Empirical Software Engineering (2027) 32:10
https://doi.org/10.1007/s10664-026-10920-5

Does road diversity really matter in testing automated driving systems?

Stefan Klikovits[1] · **Vincenzo Riccio**[2] · **Ezequiel Castellano**[3] · **Ahmet Cetinkaya**[4] · **Alessio Gambi**[5] · **Paolo Arcaini**[3]

Received: 13 January 2026 / Accepted: 29 June 2026

Abstract

Context The use of automated driving systems (ADSs) in the real world requires rigorous testing to ensure safety. To increase trust, ADSs should be tested on a large set of diverse road scenarios. Literature suggests that if a vehicle is driven along a set of geometrically diverse roads—measured using various diversity measures (DMs)—it will react in a wide range of behaviours, thereby increasing the chances of observing failures, or strengthening the confidence in its safety, if no failures are observed. However, this assumption has never been tested before, nor have road DMs been assessed for their properties.

Objective Our goal was to perform an exploratory study on 53 currently used and new, potentially promising road DMs. Specifically, our research questions looked into the road DMs themselves, to analyse their properties (e.g. *monotonicity*, *computation efficiency*), and to test correlation between DMs. Furthermore, we investigated the use of road DMs to determine whether the assumption that diverse test suites of roads expose diverse driving behaviour holds.

Method Our empirical analysis relies on a state-of-the-art, open-source ADS testing infrastructure and uses a data set containing over 97,000 individual road geometries and matching simulation data that were collected using two driving agents. By considering test suites of various sizes and measuring their roads' geometric diversity, we studied road DM properties, the correlation between road DMs, and the correlation between road DMs and the observed behaviour.

Results Our findings reveal a strong correlation between road diversity and behavioural diversity, confirming that geometrically diverse test suites systematically exercise diverse driving behaviours. We identified Dist. Entropy and Summing aggregations as most effective, with Feature Map achieving the strongest correlation of 0.95 while requiring minimal computation time. The analysed measures maintain robust correlation with behavioural diversity across test suites containing roads of varying lengths, eliminating the need for length normalisation.

Conclusions These results empirically validate the fundamental assumption underlying diversity-driven ADS testing: road geometry diversity serves as a reliable proxy for behavioural diversity. For practitioners, we recommend Feature Map or Dist. Entropy as optimal

Communicated by: Phil McMinn.

Extended author information available on the last page of the article

choices, whilst Averaging-based measures should be avoided entirely. The near-identical correlation patterns observed across architecturally different driving agents indicate that our findings generalise beyond specific ADS implementations, providing a solid foundation for diversity-driven test generation and selection.

Keywords Road diversity measure · Autonomous driving systems · Behaviour diversity · Scenario-based testing

1 Introduction

Automated driving systems (ADSs) are expected to drastically change the transportation industry by reducing the number of accidents, avoiding traffic congestion, and lowering fuel consumption. Nonetheless, reports of collisions involving ADSs (DMV California 2022; US Department of Transportation, National Highway Traffic Safety Association 2022) and fatalities (Bachman 2025) emphasise the need for extensive validation and testing before they can be safely released onto public roads.

Thorough testing explores, i.e. *covers*, different aspects of the system under test (SUT)'s behaviour (Laurent et al. 2023), and thus has the potential to find bugs and increase the confidence in the SUT's correctness (Fraser and Rojas 2019). However, due to the complexity of the SUT, it is generally difficult to directly find test inputs that maximise behavioural diversity (BD). Although initial approaches have been proposed (e.g. Cheng et al. 2023), a generally accepted approach to increase testing cost-effectiveness is to generate test suites that maximise *test diversity* (e.g. Chen et al. 2004; Bueno et al. 2014; Feldt et al. 2016). The underlying assumption of generating diverse tests is that the more diverse the tests within a test suite (TS) are, the higher the SUT's BD will be, and thus, more of its functionality will be exercised (Xie and Memon 2006). Consequently, diversity-driven approaches such as Novelty search (Lehman and Stanley 2011) have been applied to generate tests (Feldt and Poulding 2017; Zohdinasab et al. 2021).

Likewise, in the ADS domain many testing approaches aim to generate test suites of driving scenarios that feature various combinations of road structure (e.g. road geometry, lane markings), traffic participants (e.g. vehicles, pedestrians), and other environmental factors (e.g. weather, lighting) (Tang et al. 2023; Huai et al. 2023; Luo et al. 2021). Arguably, roads are one of the most fundamental aspects of a driving scenario, as other aspects will typically be expressed in reference to their geometry (e.g. "overtaking on a straight/curvy road"). Thus, a crucial problem is to create a TS of diverse road geometries, to test as many vehicle behaviours as possible. This TS should ideally cover a range of different straights, bends, curves, and turns of various degrees of sharpness. The testing of ADSs is typically coverage-based, aiming to maximise the number of tested road shapes.[1] An adequate *diversity measure (DM)* should therefore reflect the prioritisation of variety of road geometries. Figure 1 illustrates this concept over a set of simple roads. This interpretation of diversity, which is rooted in Software Testing, is different than the one from other disciplines. For instance, in Machine Learning, diversity in training data refers to "equality of distributions", which is

[1] Note that this research investigates road diversity. Test suite optimisation, such as minimality or execution efficiency, is orthogonal and considered future work.

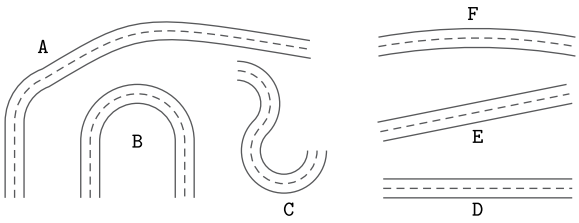


Fig. 1 Example – Different road geometries of varying diversity. The figure displays a set of six road geometries, where roads `A` , `B` , `C` and `D` will very likely provoke very different driving behaviours, due to the differing curvatures and sharpnesses of turns. Road `E` , on the other hand, is a minimally rotated version of `D` , which therefore should not change a purely curvature-based DM when added to the TS. `F` , on the other hand, is similar but slightly bent. Even though its average curvature is comparable to `D` and `E` , it will most likely induce different ADS driving behaviour ("a smooth, long, continuous turn"). An adequately sensitive DM should therefore reflect the addition of `F` to the TS

required to avoid biased datasets. Such a view is also common in some biological and natural settings (Leinster and Cobbold 2012), e.g. animal and plant population measurements.

Existing work proposed several DMs to express the diversity of roads, including measures based on *Jaccard similarity index* (Gambi et al. 2019), *Iterative Levenshtein distance* (Riccio and Tonella 2020), *discrete Fréchet distance* (Mosig and Clausen 2005), and various other road features (e.g. *curvature*, *complexity*, and *direction coverage*) (Zohdinasab et al. 2021; Sadat et al. 2021).

To the best of our knowledge, the choice of DM has only been reported in published papers, but never justified. Furthermore, the underlying concept of road diversity measures seems to have never been studied before in the context of ADS testing. This raises the paramount questions of "*how much road geometries matter in testing ADSs*?" and "*how to measure the diversity of a set of test roads?*" and motivates our research in finding a set of DMs that can be confidently used as objective measures of road diversity. Our research aims to answer the fundamental questions: *"Which DMs are well-suited for use in ADS testing?"* and *"is there any overlap (correlation) between the individual measures?"*. To answer these questions, we analyse the DMs' properties and check if they are indeed good indicators for vehicle behavioural diversity.

1.1 Desired Properties of Diversity Measures

We believe that, in order to be useful for testing, a DM should guarantee certain properties which provide guidance towards the achievement of testing goals akin traditional coverage criteria, such as code coverage, which guide automatic test generation (Fraser and Arcuri 2011). We expect adequate road DMs to have the following properties.

– *Monotonicity and growth* (Weyuker 1986): A test suite's diversity should never decrease when adding roads. Monotonicity is easily achieved by classical coverage criteria in which the test requirements to cover are fixed and known in advance. On the contrary, in domains like ADS, in which test requirements might not be known in advance, ensuring monotonicity is far from trivial. In addition to monotonicity, another aspect that is interesting to assess is *how* a DM grows as the number of test cases included in a TS increases. Indeed, a DM that grows for each individual test is able to discriminate better

among roads than a DM that, although monotonic, increases only for some specific tests. Being able to discriminate better among roads is desirable, as it can possibly lead to test the driving agent in different conditions.

- *Insensitivity to duplicates* (Weitzman 1992): Additionally, the DM should be insensitive to duplicates (also known as "twin property" (Weitzman 1992)) in TS, i.e. the DM value should neither increase nor decrease[2] if the same road is added twice.[3] Such property is important for two reasons. First, requiring that the diversity does not increase with duplicates, guarantees that one cannot inflate the diversity in an artificial way, without triggering any possible different behaviour of the ADS (see discussion on behaviour diversity in Section 1.2). Second, requiring that the diversity does not decrease with duplicates, guarantees that a test suite is not penalised for being suboptimal in terms of size. Regarding the latter point, note that the *minimality of the test suite* is another desired property in testing that, however, is an orthogonal concern that should be treated independently and should not be accounted for by a DM.
- *Efficiency*: A DM should be also efficient to compute, i.e. computing a test suite's DM should take much less time than executing the test suite. While computing the coverage of classical coverage criteria is usually fast (as it consists of checking the coverage of each test requirement individually), computing a road DM could be expensive, as it may require pairwise comparison of all roads.
- *Additivity*: Where computational complexity is inevitable, a minimal property that should be guaranteed is that the DM computation should be additive, i.e. any new road added to the test suite should not require the re-evaluation of the diversity of all the roads, but only of the currently added road.

1.2 Correlation with Behavioural Diversity

In software testing, one of the most desired properties of structural coverage criteria is the ability to expose different behaviours of the SUT (Fraser and Rojas 2019). This also holds in our context, where higher DM values should effectively correlate with the diversity of observed behaviours. If a DM is *insensitive*, it would consider roads that actually trigger different types of ADS behaviours as similar: if this is the case, by relying on the coverage of the roads alone, some of those behaviours would not be triggered. On the other hand, a road DM that is *too sensitive* would flag similar roads as different, so wrongly expecting them to trigger different ADS behaviours. Additionally, such a DM would lead users to build unnecessary large test suites that do not increase BD.

1.3 Conducted Study

With the exploratory study proposed in this paper, we aim to fill the lack of road DM research and investigate their effectiveness in ensuring an ADS's behaviour coverage. We therefore analyse the properties of a total of 53 DMs (see Section 3.2.3) that either have been used for measuring road diversity in previous research on testing, or have been used

[2] Absence of decrease is already required by monotonicity.

