simulation, is addressed. Additionally, the importance of valid measurement and reference data is pointed out and especially the challenges of repeatability and reproducibility of trajectories and measurements of perception sensors in dynamic multi-object scenarios are shown. The process to find relevant scenarios and the resulting parameter space to be examined is described. At the example of lidar point clouds, challenges with metrics with respect to the requirements are explained and exemplary evaluation results are summarized. Based on this, extensions to the state-of-the-art model validation method are provided.

2 State-of-the-Art Analysis and Derived Improvements in the Validation Method

To the knowledge of the authors, there is no generally accepted definition of the term sensor model. Mainly, a distinction is made regarding the interface of the models (e.g. [10]) not taking the modeling itself into account. For example, in [10], the categories statistical, object-based models and models relying on the geometry and material of the environment are used.

Therefore, Sec. 2.1 defines the term sensor model and its manifestations. In Sec. 2.2, state-of-the-art visual inspection as a validation instrument is described as motivation to overcome this unsatisfying state. In consequence, the recently published, sophisticated method for simulation model validation, called "objective quality assessment by statistical validation" [9], is shortly summarized to explain its different aspects later in this work.

2.1 Definition of the Term Sensor Model and its Manifestations

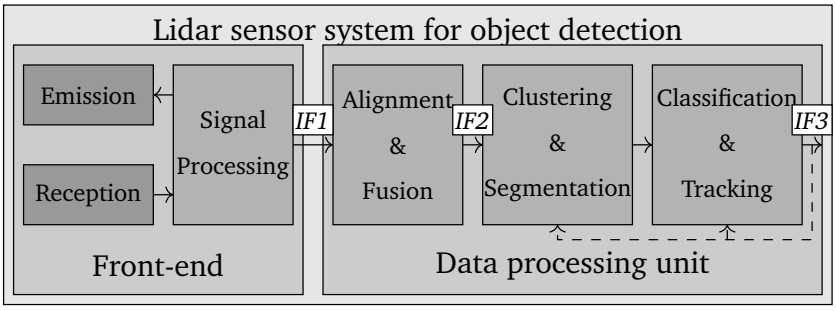

Figure 1: Lidar sensor system for object detection from [11], IF: interface

Before entering the validation concept, a few definitions are presented that are required to specify the scope of this paper and have not been explicitly defined before, to the knowledge of the authors. In general, a model is an approximation of reality that can be used for different aspects, e.g. prediction, development or testing of subsequent processing steps. In literature, there is a categorization of sensor models by the (output) interface (e.g. [10]) but not by the modeling method itself. This does not cover the diversity of sensor models. In the field of (perception) sensor data simulation, the terms sensor model and sensor system simulation have to be distinguished.

Table 1: Classification of sensor data generation by implementation, GT: Ground Truth, MR: measurement range

	GT "Models"	Idealized Models	Phenomenological Models	
			Stochastic Models	Physical Models
Principle	Transformation of global "GT" in sensor perspective "GT"	Perfect MR, (only values the sensor can actually measure)	Stochastic (probabilistic & statistic) for modeling observable effects	Modeling of signal propagation, reflection, diffraction, refraction, absorption, transmission, reception, ...
Possible specifications	None, (only transformation)	Position, orientation, ideal MR	False detections (FP/FN), noise, atmosphere, dirt, manipulation, ...	Wave lengths, material properties, surfaces, signal processing, ...
Sensor accuracy	Perfect	Perfect	Realistic overall stochastic	Realistic single measurements
Complexity	Very small	Very small	Small	Very high to infinitive

We define the term **sensor model** as a model that generates data available at the front-end of the real sensor after digitization and quantization, named signal processing in Fig. 1. Clearly, information is already reduced at this point, as e.g. thresholds are applied and counter resolution is limited. This is done with respect to the overall data size as signal processing has the goal to enable to buffer and transfer the resulting data for the first time.

The described split implies that there is no data processing (e.g. clustering, segmentation, classification and tracking) within the front-end. The actual outcome depends on the sensor technology, e.g. lidar, radar or camera. For example, a lidar sensor model can have the interface raw scan (which is transformed into a (3d) point cloud within the next processing step). Additional downstream data processing as indicated by Fig. 1 leads to the term **sensor system**. Not all stages of the illustrated data processing must be present.

There are different manifestations of sensor models and sensor system simulations. For example the simulation tools from dSPACE [12], ESI Group [13], IPG Automotive [14], TASS International [15], TESIS [16], TWT [17], or Vires Simulationstechnologie [18] offer manifestations with different names. This is done to distinguish the models in complexity and realism. In contrast to the simulation tool and model manufacturers, we define the term ground truth "model" and phenomenological model and also redefine the term idealized model. Table 1 gives an overview of the defined terms and used principles, possible specifications, modeled sensor accuracy, and complexity. The Table can be used for sensor models and sensor system models.

