electronics

Article

Modeling Stochastic Data Using Copulas for Applications in the Validation of Autonomous Driving

Katrin Lotto [1,*], Thomas Nagler [2,3] and Mladjan Radic [1]

[1] ZF Friedrichshafen AG, Research and Development, Graf-von-Soden-Platz 1, 88046 Friedrichshafen, Germany
[2] Department of Statistics, LMU Munich, 80799 Munich, Germany
[3] Munich Center for Machine Learning, 80799 Munich, Germany
* Correspondence: katrin.lotto@zf.com

Abstract: The verification and validation processes of fully automated vehicles are linked to an almost intractable challenge of reflecting the real world with all its interactions in a virtual environment. Influential stochastic parameters need to be extracted from real-world measurements and real-time data, capturing all interdependencies, for an accurate simulation of reality. A copula is a probability model that represents a multivariate distribution, examining the dependence between the underlying variables. This model is used on drone measurement data from a roundabout containing dependent stochastic parameters. With the help of the copula model, samples are generated that reflect the real-time data. The resulting applications and possible extensions are discussed and explored.

Keywords: copula; vine copula; dependance; rank correlation; bayesian reliability; risk analysis; automated driving; verification; validation

Citation: Lotto, K.; Nagler, T.; Radic, M. Modeling Stochastic Data Using Copulas for Applications in the Validation of Autonomous Driving. *Electronics* **2022**, *11*, 4154. https://doi.org/10.3390/electronics11244154

Academic Editors: Tao Zhang, Hyo-sang Shin, Xiang Xu and Gaoge Hu

Received: 10 November 2022
Accepted: 9 December 2022
Published: 13 December 2022

1. Introduction

Autonomous driving vehicles have a huge potential not only for reducing the number of road accidents but also from an economic point of view, reducing emissions as well as increasing the efficiency of traffic usage by, e.g., robotaxis or avoiding traffic jams. Thus, autonomous driving already revolutionizes our view on mobility, traffic safety, and how time is spent during car journeys. The following question arises: how do we verify that an autonomous vehicle is safe enough. There are many proposals on how many kilometers an autonomous vehicle has to drive accident-free before it may be safely released on the road. The estimation reaches from several millions to even billions of kilometers. We want to stress that this is only for one system. All updates and changes in the software, occuring errors, improvements, and transformations require a rerun. Every experiment of such dimensions, if feasible at all, would be tremendously cost-intensive. This motivates a simulation-based test of autonomous vehicles.

A simulation-based framework aims to replicate real-world traffic and to validate the automated driving functionality accordingly. In this framework, robustness, reliability, and safety can be investigated in a much quicker and cost-efficient manner. It is emphasized that the validation results and investigations in the simulated environment are only as trustworthy as the virtual environment reflects the real world [1–4].

The simulation is fed with stochastic data, e.g., velocity/acceleration of the traffic participants, extracted from the real recordings of traffic events.

A huge challenge is that the stochastic parameters, such as the velocities of every traffic participants, are not independent. There are many factors that have to be considered. The weather, visibility, and the amount of traffic participants are correlated to the driving behaviour of the participants. Unfortunately, this correlation is unidentified and cannot be assumed beforehand. In this work, dependencies in stochastic data are investigated and described by so-called copulas.

The latin word copulare may be translated into "to join" or "to connect", which already reveals the meaning and functionality of copulas. They are great tools for modeling the (nonlinear) dependence of random variables and for calculating the joint distribution. The first appearance of the term copula may be found in [5]. Since then, substantial research and the application of copulas can be seen in the fields of statistics [5–7], economics [8,9], finance [10–13], actuarial science [14,15], or risk management [16–18] to name a few.

The outline of this work is as follows. In Section 2, we examine further the subject of validating automated vehicles and the theory of copulas. The construction and usage for resimulating data are described. Section 3 illustrates how data can be extracted from observations and how the theory on a concrete numerical example and application can be applied. A conclusion and outlook for further research and development is given in Section 4.

2. Preliminaries

2.1. Safety Validation of Automated Vehicles

Unlike today's cars, autonomous vehicles require additional regulations. A distinction is made between automation levels, as in the SAE J3016 standard [19]: a *driver-only vehicle* without automation, *assisting systems*, and *semi-automated systems*. In all three levels, the driver keeps the responsibility for the behavior of the vehicle. Furthermore, we speak of *fully automated vehicles* when the driver no longer acts as a fallback [20]. In this case, the automated systems should take over vehicle control permanently and should act safely in any environment.

The international organisation for standardisation (ISO) provides ISO/PAS (Publicly Available Specification) 21448 to ensure the safety of the functionalities [21] and ISO 26262 to ensure that no hazards are caused by technical failures [22]. ISO/PAS 21448, which comprises the safety of the intended funcionality (SOTIF), focuses on predictable misuse by the driver, as well as accidents that are explicitly not caused by component failure but situations that were not planned for during development. Nevertheless, this standard does not provide a detailed strategy for identifying functional deficiencies. In contrast, ISO 26262 focuses on safety in terms of intrinsic safety (the protection of the environment from the product) and, therefore, functional safety. It is an ISO standard for safety-related electrical/electronic systems in motor vehicles. It is not enough that the function has been executed correctly, but it must also be ensured that the function has been executed in the correct context. For example, an airbag must not be triggered when driving too fast over speed bumps. However, the minimum requirements for safeguarding an autonomous vehicle are not adequately described by this standard. The goal is to minimize risks to a level that is "acceptable to society".

Consequently, the validation of autonomous vehicles cannot be based on existing ISO standards and require extension. Various approaches have been developed, starting with advanced driver assistant systems (ADAS), where function-based approaches and real-world testing operate well. In the function-based approach, requirements are defined for the operating system that are tested by simulation or on the test track. In real-world testing, a mileage-based evaluation of functionality from field tests with a driver is performed. For both methods, validating fully autonomous systems is economical infeasible in its current state. Furthermore, there is the shadow mode, presented by Wang and Winner [23], in which the automated driving function is excecuted passively in series production vehicles. The driving function receives (real) information from the sensors, but it will not act. Its actions are evaluated afterwards by simulation.