[3] As an analogy, code coverage metrics are naturally insensitive to duplicates: for example, adding a test that covers the exact statements covered by another test, does not change the statement coverage level, but only affects the minimality of the test suite size.

to measure diversity in related domains, such as measuring geometric diversity of lines and curves (Agarwal et al. 2014), or diversity in populations (Weitzman 1992). We perform our evaluation empirically on an extensive data set of roads. Our study also provides quantitative estimates of effect sizes (e.g. value growth and efficiency) using real data, which are difficult to obtain through purely analytical methods.

2 Background and Related Work

The notion of diversity has been considered in various forms for the testing of ADSs. Here, we first introduce these different methods in Section 2.1. Then, in Section 2.2, we provide an overview of DMs that can be used to quantify the diversity of road geometries in a test suite.

2.1 Diversity in ADS Testing

Several works on ADS testing aim to generate driving scenarios cost-effectively (Zhong et al. 2021). Most of these works aim to achieve driving scenarios' diversity to avoid generating too similar scenarios that might expose the same issues multiple times.

Ben Abdessalem et al. (2018); Tuncali et al. (2018); Zhu et al. (2022); Majumdar et al. (2021), and Zhong et al. (2023) define diversity in terms of differences in the scenario configuration space. The scenario configuration space includes various parameters such as the initial position and speed of vehicles and pedestrians, the road layout, the placement of scenery elements and obstacles, as well as weather and lighting conditions. While Ben Abdessalem et al. (2018) and Tuncali et al. (2018) did not measure how much scenarios differ, Zhu et al. (2022) used Euclidean distance to assess how different the scenarios are in the configuration space. Majumdar et al. (2021), instead, defined diversity based on the overall dispersion of the scenario parameters, whereas Zhong et al. (2023) proposed to discriminate diverse tests only if they are at a certain distance in the configuration space.

Riccio and Tonella (2020) designed an Iterative Levenshtein distance to evaluate test diversity. However, differently from the work mentioned above, they considered only geometric properties of roads, i.e. the road shape. Other notable works that considered only road properties as test diversity discriminant are the ones by Zohdinasab et al. (2021); Nguyen et al. (2021); Gambi et al. (2022, 2019), and Tang et al. (2021). In particular, Zohdinasab et al. (2021) computed high-level road features, like smoothness or complexity, and represented the tests into two-dimensional grids made of discretised feature values, i.e. feature maps, such that roads mapped to the same map cells are considered similar. Notably, Zohdinasab et al. (2021) also used behavioural features to identify tests that expose similar behaviour of the ADS. This map representation has been later used for test selection (Nguyen et al. 2021), and test adequacy assessment (Gambi et al. 2022). Gambi et al. (2019) and Tang et al. (2021), instead, discriminated tests based on whether they take place on roads and intersections made of similar road segments.

2.2 Diversity Measures for Road Geometry

Diversity-based testing has been a vital tool for software validation for several decades, leading to a large number of diversity measures (Elgendy et al. 2025). In the ADS context,

literature described several methods for computing DMs for roads. Most approaches that take road geometry into account measure test diversity by aggregating similarity metrics computed over pairs of roads. Furthermore, other domains such as geometry and population diversity, produced potentially viable DM methods. We categorise these methods as (1) methods that can be computed by *aggregating pairwise distances* between roads, and (2) methods that can be *computed directly* by using the information of each individual road in a test suite.

Table 1 lists the approaches in each category. Diversity computation methods that belong to the first category are formed as a pair of a distance function and an aggregation method.

In the following, we overview existing distance functions and then explain various methods that can be used to aggregate the distances between road pairs to compute test suite diversity.

2.2.1 Pairwise Distance Functions

To the best of our knowledge, the following functions are commonly used for measuring dissimilarities between roads, (typically represented as curves):

Fréchet and Discrete Fréchet Distances Fréchet (Alt and Godau 1995; Aronov et al. 2006) and discrete Fréchet (Mosig and Clausen 2005; Agarwal et al. 2014) distance functions are commonly used to measure distance between curves. Intuitively, when two vehicles move along two roads, the *maximum point-to-point distance* between them depends on the speed that they move. Fréchet distance corresponds to the shortest of all possible maximum point-to-point distances that can be obtained by varying the speeds. Discrete Fréchet distance provides an efficient way of approximating Fréchet distance for the case where the curves are described as sequences of straight segments. These distances have been used for characterising road dissimilarity (Castellano et al. 2021) and the dissimilarity of the paths vehicles can take (Fan et al. 2010; Weise and Mostaghim 2021).

Partial Curve Mapping Distance Partial curve mapping distance between curves representing two roads is defined as the sum of the discrepancy between the segments of two polygons representing normalised versions of the curves (Witowski and Stander 2012).

Table 1 Diversity measures - outline

Distance functions	Aggregation methods	Direct DMs
Discrete Fréchet distance	Weitzman aggregation	Test set diameter CS
Partial curve mapping	Distance Entropy	Convex hull of curves
Dynamic time warping	Summing	Feature map coverage
Normalised relative angle	Averaging	
Complexity vectors	Averaging of maxima	
Iterative Levenshtein distance		
Jaccard similarity index		
Area between curves		
Manhattan distance		

Dynamic Time Warping Distance Dynamic time warping was originally proposed for computing dissimilarity between temporal sequences where entries correspond to data obtained at consecutive time instants (Berndt and Clifford 1994), but it has also been commonly used for checking dissimilarity between curves (Munich and Perona 1999; Efrat et al. 2007).

Relative Angle and Normalised Relative Angle Distances Relative angles and normalised relative angles between curves as defined by Vlachos et al. (2004) provide useful methods for computing pairwise distance measures that do not change when the curve representing one of the roads is rotated. This rotation invariance property is especially useful when testing trajectory planners (Werling et al. 2010; McNaughton et al. 2011; Zhu and Aksun-Guvenc 2020) utilised in ADSs, as they depend on the sharpness of the turns on a road but not the entire orientation of the road (e.g. whether the road goes north or east).

Distance Based on Complexity Vectors In Sadat et al. (2021), roads were identified as sequences of smaller sections called frames. Then, curvatures and curvature-derivatives were used to compute the so-called complexity vectors for each frame. Given a pair of roads with multiple frames, the distance function of Sadat et al. (2021) finds the maximum of the distances between the complexity-wise closest frames of the roads.

Iterative Levenshtein Distance Curves representing roads can be described as lists of connected straight segments. The rotational differences between the orientation of consecutive segments form sequences of angles. Iterative Levenshtein distance for a pair of roads is defined in Riccio and Tonella (2020) as the Levenshtein *edit distance* (Levenshtein et al. 1966) between the sequences of angles identified for each of the roads.

Jaccard Similarity Index Jaccard similarity index (JS) for a pair of roads is defined in Gambi et al. (2019) by considering the sets of segments in those roads. JS takes a value between 0 and 1 corresponding to the ratio of the number of common segments of the roads to the total number of segments in the union of segment sets. Thus, the value $1 - \mathrm{JS}$ can be used as a distance function quantifying the dissimilarity between a pair of roads. The definition of the road segments directly affects the sensitivity of JS. A coarse definition of road segments will be less sensitive than a finer-grained one, as many road segments with similar curvature and length will be clustered together. To account for different levels of sensitivity of JS, we consider in the following $\mathrm{Jaccard}_{\mathrm{coarse}}$ and $\mathrm{Jaccard}_{\mathrm{fine}}$. $\mathrm{Jaccard}_{\mathrm{coarse}}$ considers three categories of road segments (Straight, Left-turn, and Right-turn), three levels of sharpness for left and right turns (i.e., gentle, moderate, sharp) and three classes of length (i.e., small, medium, long). $\mathrm{Jaccard}_{\mathrm{fine}}$, instead, considers the same three categories of road segments, but distinguishes between 36 levels of sharpness (turning angle between 5 and 180 degrees) and 5 classes of length, thereby enabling the metric to distinguish many different types of turns.

Area Between Curves The area between curves is a geometric measure that, casually speaking, measures "how much space fits" between two curves (Jekel et al. 2019). This measure allows fast computation of the pairwise distance between road curves.

Manhattan Distance over Feature Vectors Zohdinasab et al. (2021) used Manhattan distance between feature vectors extracted from two roads as a measure of dissimilarity between

them. The adopted features are: 1. *Smoothness* which indicates how smooth the turns are (from 0 to 5); 2. *Complexity* which characterises the shape of road, from almost straight (0) to a road with a lot of turns (5); 3. *Orientation* which indicates how many directions (i.e., N, NE, E, SE, S, SW, W, NW) are covered by the road.

2.2.2 Aggregation Methods

After a distance function is used on each pair of roads in a test suite, an aggregation method can be used for combining the distances to obtain a single value representing the DM. The properties of such a DM thus depend not only on the pairwise distance function but also on the method of aggregation. We report below aggregation methods commonly used to measure diversity:

Weitzman Aggregation Weitzman (1992) proposed an aggregation method by considering ideal properties of the relationship between diversity measures and underlying pairwise distance functions. The computation of the proposed aggregation method involves set-based recursions, and is known to be slow in general.

Aggregation Through Summing Distances between roads can be aggregated by taking their sum. This method is also used e.g. in biology to measure population diversity and corresponds to *functional attribute diversity* (Chiu and Chao 2014).

Aggregation using Distance Entropy Distance entropy was proposed by Shi et al. (2016) as a method for quantifying the diversity of a test suite for applications in software testing. When adapted to our problem setting, this method first constructs a weighted relationship graph of roads by using their pairwise distances. The aggregation is then achieved by computing the entropy of the weights in the minimum spanning tree of the relationship graph.

Aggregation Through Averaging all and Averaging Maximum Distances Averaging-based aggregation methods were previously used by Castellano et al. (2021) and Zohdinasab et al. (2021). In particular, Castellano et al. (2021) considered average of all pairwise distances between roads as an aggregation method in diversity computation. Moreover, Zohdinasab et al. (2021) considered a more general setting where pairwise distances are defined for general test cases (not just roads). In the case of roads, the aggregation method used by Zohdinasab et al. (2021) corresponds to calculating the average of the distances from each road to the road that it is most distant to (i.e. averaging maximum distances).

2.2.3 Direct Computation DMs

There are two diversity quantification methods that do not rely on distances between roads.

Test Set Diameter Compression Size (CS) Test set diameter was introduced by Feldt et al. (2016) as a method for direct computation of DMs. It is linked to normalised compression distance for multisets (Calò et al. 2020a, b), which uses Kolmogorov complexity of elements in a set. Building on this concept, we implement a version of the test set diameter that

is suited to our specific scenario. We first convert each curve representing a road into a list of integers representing its curvatures and arc-length values, and then apply zlib compression (Gailly and Adler 2026) to the concatenation of the lists for all roads. We call the size of the compressed binary data Test Set Diameter CS.

Convex Hull of Curves Area of the convex hull encompassing all road curves is a DM that can be computed directly without comparing roads with each other. Previously, convex hulls of curves have been used by Castellano et al. (2022) for checking if roads fit into a given map.