Ground truth "models" only contain the transformation of all information from world coordinates to sensor-specific coordinates, while no information is discarded. Therefore, the term "model" is set in quotes to stress the fact that no information is changed. The actual sensor (system) model with the lowest complexity and realism can be summarized as idealized sensor model. The term "perfect model" would be misleading, as an idealized model is a representation of a perfect sensor, but it is not a perfect model in most cases. An **idealized sensor system model** outputs e.g. an ideal object list that can contain all objects within the specified measurement range (MR), but does not consider occlusion of objects. We define **idealized sensor models** as a generation of features at the front-end. As mentioned previously, for a lidar sensor model, this means that a raw scan is generated. The samples are for example generated from a simple ray casting algorithm that simulates a perfect beam (i.e. a beam with an infinitely small diameter). The beams are parameterized according to the specification sheet of the sensor (e.g. number of layers/detectors, its measurement

range and resolution in range, azimuth and elevation, its update rate, and its minimum and maximum range of detection). Therefore, only simple geometries are modeled in an idealized sensor model, while all possible phenomena during beam emission, propagation, and reception are neglected. **Phenomenological models** can be modeled partly physically and partly stochastically. In contrast to idealized models, effects are modeled to generate more realistic measurements regarding consistency in stochastic or single measurement performance. **Stochastic models** have a data-driven approach and combine probabilistic and stochastic for modeling observable effects. A **physical model**, by contrast, uses equations derived from physics modeling effects, e.g. for signal propagation and reflection. If the equations have parameters that are not fundamental physical constants, a combination of both physical and sophistical modeling can be used. The accuracy of a model is defined by its precision and trueness, as defined in ISO 5725-1 [19]. Therefore, a phenomenological model is the best approach that combines the precision of the stochastic approach and the trueness (for single measurements) of the physical approach. A potential disadvantage is that the reproducibility and repeatability of scenarios and measurements can be affected by the stochastic modeling. This aspect will not be discussed in more detail here.

Finally, the term **(perception) sensor data generation** is used, whenever the simulation tool that performs environment and movement simulation, as input for the sensor system simulation, is implied, as well.

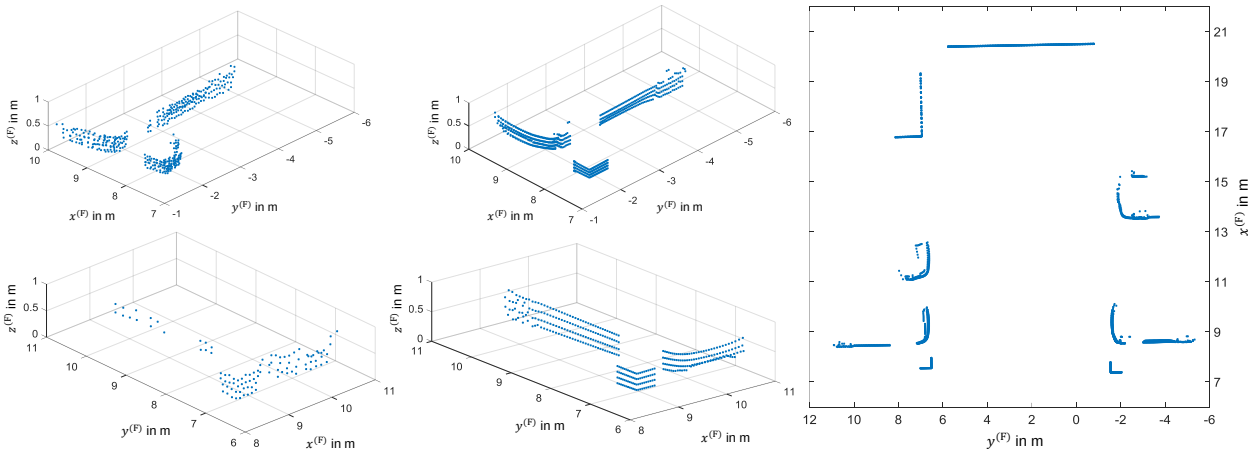

Figure 2: Visual inspection of fused Cartesian lidar point clouds of two different cars behind small obstacles from [20], left: real sensors, right: simulated sensors, rightmost: complete scene in simulation

(a) Real experiment (b) Re-simulation

Figure 3: Pictures of the performed parking scenario, from [20]

2.2 Objective Quality Assessment by Statistical Validation

Actual simulation model validation is mostly based on visual inspection of plots, performed by experts, e.g. [21]. A comprehensive summary of the state-of-the-art is given in [8]. As an example, Fig. 2, taken from [20], shows such a visual comparison of real and simulated Cartesian point clouds for two different cars. $x^{(\mathrm{F})}, y^{(\mathrm{F})}, z^{(\mathrm{F})}$ denote the Cartesian coordinates of the points with respect to the ego vehicle. Both cars are partly occluded by a cuboid-shaped obstacle at their front corner, as can be seen at the far right of Fig. 2. In addition, Fig. 3, again from [20], shows the performed simple parking space scenario in reality and simulation. The real point cloud of the car at the top left of Fig. 2 looks different compared to the simulated one. The point clouds of the car below, instead, look more similar. Nevertheless, without adequate metrics and experience in their interpretation, any type of validation stays questionable.

To overcome this unsatisfying state, Viehof has introduced a method for objective quality assessment of simulation models by statistical validation [9]. The method has already been applied in practice to vehicle dynamics simulation. It is based on the implication that simulation validation is divided into model and parametrization validation. At first, a sensitivity analysis over the parameter space that is given by the requirements on the model has to be performed. The sensitivity analysis shows all configurations that need to be tested for a selected number of repetitions and then have to be re-simulated for comparison. We will discuss this systematic approach now with respect to the simulation of perception sensor systems like lidar and radar and stress important extensions and changes.