Considering the real world as an open parameter space, with an infinite number of traffic events, the scenario-based approach tries to identify these traffic events and describe them in scenarios. It also attempts to exclude non-relevant traffic events, where neither actions nor events are observed, and to cluster similar traffic events into a representative scenario [24]. However, this leads to the question of how to find the set of representative

scenarios. A good overview of the problem of identifying critical scenarios is provided by Neurohr et al., Riedmaier et al., and Zhang et al. [25–27].

Menzel et al. [28] distinguished three categories of scenarios: functional, logical, and concrete scenarios. In the case of functional scenarios, the scenario space, in the same way as traffic events, is described at a sematic level by "a linguistic scenario annotation" [28]. For the logical scenarios, this is described at the state space level. Entities and their relationships are described using parameter ranges in the state space and optionally specified using correlations and numerical relationships. A traffic event is finally mapped explicitly to the state space given a concrete scenario. Entities and their relationships are described using concrete values for each parameter. According to Zhang, the identification of critical scenarios can be performed at all three levels of abstraction. This requires a clear definition of the operational design domain (ODD), a definition of the operating conditions under which an AD system attempts to operate, and "the formulation of a functional scenario to a logical scenario".

The German research project Pegasus (Pegasus—project for the establishment of generally accepted quality criteria, tools and methods as well as scenarios and situations for the release of highly automated driving functions; see https://www.pegasusprojekt.de/en/home, accessed on 5 November 2022) followed the approach in [29] of modeling the environment in layers and extended it to six layers for describing the environment of a highway. In the follow-up project VVM (VVM—verification and validation methods for level 4 and 5 automated vehicles; see https://www.vvm-projekt.de/, accessed on 5 November 2022), the six-layer model was refined, extended to the urban environment, and provided with guidelines [30].

To describe the operating conditions of an AD system, the parameters of the six layers are used, such as weather, the number of road users, and the number of lanes or speed of a cyclist. The assessment of whether a scenario is critical or not is based on the concrete values for these stochastic parameters. Some studies already consider realistic parameter distributions [27] that are obtained from real-life driving databases.

In scenario-based testing, parameter distributions can be used to support the search for critical scenarios within the scenario space. In doing so, they serve to model the mutuality of the scenarios. In [31], a risk index based on a Gaussian mixture model is used to efficiently select critical traffic conditions. Thereafter, the scenarios are replicated via simulation and used, for example, to evaluate the criticality of the AD system. For the simulation of the scenarios, real trajectories that are extracted from real data are used. In the same way, the characteristic criteria of scenarios are taken from the real data, e.g., parameter distributions or parameter dependencies, and used to reproduce trajectories or parameters. Thus, parameter distributions are applied in estimating the failure rate of a scenario. For example, Wagner et al. [32] predicted the behavior of road users based on conditional distributions obtained by analyzing a naturalistic driving study called euroFOT (large-scale European Field Operational Test on Active Safety Systems) and determineed the criticality for each predicted event.

2.2. Copulas

Copulas allow modelling the (nonlinear) dependence of several random variables. With the help of copulas, the multivariate distribution can be split into their marginal distributions. In the following, the definition and construction as well as the main features of copulas will be discussed and illustrated.

2.2.1. Definition

For a d-dimensional random vector $X = (X_1, \ldots, X_d)$, we denote the joint cumulative distribution function by $F_X(x) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$ and the marginal distributions by $F_k(x_k) = P(X_i \leq x_i)$. A copula is a special type of distribution function.

Definition 1 (Copulas). *A d-dimensional copula* $C\colon [0,1]^d \to [0,1]$ *is a multivariate distribution function with standard uniform marginals. If the copula is absolutely continuous, the corresponding copula density is defined as follows.*

$$c(u_1,\ldots,u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1,\ldots,u_d), \quad \forall (u_1,\ldots,u_d) \in [0,1]^d. \tag{1}$$

Sklar's theorem [5] states that any multivariate distribution can be expressed in terms of its marginal distributions and a copula.

Theorem 1 (Sklar's Theorem, 1959). *Let* $X = (X_1,\ldots,X_d)$ *be a d-dimensional random vector and let the corresponding joint distribution function be denoted by F as well as the marginal distribution functions by* F_i*, where* $i = 1,\ldots,d$*. Then, the joint distribution function is given by*

$$F(x_1,\ldots,x_d) = C(F_1(x_1),\ldots,F_d(x_d)), \tag{2}$$

where C is a d-dimensional copula. If the distributions are absolutely continuous, then copula C is unique. Moreover, the corresponding density function is given by

$$f(x_1,\ldots,x_d) = c(F_1(x_1),\ldots,F_d(x_d)) \prod_{i=1}^{d} f_i(x_i) \tag{3}$$

Sklar's Theorem can be understood as follows. The marginal distributions $F_1,\ldots,F_d$ characterize the behavior of each of the variables $X_1,\ldots,X_d$ in isolation. The copula on the other hand characterizes the dependence between them. On the one hand, we can decompose any joint distribution into marginal distributions and a copula. On the other hand, we can combine any copula with arbitrary marginal distributions to form a valid multivariate disribution function.

Example 1. *The copula corresponding to independent random variables* $X_1,\ldots,X_d$ *is* $C(u_1,\ldots,u_d) = u_1 \cdots u_d$ *and it is called an independence copula.*

2.2.2. Bivariate Parametric Copula Models

There are two major classes of parametric copula models: elliptical copulas and Archimedean copulas. Two well-known examples of elliptical copulas are the multivariate Gaussian copula and the multivariate Student's t copula. They are constructed implicitly by inverting the formula in Sklar's theorem of the following:

$$C(u_1,u_2) = F\Big(F_1^{-1}(u_1), F_2^{-1}(u_2)\Big), \tag{4}$$

and they take F to be a bivariate Gaussian or Student's t distribution. For further information, refer to, e.g., [6].