Feature Map Coverage Feature map coverage, introduced by Zohdinasab et al. (2021), is a direct diversity measure that quantifies test suite diversity by counting occupied cells in a discretised feature space. Each road is mapped to a cell in an N-dimensional grid based on interpretable features extracted from the road geometry, such as direction coverage (the number of angular sectors covered by road segments), maximum curvature, and turn count. The mapping function $x_i = \lfloor \alpha_i \cdot f_i \rfloor$ assigns each road to its corresponding cell, where α_i is a scaling factor determining the granularity. The diversity of a test suite is then simply the count of cells containing at least one road. This approach has been adopted as the primary diversity metric in the SBST/SBFT Tool Competition for cyber-physical systems testing (Gambi et al. 2022).

2.2.4 Analysis of the Relationship between Different DMs

The relationship between different DMs can be characterised through correlation of the diversity values that they assign to test suites. In some special cases, correlation coefficients can be analytically derived. For instance, for the same underlying distance function, DMs obtained with aggregation through summation and averaging are perfectly correlated (with correlation coefficient 1), since one is a scaled version of the other. Analytical correlation analysis becomes more challenging when DMs use nonlinear aggregation methods or nonlinear operations in diversity calculations. This point is further discussed in the context of diversity of species by DeBenedictis (1973). There, it is mentioned that while it is possible to show positive correlation between DMs, the exact value of the correlation coefficient is hard to derive analytically in many cases. In such cases, numerical methods are typically used (see, e.g. Mouchet et al. 2010).

Another aspect of correlation analysis of DMs is that correlation of aggregation-based DMs depends also on the relationship between the underlying distance functions. We note that for some DMs, the correlation between distance functions is preserved through aggregation. In particular, the correlation coefficient of two DMs defined by summing all pairwise distances obtained respectively with two different distance functions is equivalent to the correlation coefficient of the distance functions themselves. However, the correlation coefficient is hard to obtain analytically, since there is no straightforward transformation between distance functions for curves. In Jekel et al. (2019), some of these distance functions—partial curve mapping, discrete Fréchet, dynamic time warping, and curve length—were compared through an empirical study on an optimisation problem.

3 Research Design

Given the importance of DMs in the testing of ADSs, an objective comparison of the methods is of vital interest. Specifically, we are interested in the properties of DMs, as well as whether and how strongly they are correlated. Additionally, we also would like to investigate the relationship between (road) DMs and behavioural diversity (BD) of the vehicle. To this extent, we select 53 DMs (see Section 3.2.3) and perform the—to the best of our knowledge—first exploratory study of DMs for road geometry.

In particular, we are interested in learning more about the individual behaviour of the various DMs used in literature (cf. Section 2) and how they compare. Our analysis is structured in four primary research questions (RQs) that investigate the computational behaviour of an individual DM (***RQ1***), the correlation of DM pairs (***RQ2***), their correlation with regard to road length (***RQ3***) and their correlation with the behaviour diversity of autonomous driving agents (***RQ4***). In the following, we will provide details of the RQs (Section 3.1), followed by a description of our study setup (Section 3.2).

3.1 Research Questions

3.1.1 RQ1 – DM Properties

Monotonicity and Growth For some DMs and aggregation methods such as convex hull of curves, Weitzman aggregation and aggregation through summation, monotonicity can be mathematically proved. For these DMs and aggregation methods, we report the theoretical results known from the literature. For other DMs and aggregation methods such as test set diameter and aggregation using distance entropy, instead, assessing monotonicity is more challenging as no theoretical results are known from the literature. For this latter category, we perform an empirical evaluation.

Moreover, in case a measure is monotonic, we are also interested in assessing "how and how much" it grows; specifically, we are interested in assessing to what extent each new test increases TS diversity. Indeed, measures for which each novel (non-duplicated) test increases the DM value are better at discriminating among different tests than DMs for which only a few, highly diverse tests lead to a noticeable change of the DM value; better discrimination among tests is desirable, as it can lead to exercise different ADS behaviours (think of adding road `F` in Fig. 1). Note that in the case of DMs via aggregation, diversity largely depends on the underlying complex distance functions, and, as a result, a precise mathematical assessment of the growth is hard to achieve. This is the case even for measures for which we can analytically check monotonicity (e.g. Weitzman aggregation Weitzman (1992)). Therefore, to quantify the growth, we perform an empirical evaluation. Namely, we evaluate the relative DM growth when adding new roads to the TS. Intuitively, smaller TSs should, on average, have a larger growth in diversity, due to the larger probability of adding a highly diverse road. Nonetheless, even large test suites' DM values should reflect the addition of individuals.

Insensitivity to Duplicates Similarly to monotonicity, for some DMs, this property can be determined directly from the DM definition whereas for other DMs an empirical approach is more suitable. In particular, mathematical properties of the distance functions can be used to show that DMs obtained through Weitzman aggregation and aggregation through summation are insensitive to duplication, but DMs obtained through averaging-based aggregation methods do not possess the insensitivity property, as duplication can decrease diversity for those DMs. Moreover, in case insensitivity to duplicates is not guaranteed, it is not possible to mathematically establish to what extent a DM is sensitive to duplication; therefore, we assess this effect by means of an empirical study.

Efficiency Although asymptotic efficiency of a DM can be obtained analytically, it cannot predict concrete results on real data. Therefore, we conduct an empirical assessment to precisely compare the different DMs and provide an initial evaluation of the DMs' computational efficiency.

Additivity Additivity measures the time it takes to re-calculate a TS's DM after being extended. As for efficiency, only an asymptotic analysis of additivity is possible starting from the DMs definition. To have a more realistic comparison of the different DMs, we conduct an empirical study.

3.1.2 RQ2 – Pairwise Correlation of DMs

Using the DMs calculated for each TS, we perform correlation analyses on the totality of the data and grouped by TS size. As discussed in Section 2.2.4, analytical derivation of correlation coefficients for certain pairs of DMs is prohibitively complex. For this reason, we perform an empirical evaluation of correlations. For instance, if two DMs are not or only weakly correlated, it could mean that they measure different characteristics of road geometries and that their use in combination might be more beneficial than using either DM individually. Instead, if two DMs are strongly correlated, it may hint at redundancy, and the properties of ***RQ1*** should be used to choose a better-suited one.

3.1.3 RQ3 – Correlation of DMs and Road Length

Here we analyse the effect of road length on the individual DMs. The results of this analysis allow determining whether there is an effect of the road length on the diversity. Combined with the results of ***RQ4.b*** (see below), it helps users in selecting appropriate road lengths for ADS testing.

3.1.4 RQ4 – Correlation of DMs and BD

We analyse whether there is a correlation between the individual DMs and the BD. Unfortunately, due to the complexity of ADSs, it is typically and generally not possible to define a reliable mathematical model of an ADS's behaviour from which a correlation can be analytically deduced. Hence, we empirically study such a correlation and test the sensitivity

of ADS behaviour to road diversity. This RQ aims to discover those DMs that are actually linked to BD, thus helping testers avoid using DMs that focus on irrelevant road features.

RQ4.a First, we analyse the general correlation between road diversity (i.e. DMs) and BD. Remember that as BD is different for the individual driving agents (DAs), we perform the analyses separately.

RQ4.b We study the effect of road length on the correlation between DMs and BD, to evaluate if for TSs with short roads, there is a stronger correlation between DMs and BD.

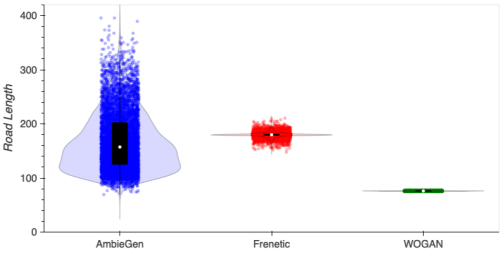
3.2 Empirical Study Setup

3.2.1 Road Data Set

In the `SBST'2022 Tool Competition` (Gambi et al. 2022), competitors provide search algorithms that generate non-intersecting two-lane roads which an autonomous driving agent should follow. Each generated road is simulated and provides timestamped observation records of the vehicle's *position*, *velocity*, *steering angle*, *brake* and *throttle* inputs. The roads—cubic interpolations of the Cartesian control points provided by the search algorithms—are handed to a simulator and allow calculation of *length*, *curvature*, *road heading*, *turn count* and aggregations thereof (e.g. *max curvature*), etc. Based on this data, we can extract further information such as *relative heading* w.r.t. the road, *total driven length*, *lateral vehicle position*, as well as *aggregate* values such as minimum, mean, maximum and standard deviation of steering input, as suggested by Jahangirova et al. (2021).

We use a large data set of 97,000 roads produced in the course of the competition as the basis for our research on road diversity. To generalise our data, we use roads generated randomly, as well as by three road generators, namely `AmbieGen`, `WOGAN` and `Frenetic` (see Gambi et al. 2022), and simulation data provided by executing these roads using two autonomous driving agents (see Section 3.2.2).

Diversity of Roads by Road Generators As the data was originally produced by the competitors of the `SBST'2022 Tool Competition`, we first assess some basic road features. For instance, we know that the three Road Generators in the competition—`AmbieGen`, `Frenetic` and `WOGAN` —produce roads of different lengths, as shown in Fig. 2.



(a) Distribution plot

	Mean	Std	Min	Median	Max
AmbieGen	168.34	54.07	69.26	157.27	395.80
Frenetic	179.70	4.66	147.62	179.79	210.59
WOGAN	76.00	0.58	75.00	75.98	77.73

(b) Statistical measures

Fig. 2 Road length by generator

Remark 1 (Online Tool & Data Availability) We refer the reader to http://stklik.github.io/2026-EMSE-Road-Diversity/, where we host digital, interactive versions of all figures, as well as additional plots.

Data Quality A preliminary look into the data showed that both the road data and the simulation information is of high quality. The created roads have been analysed for self-intersections and maximum curvature by the `SBST` pipeline at creation time. Furthermore, we also checked that the simulation data is complete, i.e. the observation data for vehicle simulation (position, velocity, acceleration, steering/throttle/brake inputs, etc.) are available for every record, and the recordings have high enough frequency (e.g. between 5Hz and 20Hz). Nonetheless, we observed that a limited number of simulations have potentially invalid data. For instance, we discovered a small number of "faulty" simulations where the vehicle started driving in the wrong direction. We thoroughly analysed the data and removed such simulations before running our experiments analysis on the data. To this extent, we implemented automatic ways to identify and remove these individuals from the data set, and we also manually checked the test samples. Furthermore, next to the road and simulation data itself, the `SBST` pipeline also produced and recorded some statistical information (driving direction coverage, road curvature measures) for each simulation. We adapted our scripts to reproduce this data using our own (independent) implementation, thereby increasing confidence in our code and correctness of the existing data.

Remark 2 (Notation) In the following evaluation section, we use the following notational styles to denote the TSs of different sizes.