3 Formulating Requirements for Sensor System Simulation

As already contained in the classic "V"-model, the first step within the development process including verification and validation is to collect the requirements for the function or system to be developed. In relation to these requirements on the top left of the "V", acceptance tests, located at the top right, can be seen as requirements validation. Viehof discusses the requirements shortly, but considers them as already given when model development and especially validation starts [9]. Additionally, in the final step of the overall statistical validation process, while results are checked at the end, requirements are analyzed to a smaller extent. This means to have a look at the required accuracy for specific metrics to determine, if it can be reached at all with respect to the actual reference data accuracy. Finally, in the case of sensor data generation for simulation-based safety validation, acceptance tests (or validation of requirements) should prohibit fatalities happening to the customer or occupants. Therefore, in summary, in the case of perception sensor system simulation, a methodology for requirements validation is necessary for the overall safety validation test architecture, while not in the scope of this particular work. To get to a proper set of requirements, we propose a stepwise approach that is actually derived from the categories defined in Sec. 2.1 and Tab. 1:

1. At first, the function or system under test (SUT) with generated perception sensor data defines the output(s) of the sensor system simulation by its own input(s).
2. Having these, the requirements engineer needs to have a catalogue of possible phenomena that can be observed on real measurement data at the selected sensor system output(s).
3. Now, the system under test has to be analyzed in detail and its sensitivity with respect to the listed possible phenomena needs to be determined. As an example, object tracking and classification is mostly insensitive to noise on point clouds, but highly sensitive to simple features like length and width of L-shapes [7].
4. The next step is to define whether a stochastic or physical approach should be used to describe the selected phenomena in particular. Here, the required fidelity and accuracy of the sensor data generation should be considered.
5. Finally, the actual accuracies of the different effects or phenomena should be determined.

It must be pointed out here that there is overall only limited experience in the field of perception sensor data generation in simulation and especially with the objective to be used in simulation for safety validation of automated driving. Therefore, formulating requirements is a totally new field of research compared to e.g. vehicle dynamics simulation. So, the already mentioned catalogues of possible phenomena and effects on perception sensor system data are not completed, yet, which substantiates the difficulty in validation of such a simulation.

4 Relevant Scenarios and the Resulting Parameter Space

Having the requirements at hand that describe the output interface of the generated perception sensor data with respect to phenomena on the data as well as physical effects that have to be considered, scenarios can be determined for collecting the measurement and reference data for parametrization and validation. We use the term scenario, short for concrete scenario [22], in this scope, as it is synonymous to parameterization or configuration, but more common in the case of safety validation of autonomous driving, which is the overall objective of the performed research. This definition includes all atmospheric parameters to be integrated in the scenario description, besides all properties of all static and dynamic objects within. It also includes different configurations of the sensor system (e.g. sampling rate or resolution), even if this is not directly implied in the term scenario at first. Here, limitations of the environment simulation come into play. At first, the phenomena and effects depend on a set of parameters of the static environment, of the atmosphere and of all dynamic objects within the scene. Now, static and dynamic parameters that can be considered by the environment simulation have to be collected as a first step of the overall perception sensor data generation. All relevant and considered parameters together constitute the parameter space for the actual experiments to be performed. This space gets huge quite fast, which makes additional methods to shrink the parameter space inevitable.

Viehof proposes a sensitivity analysis over the parameter space before planning measurements to reduce test effort in a sophisticated way [9]. Thereby, the sensitivity analysis reduces the testing effort by identifying scenarios with which falsification of the simulation is most likely, even if it seems questionable at first to use the sensor simulation itself to design its validation experiments. Specifically, in [9, p. 63ff.] the extended Fourier Amplitude Sensitivity Test (eFAST) method is proposed to be selected as sensitivity analysis method for simulation validation, as it is based on a frequency analysis of a systematically selected sample scenario. To find the relevant scenarios, the following questions have to be answered [9, p. 65f.]:

1. Relevance of changeable parameters in real experiments → Which parameters should be varied according to the sensitivity analysis and in which range?
2. Required degree of statistical validation → How granular should the parameter space be resolved? How many scenarios or configurations should be inspected?
3. Practicability → How many scenarios or configurations are feasible to be inspected or performed?

When designing scenarios for validation, it is important to check that they are different from those used for parameterization. Therefore, an analysis with respect to the parameter space coverage must be performed in order to verify that the definition "*Two concrete scenarios are equal if their respective parameter combination is situated in the same volume cell of the common parameter space.*" [23, p. 4] is falsified by all scenarios used for validation with respect to the ones used for parameterization. Additionally, there must be a metric for the scenarios that shows the difference in between such cell occupations in parameter space. This must be subject to further research, to be able to include the metric in the final validation report.

5 Derivation of Metrics with Respect to the Requirements

Besides spanning the parameter space and deriving scenarios, requirements are also needed to find and select metrics for validation. Of course, metrics have to fit to the output interface of the sensor system simulation. Additionally, they need to reflect the required phenomena and effects and their required accuracy. With respect to this, they have to be designed carefully, as they have several design parameters themselves that need to match the accuracy requirements. The performed scenarios need to be considered when designing and selecting metrics, as well. Especially the velocities of the objects, when e.g. point clouds are compared, are relevant. Finally, it must be stated again that perception sensor data generation and validation of such is quite a young research field compared to e.g. vehicle dynamics simulation. So, experience is missing to select and design metrics, but also to interpret them.