Archimedean copulas are constructed more explicitly via generator functions.

Definition 2 (Bivariate Archimedean Copulas). *Let* $\varphi : [0,1] \to [0,\infty]$ *be a continuous, strictly monotone decreasing and convex function satisfying* $\varphi(1) = 0$*; then,* φ *is the generator of the bivariate Archimedean copula*

$$C(u_1,u_2) := \varphi^{[-1]}(\varphi(u_1) + \varphi(u_2)), \tag{5}$$

where

$$\varphi^{[-1]} : [0,\infty] \to [0,1] : \qquad \varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0), \\ 0, & \varphi(0) < t \le \infty \end{cases} \tag{6}$$

is the so-called pseudo-inverse of $\varphi(\cdot)$.

For a specific choice of the generator, the bivariate copulas given in Table 1 may be constructed. Note that the BB1 and BB7 families are two-parametric famlies where the others are one-parametric Archimedean copulas, but [6] is suggested for further information.

Table 1. Classes of one- and two-parametric Archimedean copulas.

Name	Bivariate Archimedean Copula $C(u_1, u_2)$	δ	θ
Clayton	$\left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-1/\delta}$	$\delta \in (0, \infty)$	–
Gumbel	$\exp\left(-\left[(-\ln(u_1))^{\delta} + (-\ln(u_2))^{\delta}\right]^{1/\delta}\right)$	$\delta \geq 1$	–
Frank	$-\frac{1}{\delta}\ln\left(\frac{1}{1-e^{-\delta}}\left[(1-e^{-\delta}) - (1-e^{-\delta u_1})(1-e^{-\delta u_2})\right]\right)$	$\delta \neq 0$	–
Joe	$1 - \left[(1-u_1)^{\delta} + (1-u_2)^{\delta} - (1-u_1)^{\delta}(1-u_2)^{\delta}\right]^{1/\delta}$	$\delta \geq 1$	–
BB1	$\left\{1 + \left[\left(u_1^{-\theta} - 1\right)^{\delta} + \left(u_2^{-\theta} - 1\right)^{\delta}\right]^{1/\delta}\right\}^{-1/\theta}$	$\delta \geq 1$	$\theta > 0$
BB7	$1 - \left(1 - \left[\left(1-(1-u_1)^{\theta}\right)^{-\delta} + \left(1-(1-u_2)^{\theta}\right)^{-\delta} - 1\right]^{-1/\delta}\right)^{1/\theta}$	$\delta > 0$	$\theta \geq 1$

Remark 1. *Perfect dependence and independence are achieved for Clayton if* $\delta \to \infty$ *and* $\delta \to 0$ *and for Gumbel if* $\delta \to \infty$ *and* $\delta = 1$*, respectively. For the Frank copula, the indepence copula is achieved for* $\delta \to 0^+$*, and for the Joe copula, the indepence copula is achieved for* $\delta = 1$*. For the BB1 copula, the independence copula corresponds to* $\theta \to 0^+$ *and* $\delta \to 1^+$ *and for the BB7 copula, the independence copula corresponds to* $\theta = 1$ *and* $\delta = 0$*.*

Copula C is related to the distribution of random variables $U_j = F_j(X_j)$, $j = 1, \ldots, d$. We thus first estimated marginal distributions $F_1, \ldots, F_d$ by $\widehat{F}_1, \ldots, \widehat{F}_d$. Then, we generate "pseudo-observations" $\widehat{U}_j = \widehat{F}_j(X_j)$, from which the copula parameters can be estimated using likelihood methods. We refer to [6,33–35] for more information.

2.2.3. Vine Copulas

As shown above, there is a plethora of flexible parametric models for two-dimensional vectors. The extensions of elliptical and Archimedean copulas to the d-dimensional case are much less flexible however. Elliptical copulas impose strong constraints on the symmetry of dependence. For example, large observations cannot have stronger dependence than small observations. Archimedean copulas are exchangeable, i.e., all subsets of the variables must have the same dependence.

Pair-copula constructions were introduced by [36–38] to mitigate these issues. The idea is to construct a multivariate dependence structure from only bivariate copulas. Each of these bivariate building blocks describe the (conditional) dependence between a pair of variables. The bivariate copulas can be specified independently, e.g., with some Gaussian and some Archimedean with different strengths of dependence, which make these models extremely flexible.

The above construction relies on conditioning. To organize all admissible orders of conditioning, [37] introduced a graphical model called *vine*. A vine is a set of trees, where a tree $T = (V_T, E_T)$ is a connected graph that contains no cycles; the reader can refer to [39]. Regular vines (R-vines) are a nested set of $n-1$ trees $T_1, \ldots, T_{n-1}$ such that a node of the next tree is built upon the edges of the previous tree. Furthermore, the edges are only joint if they have a common node in the previous tree. If additionally the degree of each node in the first tree is at most 2, then each tree is a path, and we speak of D-vines. If each tree has

a unique node of degree $n-1$, each tree is a star, and we speak of C-vines. Hence, D- and C-vines are special cases of R-vines. An example of a four-dimensional D-vine copula can be seen in the second figure in Section 3.

Every vine graph allows the decomposition of the joint density in a different way. To perfrom this, each edge $e = (x, y)$ is assigned a label $(\mathcal{C}_x, \mathcal{C}_y | D_e)$, where $\mathcal{C}_x, \mathcal{C}_y \in \{1, \ldots, d\}$ and $D_e \subset \{1, \ldots, d\}$. Each edge is further associated with a copula c_e that captures the dependence of $(X_{\mathcal{C}_x}, X_{\mathcal{C}_y})$ that is conditional on all $\{X_k : k \in D_e\}$. The reader can refer to [6] for further details. With this notation, we obtain the following result.