- We use e.g. TS_{10} to denote a test suite of size 10.
- For evaluation of ***RQ1*** (*Monotonicity and growth*), we use e.g. $\mathrm{TS}_{50+20\%}$ to denote TS_{50}, where each TS is enlarged with 20% extra roads.
- For evaluation of ***RQ1*** (*Efficiency*), we use e.g. TS_{20+1} to denote TS_{20}, where each TS enlarged with one extra road.
- Similarly, for evaluation of ***RQ1*** (*Duplicates*), we use e.g. $\mathrm{TS}_{20+10\%-Dup}$ to denote TS_{20}, where each TS has 10 % duplicate roads.
- Finally, when referring to an individual TS, we use e.g. TS_{10}^{i} to denote the *i*-th TS of size 10.

TS Creation, Data Description and Pre-evaluation The analysis of both road and behavioural diversity (for ***RQ4***) is based on the calculation of metrics for TSs. A TS is a set of randomly selected roads of fixed size. Our TSs are sampled with varying sizes of 10, 20, 50 and 100 roads. To generalise our results, we aim to perform each computation on 100 TS of each size.[4]

From the 97,000 roads, we initially created 100 TSs. Our evaluation requires TSs of different sizes (10, 20, 50, 100) and for ***RQ1*** we further need to evaluate the growth when adding 10%, 20%, 50%, and 100% of roads to each TS. To avoid using duplicate roads, we

[4] Even though we used powerful computation infrastructure, certain DMs (e.g. those using Weitzman aggregation) are computationally (very) expensive for large TSs. We therefore set a maximum computation time of 24 hours and adjusted our analysis accordingly.

create distinct TSs, $\forall \mathrm{TS}^i, \mathrm{TS}^j : \mathrm{TS}^i \cap \mathrm{TS}^j = \varnothing$. This also holds for any variants of test suites (e.g. $\mathrm{TS}^i_{50+100\%} \cap \mathrm{TS}^j_{50+100\%} = \varnothing$). Thus, there is no overlap for any TSs of the same size.

We apply this procedure for all 100 TSs:

1) We first sample a set of 200 roads for the overall largest test suite variant required (i.e. 100 roads + 100% growth) from the total of 97,000.
2) Then, we sample the next smaller TS of roads (in this case, 100 + 50% growth) as subset of the larger TS, s.t. e.g. $\mathrm{TS}^i_{100+50\%} \subset \mathrm{TS}^i_{100+100\%}$.

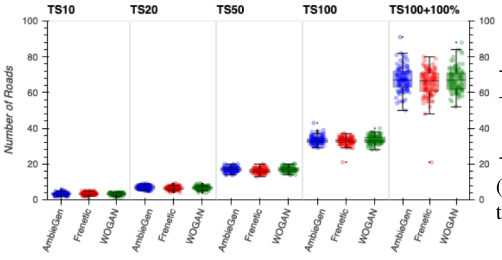
3) We follow this approach for all other TSs in decreasing TS-(variant) size, generally, by transitivity we get

$$\forall i \in [1, 100],$$
$$\mathrm{TS}^i_{10} \subset \left(\mathrm{TS}^i_{10+1} = \mathrm{TS}^i_{10+10\%}\right) \subset \left(\mathrm{TS}^i_{10+2} = \mathrm{TS}^i_{10+20\%}\right) \subset \mathrm{TS}^i_{10+50\%} \subset \mathrm{TS}^i_{10+100\%} =$$
$$\mathrm{TS}^i_{20} \subset \mathrm{TS}^i_{20+1} \subset \left(\mathrm{TS}^i_{20+2} = \mathrm{TS}^i_{20+10\%}\right) \subset \mathrm{TS}^i_{20+20\%} \subset \mathrm{TS}^i_{20+50\%} \subset \mathrm{TS}^i_{20+100\%} \subset$$
$$\mathrm{TS}^i_{50} \subset \mathrm{TS}^i_{50+1} \subset \mathrm{TS}^i_{50+2} \subset \mathrm{TS}^i_{50+10\%} \subset \mathrm{TS}^i_{50+20\%} \subset \mathrm{TS}^i_{50+50\%} \subset \mathrm{TS}^i_{50+100\%} \subset$$
$$\mathrm{TS}^i_{100} \subset \mathrm{TS}^i_{100+1} \subset \mathrm{TS}^i_{100+2} \subset \mathrm{TS}^i_{100+10\%} \subset \mathrm{TS}^i_{100+20\%} \subset \mathrm{TS}^i_{100+50\%} \subset \mathrm{TS}^i_{100+100\%}$$

As the three Road Generators produce roads of different length, we adapted our TS sampling strategy to create TSs with an even distribution of roads by each generator. This means that each TS contains (roughly) a third of its roads generated by each road generator. The empirical effects of this sampling strategy are shown in Fig. 3, which shows the distribution plot (Fig. 3a) and the mean fraction of roads by generator (Fig. 3b).

3.2.2 Driving Agents

To increase the generalisability of our results, we use driving simulations produced by two independent autonomous DAs, i.e. `BeamNG.AI` and `Dave2`. Both considered agents are widely used in the literature (Gambi et al. 2019; Riccio and Tonella 2020; Zohdinasab et al. 2021; Gambi et al. 2022; Kong et al. 2020; Zhou et al. 2020; Humeniuk and Khomh 2025) and automatically perform the lane keeping task. They are, however, conceptually different (rule-based vs deep learning-based) and, thus, show different driving behaviour.



(a) Number of roads per TS for different TS sizes

Generator	TS_{10}	TS_{20}	TS_{50}	TS_{100}	$TS_{100+100\%}$
AmbieGen	31.90 %	33.15 %	33.88 %	33.64 %	33.60 %
Frenetic	34.60 %	32.55 %	32.52 %	32.89 %	32.62 %
WOGAN	33.50 %	34.30 %	33.6 %	33.47 %	33.78 %

(b) Mean relative frequency of roads by generator per TS

Fig. 3 Frequency distribution of roads by generator for all TSs

While the results of ***RQ1–3*** are independent of the specific DAs, for the analysis of ***RQ4*** we need to take the difference between DAs into account. Thus, we separately analysed the correlation of DMs and BD for each DA.

`BeamNG.AI` is the DA shipped with the `BeamNG.tech`[5] driving simulator. It defines the ego-car's trajectory before the simulation by leveraging a perfect knowledge of the road's geometry. In particular, `BeamNG.AI` plans a trajectory that maximises the car's speed (within pre-defined speed limits), while keeping the vehicle within the right lane as much as possible.

`Dave2` exploits a deep learning architecture consisting of three convolutional and five fully-connected layers (Bojarski et al. 2016). In particular, it learns a direct mapping from the on-board sensor camera input to the steering angle value to be passed to the ego-car's actuators. This means that it does not require previous knowledge of the road geometry.

Due to their different decision logics, the two DAs can differ in the trajectories they produce for different roads. Moreover, the variety of behaviours produced by the two DAs may differ. Therefore, also the correlation with road diversity may differ. For these reasons, it is important to select such two different DAs to guarantee generalisability of the conclusions.

In our study, we performed all our analyses that take driving behaviour into account (i.e. ***RQ4***) w.r.t. each specific agent.

Behavioural diversity The metric for BD aims to express how much of a vehicle's behavioural range is covered within a TS. In this work, the behaviour of a vehicle on a road is characterised from simulation traces using the vehicle's *velocity*, *throttle*, *braking*, *steering* input and *lateral (out-of-bounds) position* on the road. Each simulation of a driving agent on a road produces records of these values in regular intervals (roughly every 0.1 seconds). Through aggregation, we calculate the *mean* (μ), *minimum*, *maximum* and *standard deviation* (σ) values of each of the five observations as reported by Jahangirova et al. (2021), yielding a total of 20 values. As the framework starts all simulations with zero throttle, brake and throttle, this value is constant for all simulations. We therefore drop these respective measures' *minimum* from the BD calculation to avoid tarnishing the results.

For the actual comparison of BD of two road simulations, we adopt these 17-dimensional feature maps (split into 10 bins each), as done in prior ADS testing work (Gambi et al. 2022). Then, we quantify BD as the number of covered bins.

Rotational Adjustment When thinking of road geometry, we typically only consider the shape of the road itself, but do not take its position and orientation into account. Thus, a perfectly straight road leading north will be seen as equivalent to a straight road of the same length leading east or west. To avoid being misled, we should therefore move all roads to the same starting location and also align them, before calculating the DMs. Nonetheless, some DAs such as `Dave2` (which was trained on image data), take positioning of the sun and ego's own shadow into account for their behaviour. Thus,

[5] https://beamng.tech/

arguably, roads with the same geometry but different heading should be distinguished when analysing DMs for `Dave2`.

For the rotation, we did *data preprocessing* in the form of a *Procrustes analysis* (Dryden and Mardia 2016), leading to two transformations. First, the roads were relocated so that their starting points match. Second, the roads were rotated around the initial point with the angle of rotation obtained through an optimisation procedure (Dryden and Mardia 2016). These steps ensure that two roads with identical shapes yield zero distance.

3.2.3 Research Subjects

The subjects of our research are the 53 diversity measures described in Section 2, by aggregating ten pairwise and three direct distance metrics using five aggregation methods. Note that we study Jaccard Similarity in two sensitivity versions (fine and coarse).

We performed our analyses on all 53 DMs, to get a complete picture of the DM landscape. Note that some properties for certain DMs can be deduced e.g. due to the nature of their aggregation function (monotonicity is for instance not assured when using *averaging*). Nonetheless, as we additionally aimed to investigate the effect size, we calculated these properties for all DMs.

Similarly, even though one might intuitively suspect certain DMs to be strongly correlated (***RQ2***), our goal is to test this hypothesis in a practical setting. The information on which and how strongly the DMs are correlated is of special interest, since it allows future users to avoid computing redundant DMs, i.e. those that produce very similar (if not equivalent) results.

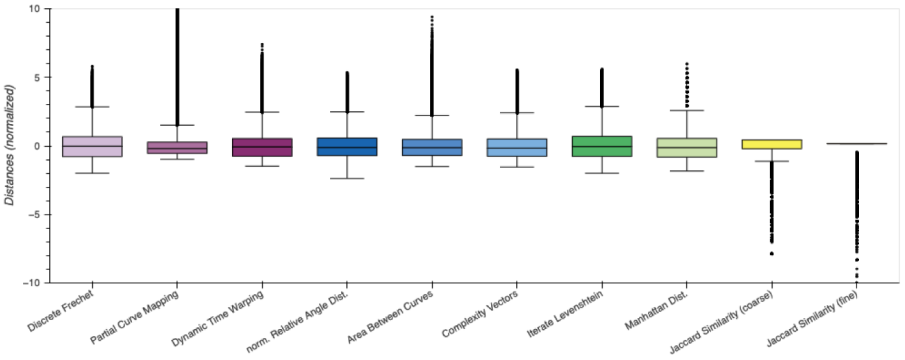
3.2.4 Pairwise Distance Measures

Before computing the full DM, we first analyse the pairwise distance measures described in Section 2: we compute for every TS, the pairwise distances between all non-duplicated roads (i.e. all 19 900 combinations in each $\mathrm{TS}_{100+100\%}$) and analyse the value distribution (Fig. 4) and computation times (Fig. 5).