To give an example and to pick up measurements from Fig. 2, metrics for comparison of lidar point clouds are presented and inspected in the following. The exemplary metric in this work is grid-based and taken from the collection in [24], which can be taken for further reading. Grid-based metrics are chosen in literature to validate (lidar) sensor system simulation, e.g. in [25], and are therefore selected here to show benefits and limitations of such metrics. In [11] it has been used, among others, to benchmark different simple lidar sensor models. In [20] it is, among others, applied to the already shown scenario. The metric is called occupied cells ratio (OCR), which is based on a cell-wise comparison of the real and simulated occupancy grid collecting the points of the point cloud into the cells. The OCR describes the relation of true classified, occupied cells in the simulated occupancy grid $\tilde{\mathbb{G}} = \{\tilde{c_1}, ..., \tilde{c_J}\}$ to the total number of occupied cells in the real occupancy grid $\mathbb{G} = \{c_1, ..., c_I\}$ [26],

$$\mathrm{OCR}^K = \frac{\sum \zeta(\tilde{c}_j)}{\sum \zeta(c_i)}, \quad \text{where} \quad \zeta(\tilde{c}_j) = \begin{cases} 1, & \text{if } P(\tilde{p_n} \text{ in } \tilde{c}_j) > 0.5 \\ 0, & \text{else} \end{cases}, \quad \zeta(c_i) = \begin{cases} 1, & \text{if } P(p_m \text{ in } c_i) > 0.5 \\ 0, & \text{else.} \end{cases}$$

For the calculation of OCR^K, K consecutive scans are considered for the generation of real and simulated occupancy grids. With these grids accumulated over time, the impact of noise is taken into account. Thus, $P(p_m \text{ in } c_i)$ indicates how often a cell was occupied during K scans. In this work, K equals 1, as a dynamic scenario is observed for which only a single scan can be taken into account for filling the grid. In consequence, $P(p_m \text{ in } c_i)$ is either 1 or 0, as it is either occupied or not, and the optimum of OCR^K is 1. Only OCR is chosen in this work to be shown, as the correlation of grid-based metrics has been described in [24] and all others are shown in [20].

The chosen metric exemplarily shows what has to be considered, when metrics are chosen and designed. The grid's cell size needs to be analyzed beforehand and in the case of static scenarios the number K of considered scans plays a role. The cell size influences the metric's calibration and has to be evaluated in any case, as shown in [20, p. 70]. Another conclusion, especially on grid-based metrics like OCR, is that they would have higher expressiveness, when in the case of the chosen dynamic scenario, everything would have been stopped at several points, which would have allowed to collect more than one scan and would result in higher K and actual $P(p_m \text{ in } c_i)$ different to 1 or 0 [20, p. 81].

6 Valid Measurement and Reference Data

The observations and conclusions on design and interpretation of metrics already show that accuracy of measurement and reference data is crucial to the overall simulation validation. Small accuracy in reference data (e.g. position and orientation of objects) results in inaccurate movements in simulation that could mislead validation (esp. as shown for grid-based metrics on point clouds). The influence of the reference data accuracy could probably be avoided by "optimization" (shifting and scaling of simulated point clouds to fit to real ones) of the simulated data with respect to known systematic deviations in the reference data, but more research is needed in this case.

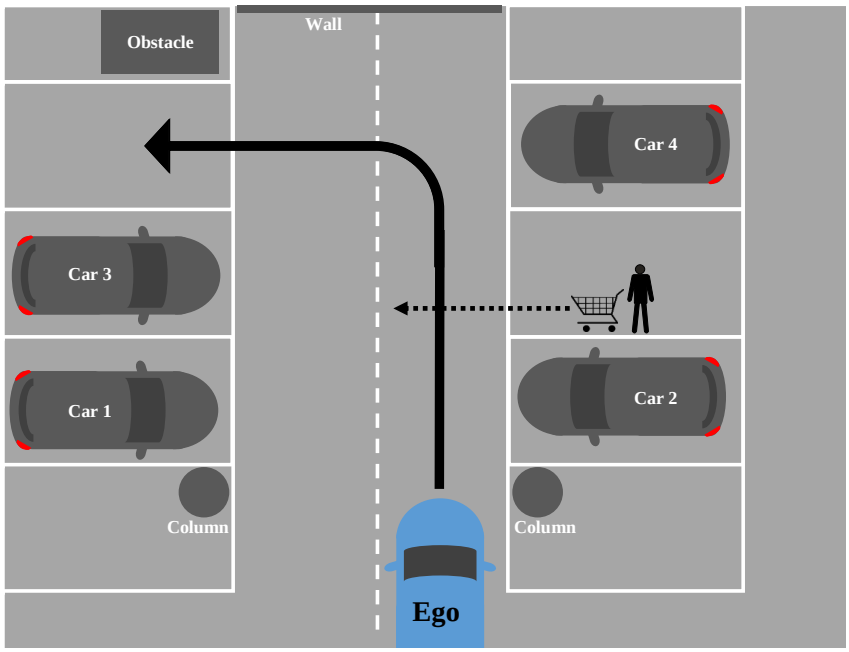

Figure 4: Performed parking area scenario including pedestrian with shopping trolley

As for all experiments, repeatability and reproducibility of trajectories are challenging, as well. Especially measurements of perception sensors in dynamic multi-object scenarios are almost impossible to repeat exactly, so the deviations need to be collected. Nevertheless, the presented objective and statistics-based validation method takes these considerations into account, as the distribution of the metrics applied to measurement data from several repetitions of the scenarios are compared to the distributions of the same metrics applied to the simulated data, resulting from the re-simulated reference data, which was collected during measurements. Thereby, the influence of repeatability is neglected.