Theorem 2. *Let $\mathcal{X} = (X_1, \ldots, X_d)$ be a sequence of random variables with continuous invertible marginal distributions $\mathcal{F} = (F_1, \ldots, F_d)$. Let $\mathcal{V}$ be a R-vine tree sequence on d elements and $\mathcal{B} = \{c_e \mid e \in E_i,\ i = 1, \ldots d-1\}$ be the set of the associated bivariate copula densities. Then, joint density f can be factorized into*

$$f(\mathbf{x}) = \left[\prod_{i=1}^{d-1} \prod_{e \in E_i} c_e\left(F_{\mathcal{C}_x|D_e}(x_{\mathcal{C}_x}|\mathbf{x}_{D_e}), F_{\mathcal{C}_y|D_e}(x_{\mathcal{C}_y}|\mathbf{x}_{D_e}); \mathbf{x}_{D_e}\right)\right] \cdot \left[\prod_{k=1}^{d} f_k(x_k)\right]. \tag{7}$$

The conditional distributions $F_{\mathcal{C}_x|D_e}, F_{\mathcal{C}_y|D_e}$ in this formula can be computed recursively from the pair-copulas c_e in earlier tree levels.

The formula can be simplified for a D- or C-vine. We refer to [6,37,38] for further reference. In this paper, we use the R package rvinecopulib [40], where the latter approaches are implemented. As we do not have any knowledge about the vine structure, we make use of the vine function, which performs parameter estimation and automatic model selection using the sequential procedure proposed by [41]. The package also allows simulations from the fitted model based on the inverse Rosenblatt transform algorithm [6,41].

2.3. Dependence Measures

One common way to measure the dependence amongst multivariate normal random variables is to use Pearson's correlation coefficient. This coefficient, $|\rho_{\text{BP}}| \in [0, 1]$, is a measure of linear dependence indicated with $|\rho_{\text{BP}}| = 1$ perfect and by $\rho_{\text{BP}} = 0$ no (linear) dependence. According to [7], the coefficient ρ_{BP} is "generally not the best measure of dependence" since it may not yield ± 1 for non-Gaussian random variables with perfect dependence. A better choice includes rank correlation coefficients, which measure monotonic dependence. Monotone associations mean a dependence, where if one variable increases, then the other tends to decrease (or increase) and vice versa. A nice property of these measures is that they are invariant relative to strictly increasing transformations on the variables [6,7]. The relation can otherwise be nonlinear. The best-known rank correlation measures are Spearman's ρ and Kendall's τ, which are explained in more detail in the following sections.

Definition 3 (Spearman's ρ_s)**.** *Let X_1, X_2 be continuous random variables with continuous marginal distributions F_1, F_2 and let $(U_1, U_2) := (F_1(X_1), F_2(X_2))$. Then, Spearman's rank correlation ρ_s for (X_1, X_2) is given by the following.*

$$\rho_s := \rho_s(X_1, X_2) := \rho_{\text{BP}}(U_1, U_2) = \rho_{\text{BP}}(F_1(X_1), F_2(X_2)) = \text{Cor}(F_1(X_1), F_2(X_2)). \tag{8}$$

Hence, Spearman's rank correlation coefficient is defnined as Pearson's correlation coefficient of the random variables $F_1(X_1)$ and $F_2(X_2)$. To compute Spearman's ρ_s empirically from a sample $(x_{i1}, x_{i2}), i = 1, \ldots, n$, one replaces marginal distributions F_1 and F_2 with the respective empirical distribution functions $\widehat{F}_1$ and $\widehat{F}_2$. The pseudo samples $u_{ij} = \widehat{F}_j(x_{ij})$ then correspond to r_{ij}/n, where r_{ij} is the rank of value x_{ij} among all $x_{1j}, \ldots, x_{nj}$. For example, $r_{ij} = 1$ if x_{ij} is the smallest value in the sample.

Definition 4 (Estimation of Spearman's ρ_s)**.** *An estimate of Spearman's ρ_s based on a sample (x_{i1}, x_{i2}) of size n with ranks r_{ij} for $j = 1, 2$ is given by*

$$\hat{\rho}_s := \hat{\rho}_s(X_1, X_2) := \frac{\sum_{i=1}^{n}(r_{i1} - \bar{r}_1)(r_{i2} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{n}(r_{i1} - \bar{r}_1)^2}\sqrt{\sum_{i=1}^{n}(r_{i2} - \bar{r}_2)^2}}, \tag{9}$$

where $\bar{r}_1 := \sum_{i=1}^{n} r_{i1}$ and $\bar{r}_2 := \sum_{i=1}^{n} r_{i2}$ are the corresponding sample rank means.

Kendall's τ on the other hand measures dependence by comparing pairs of observations.

Definition 5 (Kendall's τ)**.** *Let (X_1, X_2) be a random vector with joint distribution F and let $(X_1', X_2') \sim F$ be another vector with the same distribution, but it is independent of (X_1, X_2). Kendall's τ is defined as follows.*

$$\tau = P((X_1 - X_1')(X_2 - X_2') > 0) - P((X_1 - X_1')(X_2 - X_2') < 0). \tag{10}$$

In the above definition, we call the pair of vectors, (X_1, X_2) and (X_1', X_2'), *concordant* if the values $X_1 - X_1'$ and $X_2 - X_2'$ have the same sign. Otherwise, the pair is called *discordant*. Hence, Kendall's τ is the difference of the probabilites of concordance and discordance.

To calculate Kendall's τ empirically, one simply compares each observation and counts the number of concordances and ties.

Let N_C be the number of concordant pairs, let N_D be the number of discordant pairs, and let N_1 and N_2 be the numbers of ties in X_1 and X_2 respectively.