Pairwise Distances We easily observe in Fig. 4a that the value normalised distributions for the pairwise distances between the roads are similar for most measures. The notable difference (cf. Fig. 4b) is the Jaccard Similarity—both the coarse and fine variants—which shows comparatively low standard deviation. Moreover, these two measures are the only ones whose outliers are below the boxplot distribution.

On the other hand, we remark the "long tails" of outliers for certain pairwise measures, such as Partial Curve Mapping and Area Between Curves. Note that Partial Curve Mapping even has outliers reaching almost 75σ (cf. Fig. 4c), while the outliers of Area Between Curves "only" reach about 9.41 times the std. $\mathrm{Jaccard}_{\mathrm{fine}}$'s and $\mathrm{Jaccard}_{\mathrm{coarse}}$'s outliers reach to -16.8σ and -7.87σ on the negative side.

Finally, we look at the coefficient of variation (CV), which measures the standard deviation as a proportion of the mean value (i.e. $\frac{\sigma}{\mu}$), thus indicating "how spread out" the data is. We notice that $\mathrm{Jaccard}_{\mathrm{fine}}$ and $\mathrm{Jaccard}_{\mathrm{coarse}}$ have remarkably low CV values, indicating



(a) Z-score normalised distance measures. Note the reduced y-axis (y-range: $[-10, 10]$). Not all outliers are shown.

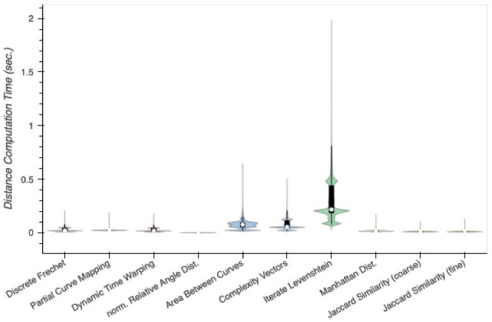
	Mean	Std	CV	Min	Median	Max
Discrete Fréchet	80.30	40.20	0.50	0.12	79.26	313.70
Partial Curve Mapping	82.59	84.41	1.02	0.11	65.70	6412.82
Dynamic Time Warping	5699.81	3845.26	0.67	3.80	5395.63	34201.66
norm. Relative Angle Dist.	0.04	0.02	0.41	0.00	0.04	0.12
Area Between Curves	3453.89	2283.10	0.66	3.28	3133.10	24937.32
Complexity Vectors	0.03	0.02	0.64	0.00	0.02	0.12
Iterative Levenshtein	87.28	43.71	0.50	0.00	84.90	332.28
Manhattan Dist.	5.40	2.94	0.54	0.00	5.00	23.00
$\text{Jaccard}_{\text{coarse}}$	0.95	0.12	0.13	0.00	1.00	1.00
$\text{Jaccard}_{\text{fine}}$	0.99	0.06	0.06	0.00	1.00	1.00

(b) Statistical measures

Min	Median	Max
-1.99	-0.03	5.81
-0.98	-0.20	74.99
-1.48	-0.08	7.41
-2.38	-0.13	5.35
-1.51	-0.14	9.41
-1.54	-0.17	5.54
-2.00	-0.05	5.61
-1.84	-0.14	5.98
-7.87	0.43	0.43
-16.80	0.16	0.16

(c) Statistical measures (norm.)

Fig. 4 Pairwise distance measures



(a) Distribution (in sec.)

	Mean	Std	CV	Min	Median	Max
Discrete Fréchet	0.03	0.02	0.59	0.01	0.02	0.20
Partial Curve Mapping	0.02	0.00	0.10	0.02	0.02	0.19
Dynamic Time Warping	0.03	0.02	0.59	0.01	0.02	0.18
norm. Relative Angle Dist.	0.00	0.00	0.27	0.00	0.00	0.01
Area Between Curves	0.08	0.04	0.47	0.02	0.08	0.63
Complexity Vectors	0.08	0.04	0.59	0.02	0.06	0.49
Iterative Levenshtein	0.30	0.18	0.59	0.08	0.22	1.93
Manhattan Dist.	0.02	0.00	0.27	0.01	0.02	0.16
$\text{Jaccard}_{\text{coarse}}$	0.01	0.00	0.33	0.00	0.01	0.10
$\text{Jaccard}_{\text{fine}}$	0.01	0.00	0.28	0.00	0.01	0.12

(b) Statistical measures (in sec.)

Fig. 5 Pairwise distance computation times (in seconds)

rather stable measurements, i.e. without much variation. Although the distance measure alone does not have direct implications for the later-calculated DMs, a "too compact" distance metric might hinder a detailed diversity analysis. We notice that Dynamic Time Warping, Area Between Curves and Complexity Vectors have rather large (> 0.6) CV, while Partial Curve Mapping even exceeds 1.0.

Computation Time In terms of computation time (see Fig. 5), we notice that the individual pairwise distance measurements on average take well below one second. Nonetheless, we note significant differences in measurement times, such that a single comparison with Iterative Levenshtein takes up to two seconds, with an average of 0.29 seconds, while the norm. Relative Angle Dist. takes at most 0.01 seconds.

This is a non-negligible difference in performance, as the pairwise distance metrics need to be calculated before aggregation to DMs can be performed. Thus, as TSs grow—and thus exponentially the number of pairwise measures between roads—the choice of distance measure is vital.

Table 2 provides a statistical evaluation for the accumulated computation time for each distance measure, aggregated for each TS size separately. A graphical representation of the distribution plots is depicted in Fig. 10 to 13 in Appendix A. We notice that Iterative Levenshtein is indeed several times slower, which, for large TSs has a significant impact. The computation of the pairwise measures alone takes on average almost 25 minutes (1461 seconds). In comparison, norm. Relative Angle Dist. takes at most 8.11 seconds for the same calculations.

We could coarsely classify the distance measures into four groups according to their computation times:

1) norm. Relative Angle Dist. is absolutely the fastest;
2) $\text{Jaccard}_{\text{coarse}}$, $\text{Jaccard}_{\text{fine}}$ and Manhattan Dist. are one order of magnitude (or roughly seven to ten times) slower than norm. Relative Angle Dist.;
3) Discrete Fréchet, Partial Curve Mapping and Dynamic Time Warping take twice to three times as much time as the previous set;
4) Area Between Curves and Complexity Vectors add another factor three compared to the previous measures;
5) Iterative Levenshtein is yet another five times slower than Area Between Curves and Complexity Vectors.

We conclude that, depending on the TS size, the computation of pairwise measures alone shows significant differences in performance. While some distance measures compute fast, certain measures—especially Iterative Levenshtein, but also Area Between Curves and Complexity Vectors —might become prohibitively slow for large test suite sizes.

4 Empirical Evaluation

In this section, we describe the empirical evaluations and provide the detailed protocol for the experiments for each RQ.

Computation Hardware All measurements were executed on JKU Linz's *relise* cluster.[6] Each compute node comprises two quad-core Intel Xeon X5570 processors (2.93 GHz), 48 GB of RAM, running Debian GNU/Linux 10.3 (Buster). To ensure reproducibility and avoid interference from parallel processing, each measurement was executed on a single

[6] https://relise.jku.at/

Table 2 Total distance computation times by test suite size (in seconds)

	(a) TS10						(b) TS20					
	Mean	Std	CV	Min	Median	Max	Mean	Std	CV	Min	Median	Max
Discrete Fréchet	1.31	0.22	0.17	0.87	1.30	2.07	5.49	0.70	0.13	3.54	5.49	7.28
Partial Curve Mapping	1.05	0.03	0.03	0.98	1.05	1.14	4.42	0.10	0.02	4.18	4.43	4.65
Dynamic Time Warping	1.21	0.21	0.17	0.80	1.20	1.89	5.05	0.64	0.13	3.27	5.04	6.71
norm. Relative Angle Dist.	0.07	0.00	0.06	0.06	0.07	0.08	0.29	0.01	0.05	0.26	0.29	0.34
Area Between Curves	3.53	0.53	0.15	2.67	3.41	5.94	14.68	1.63	0.11	11.16	14.65	20.32
Complexity Vectors	3.38	0.58	0.17	2.23	3.38	5.35	14.16	1.81	0.13	9.07	14.11	19.03
Iterative Levenshtein	13.23	2.28	0.17	8.76	13.18	21.05	55.42	6.99	0.13	35.80	55.20	74.33
Manhattan Dist.	0.73	0.06	0.08	0.60	0.74	0.93	3.08	0.20	0.07	2.57	3.09	3.68
$\text{Jaccard}_{\text{coarse}}$	0.52	0.06	0.11	0.41	0.52	0.70	2.19	0.20	0.09	1.78	2.19	2.65
$\text{Jaccard}_{\text{fine}}$	0.49	0.04	0.09	0.40	0.49	0.64	2.06	0.15	0.07	1.67	2.08	2.42
	(c) TS50						(d) TS100					
Discrete Fréchet	35.39	2.92	0.08	29.60	35.34	42.43	144.70	8.15	0.06	127.22	143.66	165.30
Partial Curve Mapping	28.49	0.45	0.02	27.46	28.51	29.46	115.30	1.51	0.01	111.74	115.54	118.88
Dynamic Time Warping	32.60	2.69	0.08	27.32	32.55	39.16	133.31	7.54	0.06	116.79	132.29	152.03
norm. Relative Angle Dist.	1.89	0.06	0.03	1.71	1.90	2.04	7.69	0.21	0.03	7.16	7.70	8.11
Area Between Curves	94.59	6.48	0.07	82.02	95.27	109.53	383.70	18.18	0.05	336.36	385.09	425.19
Complexity Vectors	91.36	7.62	0.08	77.39	91.66	108.14	373.51	21.99	0.06	328.73	370.86	433.36
Iterative Levenshtein	357.50	29.52	0.08	299.19	356.53	435.14	1461.86	82.85	0.06	1273.82	1456.14	1663.80
Manhattan Dist.	19.86	0.89	0.05	17.58	19.88	22.32	80.63	2.84	0.04	72.50	80.58	86.95
$\text{Jaccard}_{\text{coarse}}$	14.08	1.04	0.07	12.02	14.19	16.56	57.21	4.18	0.07	49.27	58.17	69.61
$\text{Jaccard}_{\text{fine}}$	13.28	0.67	0.05	11.68	13.30	14.81	53.93	2.28	0.04	47.59	53.95	58.84

node with single-threaded execution using only one of the 2 (connected) CPU cores per analysis run. We used `slurm` to execute all experiments in a similar manner. This configuration provides comparable timing results across all DMs.