Reproducibility of reference data in simulation is limited, as well. Most likely, trajectories performed in simulation differ from the real ones, even if they were to be collected with high accuracy, as the simulation tool has to interpolate the trajectories of all moving objects in between the collected points along the reference data trajectories. Besides the movements, the limitations of simulation environments also play a role in the case of environmental parameters of the atmosphere and object properties. E.g. digitization of rain intensities can be a first example for these limitations. It gets even more challenging when existing standards like OpenDrive [27] and OpenScenario [28] should be used, as they are still under development and only partly implemented in simulation tools. In conclusion, after experiments, reference data has to be validated before they are re-simulated to generate the synthetic data for comparison.

The sensitivity of metrics for point clouds, and especially grid-based metrics, to inaccuracies in the range of 10 cm has already been mentioned in Sec. 5. Reference data, even real-time kinematic (RTK) based systems for trajectory collection, show an (in-)accuracy in this range, as exemplary calculations show [20, p. 82]. Therefore, a first conclusion is that lower size than reference data accuracy is at least questionable, as cell size should be at least ten times higher than measurement and reference data accuracy for high significance. Actually in this work, even when using RTK positioning devices combined with precise gyroscopes and observers and additional measuring tape measurements for reference data collection, we propose at least 10 cm for the regular Cartesian grid's cell size. Clearly, not only local, but also temporal uncertainties in the collected data (e.g. due to different sampling rates of measurement and reference data) matter, especially for grid-based metrics, as calculations in [20, p. 82] show.

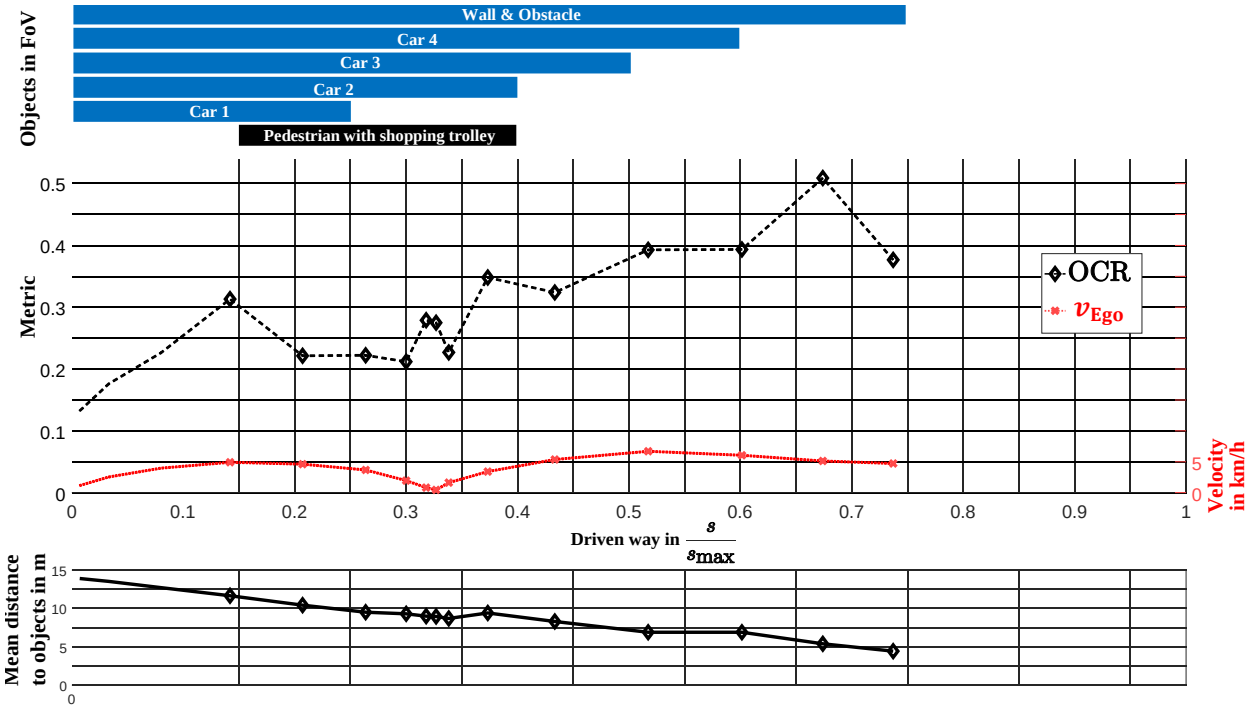

Figure 5: Objects in field of view (FoV), occupied cells ratio (OCR) during performed parking scenario, optimum would be 1, and mean distance to objects

7 First Results from Evaluation of a Simple Sensor Model

The dynamic parking scenario inspected in this work, as shown in Fig. 2, Fig. 3, and Fig. 4, is located at an exemplary parking area outside. Four cars of different size, shape, and reflectivity are already parked there and stay static. There are two small columns left and right, directly in front of the ego-car and there is an obstacle in the left back, as well as a wall in the center back. A pedestrian with a shopping trolley is walking across the planned trajectory of the ego-car, causing it to stop. The performed trajectory of the ego-car is planned to park at the third parking lot on the left side, next to car 3 and the obstacle.