Definition 6 (Empirical Kendall's τ)**.** *The empirical Kendall's τ is given by*

$$\hat{\tau}_n := \frac{N_C - N_D}{\sqrt{N_C + N_D + N_1} \times \sqrt{N_C + N_D + N_2}}. \tag{11}$$

As shown in [6,7], Kendall's τ and Spearman's ρ_s are independent of the marginal distribution and only depend on the associated copula. In particular, they can be expressed as

$$\tau = 4 \int_{[0,1]^2} C(u_1, u_2)\, dC(u_1, u_2) - 1 \quad \text{and} \quad \rho_s = 12 \int_{[0,1]^2} u_1 u_2\, dC(u_1, u_2) - 3. \tag{12}$$

3. Application to Traffic Data

Imagine a road that is busy on one day and has little traffic on another. The driving behavior of the road users is expected to differ. Considering the speed and traffic volume in terms of the number of vehicles, for example, it is reasonable to assume that these two parameters correlate. If we additionally take the time of day and weather conditions into account, we have already approached multi-dimensional dependencies. For example, splashy roads or low sun with high traffic volume would have another impact on the speed of the participants than dry roads or a cloudy sky. In order to reproduce traffic events in simulations, preferably as close to reality as possible, these multidimensional dependencies need to be considered. If we further want to sample or generate random values, according to their associated multidimensional distribution, copulas deliver the possibility to satisfying this request. The additional value generated by this coupled knowledge is a significant advantage for the simulative validation of autonomous driving systems in the context of scenario-based validation [4,26]. In particular, when calculating the probability of failure of a system, the reliability strongly depends on the distribution of the underlying parameters [42]. The more accurate the stochastic input parameters

and if their dependencies are described and therefore sampled, the more accurate the statement about the probability of critical failures and therefore the result of the reliability analysis [42]. Precisely speaking, we want to use these multidimensional dependencies on the driver's behavior as a stochastic input in the form of a probability distribution for the validation of autonomous systems and automated driving functions. To examine this, we focus on an explicit example, which is described in the following.

The rounD dataset is a drone dataset, specifically created for behavior planning for automated vehicles, that can also be used for downstream safety validation. Due to the high-density interactions between road users [43], it is of special interest for the purpose and application of this work. Due to the recordings from the air, it avoids the risk of occlusion during recordings, as well as the risk, in which road users do not adopt natural behavior in traffic when being observed [44]. This is one of the requirements that the dataset satisfies to ensure that mutual influences are reflected. To create a trajectory-based dataset, which includes more than 13746 road users, traffic was recorded at three different locations in and around Aachen, Germany. The road user types can be divided into the following classes: cars (11530), trucks (1061), vans (608), trailers (257), buses (53), and VRUs (vulnerable road user), which include pedestrians (25), bicycles (88), and motorcycles (124).

Most recordings were made in *Neuweiler*; hence, we chose this location for our application. It is a four-armed roundabout where all access roads have two lanes and the exits are single-lane roads. The fact that there is no lane marking whithin the roundabout leads to many interactions. Figure 1 illustrates a recording from this location. The trajectories of the traffic participant's movement up to this point are drawn in blue and the continued trajectories are drawn in white, respectively.

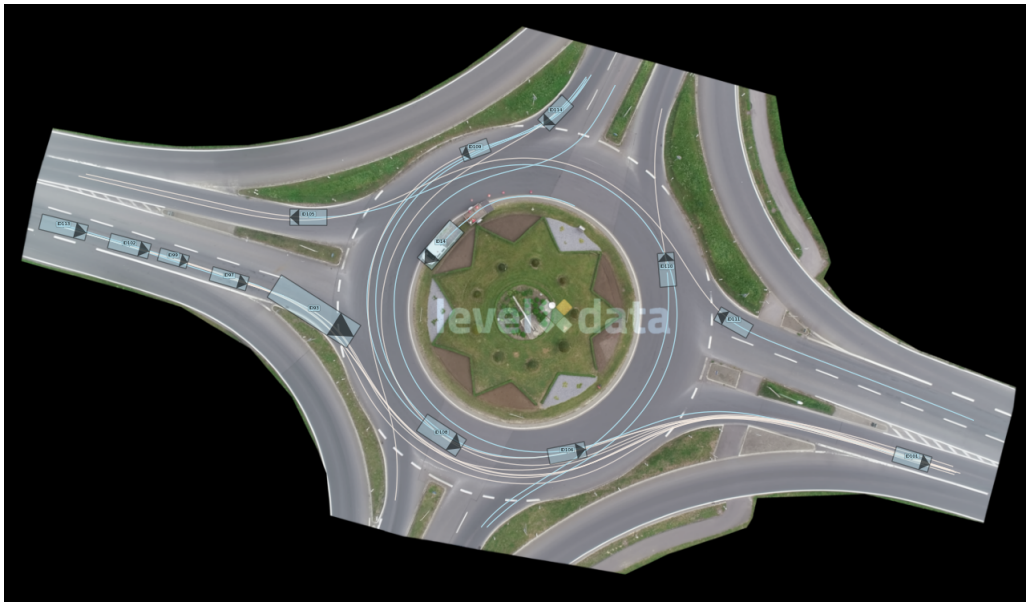

Figure 1. Recording at *Neuweiler*.

According to the report of the Insurers Accident Research [45], the most frequent accidents inside traffic circles occur due to approaching waiting vehicles too closely and disregarding the right of way when entering the traffic circle.

Based on this evaluation, we derive the four parameters shown in Table 2. We consider the speed behavior (VelCar) per frame, which is already provided by the RounD dataset. To determine the traffic density (TrafficCar), we count the vehicles within a radius of 10 m around each car per frame. Furthermore, we consider the time of standstill (WaitTime) for each vehicle before entering the roundabout as well as inside the roundabout. Finally, we take the minimal distance (DistCar) per frame from one car to all cars in its associated surrounding. A total number of 132,409 samples result from the 22 Neuweiler recordings. Each recording is taken with 25 frames per second and has a length between 19 and 23 min.

Since we count the vehicles for the representation of the traffic density, TrafficCar is a discrete parameter, whereas VelCar, WaitTime, and DistCar are continuous parameters.

Table 2. Parameters.