4.1 *RQ1* – DM Properties

The first RQ analyses the properties of the various DMs individually. More specifically, we observe the DMs' behaviours in terms of growth, insensitivity to duplicates, efficiency and additivity, as described in Section 3.

4.1.1 *RQ1a*: Monotonicity and growth

To quantify the growth empirically, we proceeded as follows:

1) For each of the TSs (grouped by size), we computed the DMs.
2) We added n new roads to each TS, where $n \in \{10\%, 20\%, 50\%, 100\%\}$ of the TS's size and re-compute the DMs.
3) We then analysed the difference before and after extension, to evaluate typical statistical metrics (min, median, the first quartile Q_1) for each DM and TS size.

Thus, the evaluation compares the enlarged TS over the default TS's DM measure for each DM and TS. For simplicity, we computed the median and minimum growth values, as shown in Table 3, and 4, respectively. Blue cells indicate growing values, orange cells shrinking, and white cells negligible fluctuations ($< 0.1\%$ change).

Median Growth Table 3 shows that, on average, the median value grows for most measures, especially those using Dist. Entropy aggregation and Summing. Weitzman also appears to be a good aggregator, even though for larger TS sizes we could not obtain values due to computation timeout. It also appears that the Direct DMs Feature Map, Test Set Diameter CS and Convex Hull are monotonic. As expected, Averaging and Avg. Maxima are not monotonic and decrease for many measures or experience only negligible changes.

We also notice that Dist. Entropy, Summing and Avg. Maxima are moderately growing across all distance metrics and TS sizes, even though Avg. Maxima only grows for certain ones, such as Discrete Fréchet, Partial Curve Mapping, and Dynamic Time Warping.

Minimum Growth Table 4, in contrast, displays the minimum observed values for DM growth. This means that in this table we may observe which DMs experienced monotonicity and growth for all TSs. Here, it is evident that many measures which expose monotonic behaviour when looking at the median value, may, in fact, have reduced values. A notable example is the Test Set Diameter CS, which appears to be monotonic in Table 3, but not in Table 4. Similarly, we may observe that none of the Averaging and Avg. Maxima are, in fact, monotonic. Moreover, Dist. Entropy may be non-monotonic for small increases in TS size (e.g. +10%) for e.g. Partial Curve Mapping and Dynamic Time Warping. Furthermore, it appears that combining Manhattan Dist. and Dist. Entropy causes monotonicity problems.

Table 3 RQ1a – Relative growth (median values)

		TS10				TS20				TS50				TS100			
		+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%
Discrete Fréchet	Dist. Entropy	1.1	1.2	1.51	2.02	1.1	1.18	1.47	1.94	1.09	1.18	1.42	1.89	1.09	1.19	1.45	1.88
	Summing	1.2	1.44	2.27	4.11	1.21	1.44	2.28	4.07	1.21	1.45	2.25	4.07	1.21	1.45	2.26	4.03
	Averaging	0.98	0.98	0.97	0.97	0.99	0.99	0.99	0.99	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.02	1.04	1.0	1.0	1.02	1.05	1.0	1.01	1.02	1.04	1.0	1.0	1.02	1.04
	Weitzman	1.06	1.12	1.31	1.61	1.07	1.12	—	—	—	—	—	—	—	—	—	—
Partial Curve Mapping	Dist. Entropy	1.09	1.19	1.49	1.92	1.1	1.17	1.44	1.95	1.09	1.18	1.43	1.86	1.09	1.19	1.45	1.88
	Summing	1.19	1.44	2.29	4.11	1.2	1.43	2.26	4.01	1.22	1.45	2.27	4.13	1.22	1.45	2.28	4.04
	Averaging	0.98	0.98	0.98	0.97	0.99	0.98	0.99	0.98	1.01	1.0	1.0	1.02	1.0	1.01	1.01	1.0
	Avg. Maxima	1.0	1.01	1.09	1.14	1.0	1.0	1.06	1.11	1.01	1.03	1.07	1.15	1.01	1.03	1.08	1.13
	Weitzman	1.04	1.09	1.3	1.63	1.04	1.08	—	—	—	—	—	—	—	—	—	—
Dynamic Time Warping	Dist. Entropy	1.09	1.21	1.52	2.03	1.1	1.18	1.5	1.95	1.09	1.18	1.44	1.92	1.09	1.2	1.46	1.9
	Summing	1.2	1.43	2.27	4.11	1.2	1.42	2.3	4.1	1.21	1.44	2.25	4.12	1.21	1.46	2.27	4.03
	Averaging	0.98	0.98	0.97	0.97	0.99	0.98	1.0	1.0	1.0	1.0	0.99	1.02	1.0	1.01	1.01	1.0
	Avg. Maxima	1.0	1.0	1.02	1.06	1.0	1.0	1.02	1.07	1.0	1.01	1.01	1.05	1.0	1.01	1.02	1.05
	Weitzman	1.04	1.11	1.32	1.62	1.06	1.11	—	—	—	—	—	—	—	—	—	—
norm. Relative Angle Dist.	Dist. Entropy	1.13	1.27	1.66	2.36	1.11	1.22	1.56	2.13	1.1	1.21	1.51	2.03	1.1	1.2	1.51	2.0
	Summing	1.22	1.46	2.34	4.21	1.22	1.45	2.29	4.13	1.21	1.45	2.27	4.05	1.21	1.44	2.25	4.02
	Averaging	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.0	1.0	1.0	0.99	1.0
	Avg. Maxima	1.0	1.01	1.04	1.08	1.0	1.01	1.04	1.06	1.0	1.01	1.02	1.04	1.0	1.0	1.01	1.02
	Weitzman	1.08	1.16	1.41	1.82	1.08	1.15	—	—	—	—	—	—	—	—	—	—
Area Between Curves	Dist. Entropy	1.1	1.23	1.56	2.14	1.1	1.2	1.55	2.05	1.1	1.2	1.47	2.0	1.1	1.21	1.48	1.93
	Summing	1.2	1.44	2.28	4.12	1.21	1.43	2.3	4.12	1.21	1.44	2.24	4.11	1.21	1.46	2.27	4.04
	Averaging	0.98	0.98	0.98	0.98	0.99	0.99	1.01	1.0	1.0	1.0	0.99	1.02	1.0	1.01	1.0	1.0
	Avg. Maxima	1.0	1.0	1.03	1.07	1.0	1.0	1.03	1.09	1.0	1.0	1.02	1.06	1.0	1.01	1.02	1.04
	Weitzman	1.06	1.13	1.37	1.7	1.07	1.12	—	—	—	—	—	—	—	—	—	—
Complexity Vectors	Dist. Entropy	1.09	1.23	1.57	2.12	1.08	1.17	1.46	1.89	1.08	1.16	1.4	1.8	1.09	1.17	1.42	1.82
	Summing	1.21	1.45	2.39	4.33	1.21	1.46	2.33	4.23	1.21	1.44	2.27	4.03	1.2	1.44	2.24	3.98
	Averaging	0.99	0.99	1.03	1.03	0.99	1.0	1.02	1.03	1.0	1.0	1.0	1.0	0.99	0.99	0.99	0.99
	Avg. Maxima	1.0	1.01	1.08	1.15	1.0	1.01	1.05	1.12	1.0	1.0	1.02	1.05	1.0	1.0	1.01	1.02
	Weitzman	1.06	1.14	1.38	1.74	1.05	1.1	—	—	—	—	—	—	—	—	—	—
Iterative Levenshtein	Dist. Entropy	1.13	1.26	1.67	2.36	1.12	1.22	1.6	2.24	1.11	1.22	1.55	2.14	1.11	1.23	1.56	2.13
	Summing	1.21	1.44	2.28	4.13	1.2	1.44	2.26	4.08	1.21	1.44	2.25	4.07	1.21	1.44	2.26	4.02
	Averaging	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.99	1.0	1.0	0.99	1.01	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.01	1.02	1.03	1.0	1.01	1.01	1.04	1.0	1.0	1.01	1.02	1.0	1.0	1.01	1.01
	Weitzman	1.07	1.15	1.37	1.74	1.08	1.15	—	—	—	—	—	—	—	—	—	—
Manhattan Distance	Dist. Entropy	1.1	1.22	1.53	1.98	1.08	1.17	1.41	1.79	1.08	1.14	1.33	1.58	1.05	1.12	1.27	1.47
	Summing	1.2	1.45	2.32	4.25	1.2	1.44	2.27	4.08	1.21	1.44	2.25	4.03	1.2	1.43	2.25	3.99
	Averaging	0.98	0.99	1.0	1.01	0.99	0.99	0.99	0.99	1.0	1.0	0.99	1.0	0.99	0.99	1.0	0.99
	Avg. Maxima	1.0	1.01	1.03	1.06	1.0	1.01	1.03	1.06	1.0	1.0	1.01	1.04	1.0	1.0	1.02	1.04
	Weitzman	1.05	1.14	1.35	1.67	1.06	1.11	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (coarse)	Dist. Entropy	1.17	1.33	1.77	2.59	1.13	1.25	1.63	2.28	1.11	1.22	1.52	2.05	1.11	1.21	1.51	2.0
	Summing	1.23	1.47	2.34	4.22	1.22	1.45	2.29	4.11	1.21	1.44	2.26	4.04	1.21	1.44	2.26	4.02
	Averaging	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.12	1.22	1.51	2.0	1.1	1.18	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (fine)	Dist. Entropy	1.17	1.33	1.85	2.74	1.14	1.28	1.7	2.42	1.12	1.24	1.62	2.24	1.12	1.24	1.59	2.19
	Summing	1.22	1.47	2.33	4.22	1.22	1.45	2.29	4.1	1.21	1.44	2.27	4.04	1.21	1.44	2.26	4.02
	Averaging	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.11	1.22	1.54	2.07	1.11	1.2	—	—	—	—	—	—	—	—	—	—
Direct DMs	Feature Map	1.1	1.2	1.5	1.9	1.11	1.18	1.42	1.8	1.08	1.15	1.35	1.67	1.06	1.13	1.3	1.55
	Test Set Diameter CS	1.08	1.16	1.39	1.79	1.09	1.17	1.44	1.85	1.1	1.19	1.47	1.94	1.1	1.2	1.48	1.93
	Convex Hull	1.0	1.02	1.12	1.24	1.0	1.03	1.12	1.21	1.01	1.02	1.08	1.13	1.0	1.02	1.05	1.11

Nonetheless, Dist. Entropy shows monotonic behaviour for most distance metrics and TS sizes. Similarly, as expected, Summing is monotonic across all distance metrics. As for Direct DMs, Feature Map appears to be the only one that is monotonic, except for its apparent insensitivity to smaller TS size increases.

Remark 3 For reference, we provide a similar analysis of the 25%-quantile in Table 9 in Appendix B. Readers can refer to the table to see which DMs experience growth in 75% of the measurements, which may offer a tradeoff solution that is growing in most cases.

A richer, more detailed and interactive representation of these growth metrics is available at http://stklik.github.io/2026-EMSE-Road-Diversity/.