Fig. 5 shows the temporal progress of the scenario with respect to the objects in the measurement range, visualized at the top, the velocity of the ego-car v_{Ego}, and the mean distance to the objects in the measurement range, shown at the bottom. Focus of Fig. 5 is to describe how the OCR metric evolves during the experiment, correlated to the mean distance to the objects, while the velocity stays below 10 km/h. Due to the performed trajectory, no objects are visible in the sensors' measurement ranges after 75 % of the overall path. Therefore, metric, velocity, and distance calculation in Fig. 5 end there.

The tendency of worse results of the OCR metric with higher distance, as visible in Fig. 5, confirms the results for the performed benchmarking of idealized lidar sensor models in [11] for a different scenario. As reported in [11], metrics applied to real and synthetic point cloud data from simple ray casting or Z-buffer clearly show worse results with higher distance to objects. Of course, the investigation here does not claim to be a representative validation study, but only an exemplary evaluation to draw and show first results and conclusions. So, against first impressions by visual inspection in Sec. 2, objective metrics prove that all inspected simulations are falsified, if requirements state that grid-based metrics used for point clouds should show high overlap of real and synthetic point clouds up to high distance.

Nevertheless, reference data and its repeatability in simulation are crucial to the overall evaluation. In the presented case, the accuracies of trajectory recording and position measurement of static objects are of about 10 cm, while measurement accuracy of the sensor system and cell size of the grid metric are in the same range. Orientation measurements of static objects show similar accuracy and several more uncertainties exist, e.g. through transformation of coordinates for re-simulation into the world

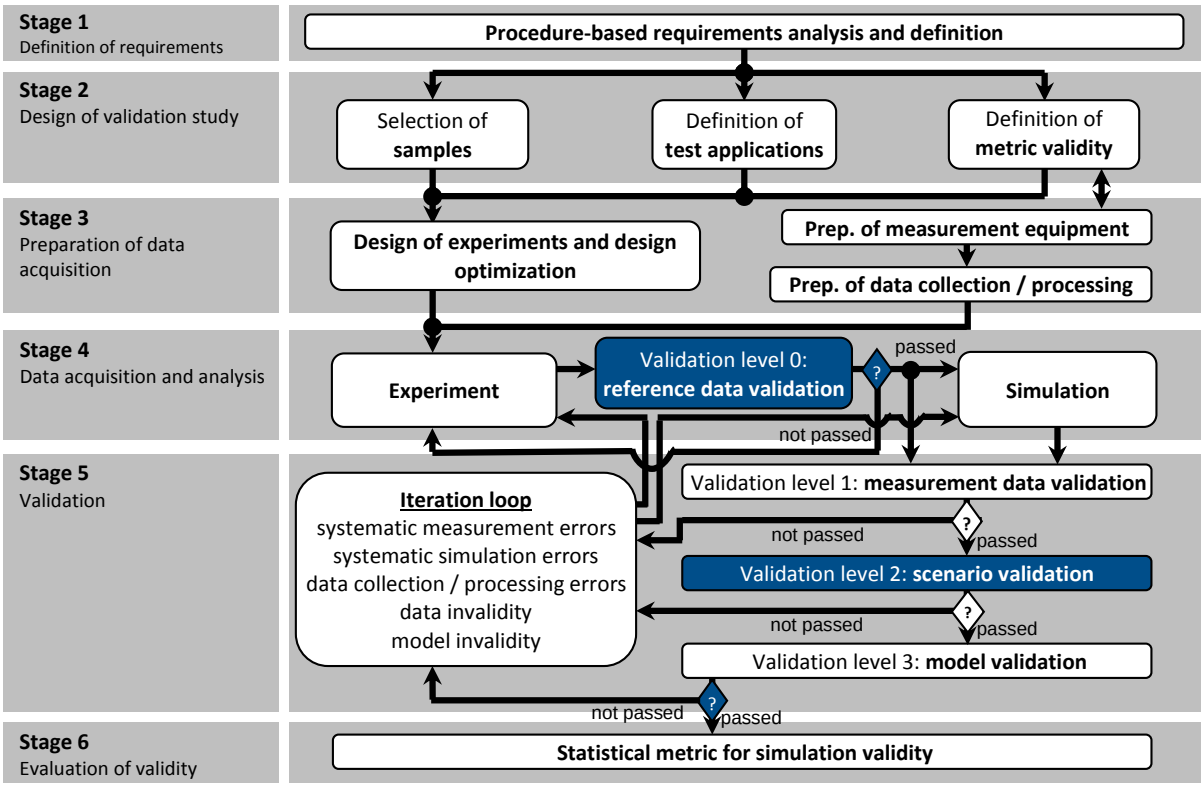

Figure 6: Objective quality assessment by statistical validation, blue: changes to the method in [9, p. 47]

frame of the simulated representation of the parking area. This results in concerns regarding the overall qualification of the here and often used grid-based metrics. Therefore, a main topic for further investigation is to either find new metrics that are more appropriate, or to calibrate and qualify grid-based metrics for the intended use, besides to aim for higher reference data accuracy, of course.

8 Resulting Extensions to the State-of-the-art Model Validation Method

After statistical validation of sensor system simulation has been discussed, findings conclude in the extensions to the existing method from Viehof [9, p. 47], as they are visualized in Fig. 6. The first stage that involves requirements analysis and definition has been addressed in Sec. 3. In the second stage, three tasks are included. The selection of samples from the parameter space, definition of tests, and definition of metric validity have been discussed in Sec. 4 and Sec. 5. Stage 3 has not been addressed here, as it mainly involves experiment design and preparation (including equipment) that result from requirements on measurement and reference data. All three stages are ordered perfectly fine within the existing method and no structural changes are needed, while important points have been described.