Parameter	Description	Unit
VelCar	velocity of cars per frame	(m/s)
TrafficCar	traffic density per car per frame	(−)
WaitTime	waiting time per car before entering roundabout	(s)
DistCar	minimal distance for one car to the other around him	(m)

In the lower triangular matrix of Table 3, Kendall's τ is depicted and Spearman's ρ_s in the upper triangular matrix. The values in Table 3 of the rank correlations confirm what can be intuitively surmised. If the amount of traffic participants increase (TrafficCar), vehicle speeds (VelCar) decrease, $\rho_s = -0.45$ and $\tau = -0.36$ (i.e., negative correlation), along with the distances (DistCar) between vehicles, $\rho_s = -0.79$ and $\tau = -0.64$; waiting times (WaitTime) increase, $\rho_s = 0.43$ and $\tau = 0.41$ (i.e., positive correlation). Analogously, the speeds (VelCar) decrease when the waiting times (WaitTime) increase, $\rho_s = -0.51$ and $\tau = -0.41$, and the distances (DistCar) between the vehicles decrease, $\rho_s = 0.49$ and $\tau = -0.34$. Furthermore, the distances (DistCar) decrease with increased waiting times (WaitTime), $\rho_s = -0.39$ and $\tau = -0.31$.

Table 3. Rank correlation coefficients, Kendall's τ and Spearman's ρ.

τ \ ρs	TrafficCar	VelCar	WaitTime	DistCar
TrafficCar	1.0	−0.45	0.43	−0.79
VelCar	−0.36	1.0	−0.51	0.49
WaitTime	0.41	−0.41	1.0	−0.39
DistCar	−0.64	0.34	−0.31	1.0

Subsequently, the copula is estimated from the four parameters using the vine function from the rvinecopulib package [40]. The resulting vine structure is shown in Figure 2. Therein, node 1 represents parameter TrafficCar, node 2 = VelCar, node 3 = WaitTime, and node 4 represents DistCar.

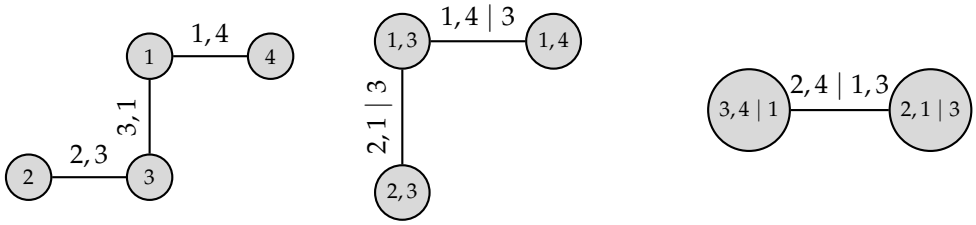

Figure 2. Vine tree structure.

The bivariate unconditional and conditional copula densities can be reconstructed from the tree structure such that the joint copula density c can be written as follows:

$$c(u_1, \ldots, u_4) = c_{23} \cdot c_{31} \cdot c_{14} \cdot c_{21|3} \cdot c_{34|1} \cdot c_{24|13}$$

with u_1 = TrafficCar, u_2 = VelCar, u_3 = WaitTime, and u_4 = DistCar. The unconditioned copulas c_{23}, c_{31} and c_{14} and conditioned copulas $c_{34|1}$ and $c_{24|13}$ were estimated using the local-likelihood transformation estimator (TLL). The conditioned bivariate copula

$c_{21|3}$ results in a (Survival) Clayton–Gumbel (BB1) family with parameters par $\in (0, \infty)$, par2 $\in [1, \infty)$ and internal coding family: 7, 17 [40]. As mentioned before, the generation of samples similar to real data is an important requirement for reliability and risk analyses. Figure 3 shows in black 5000 samples generated from the fitted vine copula model. They are plotted according to real data samples, which are depicted in grey and generated from the rounD dataset. Since the simulated data matche the original data sufficiently, the generated copula reflects the dependency structure of the four parameters and may therefore be used for further computation, such as the risk and reliability analysis. Further analysis is not performed in this study. In the next chapter, we will give an outlook for ongoing investigation and stress tests.

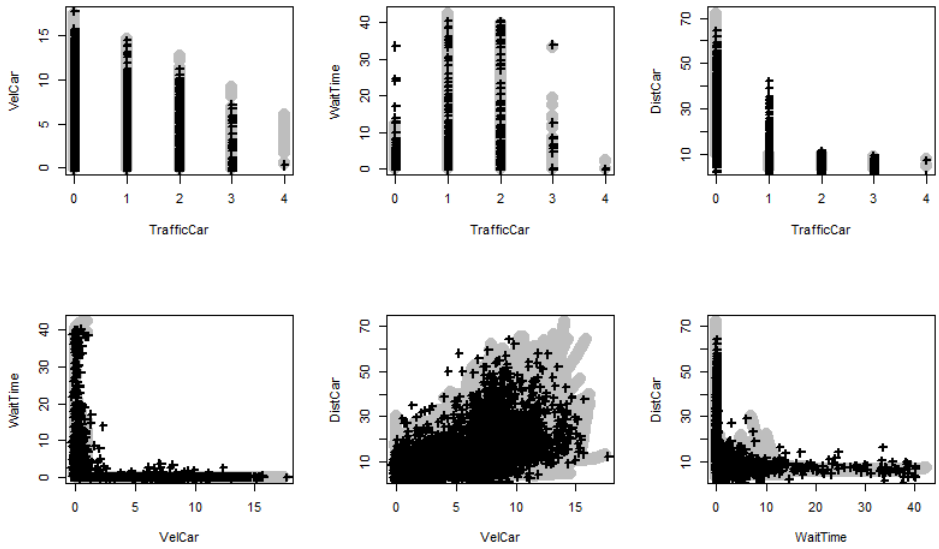

Figure 3. Measured data (grey) and associated simulated data (+) generated with the underlying copula.

4. Conclusions

When it comes to describing stochastic phenomena, such as the traffic behavior, the correlation between the corresponding parameters cannot be neglected, particularly if the stochastic descriptions of the parameters are needed for ongoing analysis, such as for the validation and verification of autonomous driving and assistance systems. In this study, copulas are used as an approach for not only measuring the dependence and correlation but also to generate and calculate the joint distribution function of multivariate stochastic variables, taking the dependence into account.