Table 4 RQ1a – Relative growth (minimum observed values)

		TS10				TS20				TS50				TS100			
		+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%
Discrete Fréchet	Dist. Entropy	1.02	1.06	1.2	1.6	1.01	1.05	1.15	1.5	1.03	1.09	1.31	1.62	1.03	1.1	1.3	1.6
	Summing	1.15	1.33	1.93	3.55	1.16	1.34	1.98	3.56	1.17	1.36	2.12	3.69	1.18	1.39	2.11	3.66
	Averaging	0.94	0.9	0.83	0.84	0.96	0.92	0.86	0.87	0.96	0.94	0.94	0.91	0.97	0.96	0.94	0.91
	Avg. Maxima	0.97	0.95	0.93	0.92	0.98	0.96	0.94	0.92	0.98	0.98	0.97	0.97	0.99	0.98	0.98	0.97
	Weitzman	1.0	1.04	1.1	1.32	1.01	1.04	—	—	—	—	—	—	—	—	—	—
Partial Curve Mapping	Dist. Entropy	0.99	1.03	1.13	1.51	1.0	1.04	1.16	1.44	1.01	1.06	1.25	1.6	1.03	1.11	1.3	1.58
	Summing	1.08	1.18	1.64	2.63	1.13	1.28	1.93	3.32	1.15	1.35	2.01	3.35	1.17	1.38	2.01	3.34
	Averaging	0.88	0.8	0.7	0.62	0.93	0.88	0.84	0.81	0.95	0.93	0.89	0.83	0.96	0.95	0.89	0.83
	Avg. Maxima	0.93	0.87	0.78	0.77	0.94	0.92	0.89	0.84	0.95	0.94	0.87	0.88	0.96	0.96	0.94	0.86
	Weitzman	1.0	1.02	1.04	1.12	1.01	1.02	—	—	—	—	—	—	—	—	—	—
Dynamic Time Warping	Dist. Entropy	0.97	0.98	1.1	1.42	0.99	1.01	1.11	1.43	1.01	1.08	1.25	1.51	1.02	1.08	1.28	1.58
	Summing	1.12	1.28	1.87	3.29	1.13	1.29	1.9	3.33	1.14	1.33	2.06	3.47	1.16	1.36	2.06	3.53
	Averaging	0.92	0.87	0.8	0.78	0.93	0.88	0.83	0.81	0.94	0.92	0.91	0.86	0.95	0.95	0.91	0.88
	Avg. Maxima	0.96	0.95	0.93	0.92	0.97	0.96	0.94	0.93	0.98	0.98	0.96	0.97	0.99	0.99	0.98	0.98
	Weitzman	1.0	1.02	1.07	1.2	1.01	1.02	—	—	—	—	—	—	—	—	—	—
norm. Relative Angle Dist.	Dist. Entropy	0.98	1.1	1.21	1.73	0.96	1.05	1.22	1.7	1.03	1.11	1.4	1.84	1.03	1.12	1.38	1.81
	Summing	1.15	1.32	1.96	3.49	1.15	1.36	2.08	3.77	1.17	1.38	2.15	3.78	1.18	1.39	2.17	3.84
	Averaging	0.94	0.9	0.84	0.83	0.95	0.94	0.91	0.92	0.97	0.95	0.95	0.94	0.97	0.97	0.96	0.95
	Avg. Maxima	0.97	0.95	0.94	0.9	0.98	0.97	0.96	0.94	0.98	0.97	0.97	0.96	0.99	0.98	0.97	0.98
	Weitzman	1.02	1.08	1.19	1.45	1.04	1.08	—	—	—	—	—	—	—	—	—	—
Area Between Curves	Dist. Entropy	0.92	1.01	1.16	1.5	1.0	1.02	1.17	1.49	1.02	1.09	1.28	1.57	1.03	1.11	1.31	1.62
	Summing	1.12	1.27	1.83	3.2	1.13	1.27	1.86	3.22	1.13	1.32	2.07	3.48	1.15	1.35	2.03	3.46
	Averaging	0.92	0.87	0.78	0.76	0.93	0.88	0.81	0.78	0.94	0.91	0.91	0.86	0.95	0.94	0.9	0.86
	Avg. Maxima	0.96	0.94	0.91	0.91	0.97	0.96	0.93	0.94	0.97	0.98	0.97	0.97	0.98	0.98	0.98	0.98
	Weitzman	1.0	1.02	1.1	1.27	1.0	1.02	—	—	—	—	—	—	—	—	—	—
Complexity Vectors	Dist. Entropy	1.03	1.04	1.26	1.56	1.03	1.08	1.26	1.55	1.05	1.11	1.3	1.58	1.06	1.13	1.35	1.7
	Summing	1.15	1.31	1.94	3.15	1.15	1.3	1.86	3.12	1.17	1.35	2.06	3.56	1.17	1.36	2.04	3.6
	Averaging	0.94	0.89	0.83	0.75	0.94	0.9	0.81	0.76	0.96	0.93	0.91	0.88	0.97	0.94	0.9	0.9
	Avg. Maxima	0.96	0.95	0.92	0.9	0.97	0.96	0.93	0.92	0.98	0.98	0.97	0.96	0.99	0.98	0.97	0.97
	Weitzman	1.01	1.03	1.12	1.25	1.02	1.04	—	—	—	—	—	—	—	—	—	—
Iterative Levenshtein	Dist. Entropy	1.04	1.12	1.34	1.86	1.03	1.12	1.35	1.86	1.05	1.15	1.44	1.9	1.05	1.15	1.42	1.82
	Summing	1.16	1.37	2.05	3.54	1.17	1.36	2.06	3.56	1.18	1.39	2.14	3.75	1.19	1.4	2.15	3.76
	Averaging	0.95	0.93	0.88	0.84	0.97	0.94	0.9	0.87	0.98	0.96	0.95	0.93	0.98	0.97	0.95	0.94
	Avg. Maxima	0.97	0.97	0.95	0.95	0.98	0.96	0.95	0.94	0.98	0.97	0.96	0.98	0.98	0.98	0.98	0.97
	Weitzman	1.02	1.04	1.15	1.38	1.02	1.06	—	—	—	—	—	—	—	—	—	—
Manhattan Distance	Dist. Entropy	0.9	0.97	1.0	1.16	0.91	0.95	1.06	1.39	0.96	1.01	1.08	1.34	0.97	0.97	1.11	1.23
	Summing	1.15	1.31	1.93	3.14	1.15	1.34	1.92	3.42	1.17	1.36	2.07	3.58	1.18	1.38	2.11	3.77
	Averaging	0.94	0.89	0.83	0.74	0.94	0.92	0.84	0.83	0.96	0.94	0.92	0.89	0.97	0.96	0.93	0.94
	Avg. Maxima	0.97	0.95	0.93	0.9	0.98	0.97	0.96	0.96	0.99	0.98	0.96	0.97	0.99	0.99	0.97	0.98
	Weitzman	1.0	1.0	1.08	1.18	1.0	1.02	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (coarse)	Dist. Entropy	1.0	1.14	1.39	1.81	1.0	1.06	1.33	1.73	1.06	1.14	1.41	1.82	1.05	1.14	1.41	1.82
	Summing	1.19	1.42	2.2	4.01	1.2	1.43	2.23	4.0	1.2	1.43	2.23	3.95	1.2	1.43	2.24	3.97
	Averaging	0.97	0.97	0.94	0.95	0.99	0.98	0.97	0.98	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.0	1.08	1.28	1.55	1.0	1.04	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (fine)	Dist. Entropy	1.0	1.13	1.5	2.28	1.07	1.19	1.49	2.12	1.05	1.15	1.48	2.05	1.07	1.18	1.5	2.05
	Summing	1.2	1.44	2.27	4.12	1.21	1.44	2.27	4.06	1.21	1.44	2.25	4.02	1.21	1.44	2.25	4.01
	Averaging	0.98	0.98	0.97	0.98	0.99	0.99	0.99	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.0	1.09	1.33	1.8	1.05	1.15	—	—	—	—	—	—	—	—	—	—
Direct DMs	Feature Map	1.0	1.1	1.3	1.5	1.0	1.0	1.18	1.58	1.0	1.04	1.22	1.43	1.0	1.04	1.13	1.28
	Test Set Diameter CS	0.81	0.88	1.0	1.1	0.81	0.88	1.1	1.23	0.89	0.98	1.16	1.43	0.87	0.93	1.12	1.45
	Convex Hull	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.01

Answer to *RQ1a*: Overall, we notice that monotonicity—an important property of a metric—is not absolutely given for many of the analysed measures, as shown by Table 4. Nonetheless, considering the typical behaviour of the DMs, as depicted by their median (and moreover their 25%-quantile) behaviour, we may consider many of them as useful indicators.

Still, there are numerous DMs—especially those building on Dist. Entropy, Summing and Weitzman, but also the Feature Map Direct DM —that provide monotonicity, and adequate growth behaviour.

4.1.2 *RQ1b*: Insensitivity to duplicates

To assess the insensitivity to duplicates, and empirically evaluate its behaviour, we proceeded as follows:

1) We calculate the TSs' DMs.
2) We then randomly duplicate n roads in each TSs, where n corresponds to $n \in \{10\%, 20\%\}$ of the TS's size and re-compute the DMs.
3) We then check if any DM's value changed and analyse the difference in magnitude before and after the extension, using standard descriptive statistics (mean $\pm$ standard deviation).

Table 5 displays the mean and standard deviation of the relative DM value change after adding the duplicates. A completely insensitive DM would thus show a blue cell valued 1.0 ± 0.0 in the table.[7] When analysing the table, we clearly see that most DMs values change when we add 10% or 20% duplicated roads. Evidently, this is not surprising for aggregations such as Summing or Averaging. We also notice that the Avg. Maxima aggregator—despite being sensitive to duplicates—only experiences minor shifts. This suggests that, despite its problematic characteristics, it may still be useful in certain situations. Similarly, the combination of Avg. Maxima and $\text{Jaccard}_{\text{fine}}$ /$\text{Jaccard}_{\text{coarse}}$ appears to be insensitive to the addition of duplicates.

We furthermore observe that Dist. Entropy and Weitzman are (mostly) insensitive to duplicates. Interestingly, however, the combinations of Dist. Entropy and Manhattan Dist., and Dist. Entropy and Partial Curve Mapping do change when duplicates are added. As for the Direct DMs, we see that Feature Map and Convex Hull are insensitive to duplicates, while the Test Set Diameter CS changes significantly.

> **Answer to *RQ1b*:** Generally, it appears that Dist. Entropy and Weitzman are good aggregators for most DMs, even though for specific distance measures (i.e. Partial Curve Mapping, Manhattan Dist.) they experience duplicate sensitivity. Other aggregators (Summing, Averaging, Avg. Maxima) typically change when adding duplicates. Feature Map and Convex Hull appear to be insensitive to duplicates.