After experiments are performed, as seen in stage 4, we propose to insert a first evaluation block for the recorded reference data, as marked in blue. If accuracies of reference data in time and space are according to the requirements, (re-)simulation can take place, located at stage 4. Afterwards, in stage 5, three different validation blocks take place. The first one, on measurement and simulation data, checks for systematic measurement and simulation errors, which could be e.g. differences between reference and simulated trajectories. Here, data collection and processing is investigated, as well. If measurement validation is passed, scenario (originally: parameter) validation takes place. This includes checking the selection of samples and if the expectations in the sensitivity analysis do hold. We propose a change of wording with respect to [9], which has already been explained in Sec. 4. Finally, model validation is performed, which leads to the overall simulation validation that results in the statistical metric for

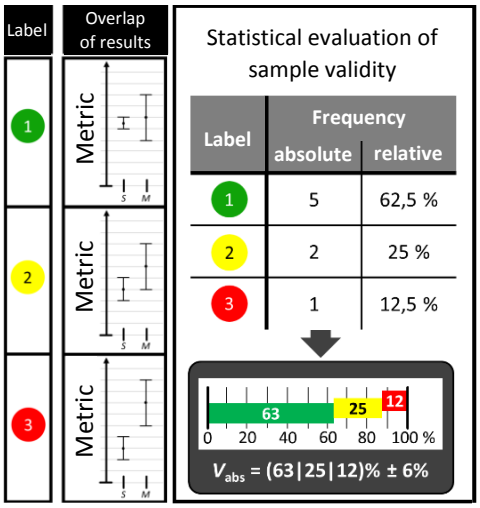

Figure 7: Statistical evaluation for absolute comparison of simulated and real data, adapted from [9]

simulation validity in stage 5. As shown in Fig. 6, we propose an additional, optional iteration loop after model validation to visualize the option to improve the model, after sample validation has taken place.

At the end, as an outlook towards further investigations, a short look on the statistical validation process is provided. We focus on the absolute comparison of real against simulated data, as shown in Fig. 7. The final evaluation chart will show for each sample and different metrics, how the metrics applied to several repetitions for respective samples overlap in range. The distributions result from deviations in reference data for different repetitions, even if no noise is considered in the model. The colors show if the metric applied to simulated data after several repetitions lies within the range of the real data (1, green), overlaps only partly (2, yellow), or has no overlap (3, red). Here, requirements have to be considered, as sensor system simulation e.g. aims for worst-case simulation due to test requirements. If it should still be possible to show this with this color schematic, either metrics have to be adapted or the labeling must be designed differently, e.g. to only label green if both distributions show similar variance. The statistical evaluation after all experiments and simulations delivers the overall trust value for the model in the inspected parameter space. If this evaluation is mostly green and no red or black labels are indicated, sample validity of the model is shown for the inspected parameter space.

9 Conclusion

A generally accepted validation methodology for perception sensor system simulation has not yet been established [7]. In this work, progress concerning the validation method is presented, as findings and improvements on an existing methodology as well as concrete experience in comparison of real and simulated data are shown. Definitions have been written up for several terms in the field. A first guideline for formulating requirements for sensor system simulation has been proposed. Questions to be answered when scenarios for parametrization and validation of such simulation are listed again and important points have been highlighted. Metrics for point clouds have been applied to a dynamic scenario to point out actual challenges in design and interpretation. The importance of reference data validation when validation experiments are performed and re-simulated has been shown. Finally, the state-of-the-art method for model validation has been extended for application to perception sensor system simulation. This work highlights recent challenges and improvements of assessment by statistical validation towards a generally accepted methodology. However, some aspects like deriving requirements from a concrete intended function (e.g. an assistance system) or selection and representation of scenarios, especially edge cases, as part of the whole process were not addressed herein and are part of future work.

Acknowledgments

This paper is part of the work in the project PEGASUS funded by the German Federal Ministry for Economic Affairs and Energy based on a decision of the Deutsche Bundestag.
The authors would like to thank Michael Viehof for the valuable discussions.

References

[1] German Aerospace Center. *PEGASUS Research Project: Securing Automated Driving efficiently*. `http://www.pegasus-projekt.info/en/about-PEGASUS`, 2017. Accessed: 2019-01-25.

[2] AVL LIST GMBH. *Enable-S3: European Initiative to Enable Validation for Highly Automated Safe and Secure Systems*. `https://www.enable-s3.eu/about-project/`, 2017. Accessed: 2019-01-25.

[3] Fabian Schuldt. *Ein Beitrag für den methodischen Test von automatisierten Fahrfunktionen mit Hilfe von virtuellen Umgebungen*. Ph.D. thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2017.

[4] Christian Amersbach and Hermann Winner. *Functional Decomposition: An Approach to Reduce the Approval Effort for Highly Automated Driving*. In *8. Tagung Fahrerassistenz, Einführung hochautomatisiertes Fahren*, München, Germany, 2017. TÜV Süd Akademie GmbH.

[5] Walther Wachenfeld and Hermann Winner. *Die Freigabe des autonomen Fahrens*. In *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, pages 439–464, Berlin, Heidelberg, Germany, 2015. Springer Berlin Heidelberg.

[6] Philipp Rosenberger, Martin Holder, Marina Zirulnik, and Hermann Winner. *Analysis of Real World Sensor Behavior for Rising Fidelity of Physically Based Lidar Sensor Models*. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, Suzhou, China, 2018.