The numerical example given in Section 3 describes the application of the theory of copulas on a concrete observation and real-time measurements of a roundabout. Despite considering only four parameters, the speed, traffic density, waiting time (before entering the roundabout and within the roundabout), and the distance to other traffic participants, the usefulness of this approach is illustrated. Additionally, the plausibility of the negative and positive correlation of the parameters is discussed, which is captured via the use of a copula.

It is emphasized that more stochastic parameters can be taken into account, such as the time, since the traffic density may correlate to the time of day, e.g., rush hour at the end of the working day to give an example. Moreover, the weather, road quality, or visibility might be important parameters depending on which questions the downstream analysis tries to address. It is worth mentioning that more data and therefore more hours of recording are necessary to extract these parameters and to consider their interferences. The given approach may therefore be seen as an intial attempt to generate random variables or samples according to their dependencies, extracted from real traffic data, without the claim of the completeness of all influential parameters.

Based on the given approach, a further analysis has to be performed in the future and is not considered in this work. To provide an example of how often an autonomous vehicle causes a crash, it can be investigated with a simulated reliability analysis via, e.g., Monte Carlo simulations. It is expected that by considering the joint distribution with all correlations extracted from the real-time data, the reliability analysis is more accurate and closer to reality. However, as there are no accidents in the drone dataset, the comparison between simulation and reality is not feasible. This highlights the need for more data and more recordings.

This approach can also be helpful when it comes to studying the behavior of swarms of traffic participants and their interactions, similarly to [46], especially when transfering this behavior from real-world to a virtual or simulated environment.

Author Contributions: Writing—original draft, K.L., T.N. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "Verifikations- und Validierungsmethoden automatisierter Fahrzeuge im urbanen Umfeld". The authors would like to thank the consortium for the successful cooperation.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Behrensdorf, J.; Broggi, M.; Beer, M. Reliability Analysis of Networks Interconnected With Copulas. *ASME J. Risk Uncertain. Part B Dec.* **2019**, *5*, 041006. [CrossRef]
2. Andrade, D.F.; Barbetta, P.A.; De Freitas Filho, P.J.; De Mello Zunino, N.A.; Jacinto, C.M.C. Using Copulas in Risk Analysis. In Proceedings of the 2006 Winter Simulation Conference, Monterey, CA, USA, 3–6 December 2006; pp. 727–732.
3. Tang, X.S.; Li, D.Q.; Zhou, C.B.; Zhang, L.M. Bivariate distribution models using copulas for reliability analysis. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **2013**, *227*, 499–512. [CrossRef]
4. Zhao, D.; Lam, H.; Peng, H.; Bao, S.; LeBlanc, D.J.; Nobukawa, K.; Pan, C.S. Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 595–607. [CrossRef] [PubMed]
5. Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **1959**, *8*, 229–231.
6. Czado, C. Analyzing Dependent Data with Vine Copulas. In *Lecture Notes in Statistics*; Springer: Berlin/Heidelberg, Germany, 2019.
7. Joe, H. *Dependence Modeling with Copulas*; Taylor & Francis: Abingdon, UK, 2014.
8. Li, Y.; Cai, Y.; Fu, Q.; Wang, X.; Li, C.; Liu, Q.; Xu, R. A stochastic modeling approach for analyzing water resources systems. *J. Contam. Hydrol.* **2021**, *242*, 103865. [CrossRef] [PubMed]
9. Kielmann, J.; Manner, H.; Min, A. *Stock Market Returns and Oil Price Shocks: A CoVaR Analysis Based on Dynamic Vine Copula Models*; Graz Economics Papers; Department of Economics and Department of Public Economics and University of Graz: Graz, Austria, 2021. Available online: https://ideas.repec.org/p/grz/wpaper/2021-01.html (accessed on 5 November 2022).
10. Brechmann, E.C.; Czado, C.; Aas, K. Truncated regular vines in high dimensions with application to financial data. *Can. J. Stat.* **2012**, *40*, 68–85. [CrossRef]
11. Rank, J. *Copulas: From Theory to Application in Finance*; Bloomberg Financial; Wiley: New York, NY, USA, 2007.
12. Cherubini, U.; Luciano, E.; Vecchiato, W. *Copula Methods in Finance*; The Wiley Finance Series; Wiley: Hoboken, NJ, USA, 2004.
13. Cherubini, U.; Mulinacci, S.; Gobbi, F.; Romagnoli, S. *Dynamic Copula Methods in Finance*; The Wiley Finance Series; Wiley: Hoboken, NJ, USA, 2011.
14. Tran, Q.H. *Copulas im Risikomanagement von Versicherungsunternehmen*; Universität Mannheim: Baden-Württemberg, Germany, 2009. Available online: https://www.grin.com/document/ (accessed on 5 November 2022).
15. Gschlößl, S.; Czado, C. Spatial modelling of claim frequency and claim size in non-life insurance. *Scand. Actuar. J.* **2007**, *2007*, 202–225. [CrossRef]
16. Goodwin, B.K.; Hungerford, A. Copula-based models of systemic risk in US agriculture: Implications for crop insurance and reinsurance contracts. *Am. J. Agric. Econ.* **2015**, *97*, 879–896. [CrossRef]
17. Zhang, X.; Jiang, H. Application of Copula function in financial risk analysis. *Comput. Electr. Eng.* **2019**, *77*, 376–388. [CrossRef]
18. Karimalis, E.N.; Nomikos, N.K. Measuring systemic risk in the European banking sector: A copula CoVaR approach. *Eur. J. Financ.* **2018**, *24*, 944–975. [CrossRef]