[7] Martin Holder, Philipp Rosenberger, Felix Bert, and Hermann Winner. *Data-driven Derivation of Requirements for a Lidar Sensor Model*. In *2018 Graz Symposium Virtual Vehicle (GSVF)*, Graz, Austria, 2018.

[8] Michael Viehof and Hermann Winner. *Stand der Technik und der Wissenschaft: Modellvalidierung im Anwendungsbereich der Fahrdynamiksimulation: Forschungsbericht*. Darmstadt, Germany, 2017. Technische Universität Darmstadt.

[9] Michael Viehof. *Objektive Qualitätsbewertung von Fahrdynamiksimulationen durch statistische Validierung*. Ph.D. thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2018.

[10] Nils Hirsenkorn, Paul Subkowski, Timo Hanke, Alexander Schaermann, Andreas Rauch, Ralph Rasshofer, and Erwin Biebl. *A ray launching approach for modeling an FMCW radar system*. In *2017 International Radar Symposium (IRS)*, Prague, Czech Republic, 2017.

[11] Philipp Rosenberger, Martin Holder, Sebastian Huch, Hermann Winner, Tobias Fleck, Marc René Zofka, Johann Marius Zöllner, Thomas D'Hondt, and Benjamin Wassermann. *Benchmarking and Functional Decomposition of Automotive Lidar Sensor Models*. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, 2019. Accepted for presentation.

[12] dSpace GmbH. Probabilistic Sensor Models for ADAS/AD Simulations. `https://www.dspace.com/en/ltd/home/products/systems/simulationmodels/simulation_models_use_cases/probabilisticsensormodels.cfm`, 2019. Accessed: 2019-04-08.

[13] Jean-Claude Kedzia, Philippe de Souza, and Dominique Gruyer. Advanced RADAR Sensors Modeling for Driving Assistance Systems Testing. `https://www.esi-group.com/de/resources/technical-paper/advanced-radar-sensors-modeling-driving-assistance-systems-testing`, 2016. Accessed: 2019-04-08.

[14] IPG Automotive GmbH. Many Sensors, One Solution: Virtual Test Driving with CarMaker. https://ipg-automotive.com/news/article/many-sensors-one-solution-virtual-test-driving-with-carmaker/langswitch/1/, 2017. Accessed: 2019-04-08.

[15] TASS Internatinal. PreScan Sensors. https://tass.plm.automation.siemens.com/prescansensors, 2019. Accessed: 2019-04-08.

[16] TESIS GmbH. Sensor simulation: Physical simulation of environment sensors for ADAS & AD. https://www.tesis.de/en/sensorsimulation/, 2019. Accessed: 2019-04-08.

[17] TWT GmbH. TRONIS SENSORS. https://www.tronis.de/features, 2019. Accessed: 2019-04-08.

[18] VIRES Simulationstechnologie GmbH. VTD - VIRES Virtual Test Drive. https://vires.com/vtd-vires-virtual-test-drive/, 2019. Accessed: 2019-04-08.

[19] ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results Part 1: General principles and definitions. Standard ISO 5725-1:2013, International Organization for Standardization, 2013.

[20] Moritz Berghöfer. *Generierung realer und synthetischer Sensordaten zur Validierung von Sensormodellen für die simulationsbasierte Absicherung der Valet Parking Funktion*. B.Sc. thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2019.

[21] Erwin Roth, Tobias Dirndorfer, Kilian v. Neumann-Cosel, Marc-Oliver Fischer, Thomas Ganslmeier, Andreas Kern, and Alois Knoll. Analysis and validation of perception sensor models in an integrated vehicle and environment simulation. In *Proceedings of the 22nd Enhanced Safety of Vehicles Conference (ESV)*, 2011.

[22] Till Menzel, Gerrit Bagschik, and Markus Maurer. *Scenarios for Development, Test and Validation of Automated Vehicles*. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, Suzhou, China, 2018.

[23] Christian Amersbach and Hermann Winner. *Defining Required and Feasible Test Coverage for Scenario-Based Validation of Highly Automated Vehicles*. In *2019 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Auckland, New Zealand, 2019. Author-submitted.

[24] Sebastian Huch. *Entwicklung einer umfassenden Metrik für die Bewertung einer Lidar-Sensor-Simulation durch Betrachtung mehrerer aufeinander folgender Verarbeitungsebenen*. M.Sc. thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2018.

[25] A. Schaermann, A. Rauch, N. Hirsenkorn, T. Hanke, R. Rasshofer, and E. Biebl. Validation of vehicle environment sensor models. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 405–411, Redondo Beach, CA, USA, 2017.

[26] Ralph Grewe, Matthias Komar, Andree Hohm, Stefan Lüke, and Hermann Winner. Evaluation method and results for the accuracy of an automotive occupancy grid. In *2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, pages 19–24, 2012.

[27] Marius Dupuis, Esther Hekele, and Andreas Biehn et al. OpenDRIVE Format Specification, Rev. 1.5. http://www.opendrive.org/docs/OpenDRIVEFormatSpecRev1.5M.pdf, 2019. Accessed: 2019-04-08.

[28] VIRES Simulationstechnologie GmbH. OpenSCENARIO v0.9.1 XML Schema. http://www.vires.com/OpenSCENARIO/OpenSCENARIO_v0.9.1/OpenSCENARIO_v0.9.1_specification.zip, 2017. Accessed: 2019-04-08.