19. Sae, T. *Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*; J3016, SAE International Standard; SAE Pub. Inc.: Warrendale, PA, USA, 2014.
20. Wachenfeld, W.; Winner, H. Die Freigabe des autonomen Fahrens. In *Autonomes Fahren: Technische, Rechtliche und gEsellschaftliche Aspekte*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 439–464.
21. International Organization for Standardization. *ISO/PAS 21448: Road Vehicles Safety of the Intended Functionality*; International Organization for Standardization: Geneva, Switzerland, 2019.
22. ISO 26262: Road Vehicles Functional Safety. 2018. Available online: https://www.iso.org/standard/68383.html (accessed on 5 November 2022).
23. Wang, C.; Storms, K.; Winner, H. Online Safety Assessment of Automated Vehicles Using Silent Testing. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 13069–13083. [CrossRef]
24. Nalic, D.; Mihalj, T.; Bäumler, M.; Lehmann, M.; Eichberger, A.; Bernsteiner, S. Scenario based testing of automated driving systems: A literature survey. In Proceedings of the FISITA Web Congress, Online, 24 November 2020.
25. Neurohr, C.; Westhofen, L.; Henning, T.; de Graaff, T.; Möhlmann, E.; Böde, E. Fundamental Considerations around Scenario-Based Testing for Automated Driving. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 121–127.
26. Riedmaier, S.; Ponn, T.; Ludwig, D.; Schick, B.; Diermeyer, F. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access* **2020**, *8*, 87456–87477. [CrossRef]
27. Zhang, X.; Tao, J.; Tan, K.; Törngren, M.; Sánchez, J.M.G.; Ramli, M.R.; Tao, X.; Gyllenhammar, M.; Wotawa, F.; Mohan, N.; et al. Finding Critical Scenarios for Automated Driving Systems: A Systematic Literature Review. *arXiv* **2021**, arXiv:2110.08664..
28. Menzel, T.; Bagschik, G.; Maurer, M. Scenarios for Development, Test and Validation of Automated Vehicles. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1821–1827.
29. Schuldt, F.; Saust, F.; Lichte, B.; Maurer, M.; Scholz, S. Effiziente systematische Testgenerierung für Fahrerassistenzsysteme in virtuellen Umgebungen. In *Automatisierungssysteme, Assistenzsysteme und Eingebettete Systeme Für Transportmittel*; 2013. Available online: https://www.semanticscholar.org/paper/Effiziente-systematische-Testgenerierung-f%C3%BCr-in-Schuldt-Saust/5b0f718c6f1f4c6e98861b2a0f6f8f4449f0255f (accessed on 5 November 2022).
30. Scholtes, M.; Westhofen, L.; Turner, L.R.; Lotto, K.; Schuldes, M.; Weber, H.; Wagener, N.; Neurohr, C.; Bollmann, M.; Körtke, F.; et al. 6-Layer Model for a Structured Description and Categorization of Urban Traffic and Environment. *IEEE Access* **2021**, *9*, 59131–59147. [CrossRef]
31. Akagi, Y.; Kato, R.; Kitajima, S.; Antona-Makoshi, J.; Uchida, N. A Risk-index based Sampling Method to Generate Scenarios for the Evaluation of Automated Driving Vehicle Safety. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 667–672.
32. Wagner, S.; Groh, K.; Kuhbeck, T.; Dorfel, M.; Knoll, A. Using Time-to-React based on Naturalistic Traffic Object Behavior for Scenario-Based Risk Assessment of Automated Driving. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1521–1528.
33. Romeo, J.S.; Tanaka, N.I.; Pedroso-de Lima, A.C. Bivariate survival modeling: A Bayesian approach based on Copulas. *Lifetime Data Anal.* **2006**, *12*, 205–222. [CrossRef]
34. Biau, G.; Wegkamp, M. A note on minimum distance estimation of copula densities. *Stat. Probab. Lett.* **2005**, *73*, 105–114. [CrossRef]
35. Fermanian, J.D.; Scaillet, O. Nonparametric Estimation of Copulas for Time Series. *J. Risk* **2003**, *5*. [CrossRef]
36. Joe, H. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In *Distributions with Fixed Marginals and Related Topics (Seattle, WA, 1993)*; IMS Lecture Notes Monogr. Ser.; Inst. Math. Statist.: Hayward, CA, USA, 1996; Volume 28, pp. 120–141.
37. Bedford, T.; Cooke, R.M. Vines–a new graphical model for dependent random variables. *Ann. Stat.* **2002**, *30*, 1031–1068. [CrossRef]
38. Aas, K.; Czado, C.; Frigessi, A.; Bakken, H. Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **2009**, *44*, 182–198. [CrossRef]
39. Wilson, R.J. *Introduction to Graph Theory*; Prentice Hall/Pearson: New York, NY, USA, 2010.
40. Nagler, T.; Vatter, T. *Rvinecopulib: High Performance Algorithms for Vine Copula Modeling*; R Package Version 0.6.2.1.0; 2022. Available online: https://rdrr.io/cran/rvinecopulib/man/vinecop_dist.html (accessed on 5 November 2022).
41. Dißmann, J.; Brechmann, E.; Czado, C.; Kurowicka, D. Selecting and estimating regular vine copulae and application to financial returns. *Comput. Stat. Data Anal.* **2013**, *59*, 52–69. [CrossRef]
42. Wang, F.; Li, H. Distribution modeling for reliability analysis: Impact of multiple dependences and probability model selection. *Appl. Math. Model.* **2018**, *59*, 483–499. [CrossRef]
43. Krajewski, R.; Moers, T.; Bock, J.; Vater, L.; Eckstein, L. The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–6.
44. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1929–1934.

45. Bondzio, L.; Ortlepp, J.; Scheit, M.; Voß, H.; Weinert, R. Verkehrssicherheit innerörtlicher Kreisverkehre. Forschungsbericht vi 05, Gesamtverband der Deutschen Versicherungswirtschaft e. V., Unfallforschung der Versicherer, 2012. Available online: https://www.udv.de/resource/blob/79712/5df1ae6ec093c99b3cefe818daa703de/18-verkehrssicherheit-inneroertlicher-kreisverkehre-data.pdf (accessed on 5 November 2022).
46. Mba, J.C.; Mai, M.M. A Particle Swarm Optimization Copula-Based Approach with Application to Cryptocurrency Portfolio Optimisation. *JRFM* **2022**, *15*, 285. [CrossRef]