A more detailed analysis of each DM, including individual violin plots, is provided online: http://stklik.github.io/2026-EMSE-Road-Diversity/.

4.1.3 *RQ1c*: Efficiency

For each TS, we recorded the time it took to compute each DM. Due to the total of 100 TSs of each size, we then confidently analysed and compared the DM computation times by size, listing their aggregated information (mean, standard deviation, median).

We point out that our evaluation does not aim at the theoretical best/worst case complexity (e.g. in $\mathcal{O}$-notation), but instead compares the effective computation time, as displayed in Table 6. For pairwise-distance-based metrics, the table entries combine the computation of the respective pairwise measures plus the time for their aggregation. For "aggregator-only"

[7] Note that some cells show 1.0 ± 0.0, yet are coloured in orange. This is due to rounding.

Table 5 RQ1b – Insensitivity to duplicates — Relative change (mean ± stdev)

		TS10		TS20		TS50		TS100	
		+10%	+20%	+10%	+20%	+10%	+20%	+10%	+20%
Discrete Fréchet	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.04	1.43 ± 0.06	1.2 ± 0.03	1.43 ± 0.04	1.21 ± 0.02	1.43 ± 0.03	1.21 ± 0.02	1.44 ± 0.02
	Averaging	0.98 ± 0.03	0.98 ± 0.04	0.99 ± 0.03	0.98 ± 0.03	0.99 ± 0.02	0.99 ± 0.02	1.0 ± 0.01	1.0 ± 0.02
	Avg. Maxima	1.0 ± 0.01	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.0	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Partial Curve Mapping	Dist. Entropy	0.99 ± 0.02	0.97 ± 0.02	0.99 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.99 ± 0.0	0.99 ± 0.01
	Summing	1.2 ± 0.08	1.42 ± 0.15	1.21 ± 0.08	1.46 ± 0.18	1.21 ± 0.05	1.45 ± 0.08	1.21 ± 0.04	1.44 ± 0.05
	Averaging	0.98 ± 0.06	0.97 ± 0.11	1.0 ± 0.06	1.01 ± 0.13	1.0 ± 0.04	1.0 ± 0.05	1.0 ± 0.03	1.0 ± 0.04
	Avg. Maxima	1.02 ± 0.1	1.02 ± 0.17	1.03 ± 0.15	1.09 ± 0.32	1.04 ± 0.17	1.08 ± 0.21	1.04 ± 0.18	1.07 ± 0.22
	Weitzman	1.03 ± 0.09	1.04 ± 0.11	1.03 ± 0.09	1.07 ± 0.18	—	—	—	—
Dynamic Time Warping	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.06	1.43 ± 0.08	1.2 ± 0.05	1.43 ± 0.07	1.2 ± 0.03	1.43 ± 0.05	1.2 ± 0.02	1.43 ± 0.03
	Averaging	0.98 ± 0.05	0.98 ± 0.06	0.99 ± 0.04	0.98 ± 0.05	0.99 ± 0.03	0.99 ± 0.03	0.99 ± 0.02	0.99 ± 0.02
	Avg. Maxima	1.0 ± 0.01	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.0	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
norm. Relative Angle Dist.	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.19 ± 0.03	1.41 ± 0.04	1.2 ± 0.02	1.43 ± 0.03	1.2 ± 0.01	1.43 ± 0.02	1.21 ± 0.01	1.44 ± 0.01
	Averaging	0.98 ± 0.02	0.96 ± 0.03	0.99 ± 0.02	0.98 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	1.0 ± 0.01	1.0 ± 0.01
	Avg. Maxima	1.0 ± 0.02	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.0	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Area Between Curves	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.07	1.43 ± 0.09	1.2 ± 0.05	1.43 ± 0.08	1.2 ± 0.03	1.43 ± 0.05	1.2 ± 0.02	1.43 ± 0.03
	Averaging	0.98 ± 0.05	0.98 ± 0.06	0.99 ± 0.04	0.98 ± 0.05	0.99 ± 0.03	0.99 ± 0.03	0.99 ± 0.02	0.99 ± 0.02
	Avg. Maxima	1.0 ± 0.02	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Complexity Vectors	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.19 ± 0.05	1.41 ± 0.07	1.2 ± 0.04	1.42 ± 0.06	1.21 ± 0.03	1.43 ± 0.04	1.21 ± 0.02	1.44 ± 0.03
	Averaging	0.98 ± 0.04	0.96 ± 0.05	0.99 ± 0.03	0.98 ± 0.04	0.99 ± 0.02	0.99 ± 0.03	1.0 ± 0.02	1.0 ± 0.02
	Avg. Maxima	1.0 ± 0.02	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Iterative Levenshtein	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.03	1.42 ± 0.05	1.2 ± 0.02	1.43 ± 0.04	1.2 ± 0.01	1.43 ± 0.03	1.21 ± 0.01	1.44 ± 0.02
	Averaging	0.98 ± 0.03	0.97 ± 0.04	0.99 ± 0.02	0.98 ± 0.03	0.99 ± 0.01	0.99 ± 0.02	1.0 ± 0.01	1.0 ± 0.01
	Avg. Maxima	1.0 ± 0.01	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.0	1.0 ± 0.01
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Manhattan Distance	Dist. Entropy	0.99 ± 0.02	0.99 ± 0.03	0.99 ± 0.02	0.98 ± 0.03	0.99 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	0.97 ± 0.02
	Summing	1.2 ± 0.06	1.43 ± 0.08	1.2 ± 0.04	1.43 ± 0.06	1.21 ± 0.03	1.44 ± 0.03	1.21 ± 0.02	1.44 ± 0.02
	Averaging	0.98 ± 0.05	0.97 ± 0.05	0.99 ± 0.03	0.99 ± 0.04	1.0 ± 0.02	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.02
	Avg. Maxima	1.0 ± 0.02	1.0 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.01
	Weitzman	1.0 ± 0.01	1.01 ± 0.02	1.0 ± 0.01	1.01 ± 0.01	—	—	—	—
Jaccard Similarity (coarse)	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.01	1.42 ± 0.01	1.2 ± 0.01	1.43 ± 0.01	1.21 ± 0.0	1.44 ± 0.0	1.21 ± 0.0	1.44 ± 0.0
	Averaging	0.98 ± 0.01	0.97 ± 0.01	0.99 ± 0.0	0.98 ± 0.01	1.0 ± 0.0	0.99 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Avg. Maxima	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Jaccard Similarity (fine)	Dist. Entropy	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Summing	1.2 ± 0.0	1.42 ± 0.0	1.21 ± 0.0	1.43 ± 0.0	1.21 ± 0.0	1.44 ± 0.0	1.21 ± 0.0	1.44 ± 0.0
	Averaging	0.98 ± 0.0	0.97 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	1.0 ± 0.0	0.99 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Avg. Maxima	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Weitzman	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	—	—	—	—
Direct DMs	Feature Map	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	Test Set Diameter CS	1.07 ± 0.03	1.13 ± 0.03	1.08 ± 0.02	1.17 ± 0.03	1.09 ± 0.01	1.19 ± 0.02	1.1 ± 0.01	1.2 ± 0.01
	Convex Hull	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

timings, we refer to Table 10 in Appendix C which separates distance and aggregation computations. Finally, Direct DMs naturally show the total computation time only.

Note on Weitzman Aggregator We may right away notice that for the aggregated DMs, the timing differences between Dist. Entropy, Summing, Averaging and Avg. Maxima are negligible. In fact, only Weitzman aggregation shows significant timing differences to the other aggregators across all pairwise measures. The Weitzman aggregator's computational complexity is so large, however, that it is effectively only usable on smaller TS sizes

Table 6 RQ1c – Efficiency – Mean, standard deviation and median of total computation time (in seconds)

		TS10			TS20			TS50			TS100		
		Mean	Std	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median
Discrete	Dist. Entropy	1.313	0.224	1.307	5.488	0.696	5.489	35.392	2.915	35.343	144.709	8.150	143.659
Fréchet	Summing	1.312	0.224	1.305	5.486	0.696	5.488	35.390	2.915	35.341	144.705	8.150	143.656
	Averaging	1.312	0.224	1.305	5.486	0.696	5.488	35.390	2.915	35.341	144.705	8.150	143.656
	Avg. Maxima	1.321	0.224	1.314	5.505	0.696	5.507	35.437	2.915	35.388	144.802	8.149	143.752
	Weitzman	1.394	0.224	1.394	221.026	14.048	219.414	—	—	—	—	—	—
Partial	Dist. Entropy	1.052	0.030	1.051	4.425	0.101	4.435	28.495	0.446	28.510	115.301	1.512	115.542
Curve	Summing	1.050	0.030	1.049	4.422	0.101	4.433	28.492	0.446	28.508	115.296	1.512	115.538
Mapping	Averaging	1.050	0.030	1.049	4.422	0.101	4.433	28.492	0.446	28.508	115.296	1.512	115.538
	Avg. Maxima	1.060	0.030	1.059	4.441	0.101	4.451	28.540	0.446	28.555	115.394	1.512	115.633
	Weitzman	1.136	0.033	1.134	199.575	8.388	197.663	—	—	—	—	—	—
Dynamic	Dist. Entropy	1.209	0.207	1.206	5.056	0.643	5.043	32.604	2.689	32.556	133.317	7.537	132.292
Time	Summing	1.208	0.207	1.204	5.054	0.643	5.041	32.602	2.689	32.554	133.313	7.537	132.288
Warping	Averaging	1.208	0.207	1.204	5.054	0.643	5.041	32.602	2.689	32.554	133.313	7.537	132.288
	Avg. Maxima	1.217	0.207	1.213	5.073	0.643	5.060	32.649	2.689	32.602	133.410	7.537	132.385
	Weitzman	1.290	0.207	1.285	220.619	14.472	218.844	—	—	—	—	—	—
norm.	Dist. Entropy	0.071	0.004	0.071	0.296	0.013	0.295	1.897	0.060	1.902	7.695	0.206	7.702
Relative	Summing	0.070	0.004	0.070	0.294	0.013	0.293	1.895	0.060	1.900	7.691	0.206	7.698
Angle	Averaging	0.070	0.004	0.070	0.294	0.013	0.293	1.895	0.060	1.900	7.691	0.206	7.698
Dist.	Avg. Maxima	0.079	0.004	0.079	0.313	0.013	0.311	1.942	0.060	1.947	7.788	0.206	7.794
	Weitzman	0.152	0.007	0.152	215.614	15.920	212.012	—	—	—	—	—	—
Area	Dist. Entropy	3.532	0.527	3.417	14.686	1.633	14.656	94.596	6.480	95.274	383.706	18.185	385.097
Between	Summing	3.530	0.527	3.415	14.685	1.633	14.654	94.594	6.480	95.272	383.702	18.185	385.093
Curves	Averaging	3.530	0.527	3.415	14.685	1.633	14.654	94.594	6.480	95.272	383.702	18.185	385.093
	Avg. Maxima	3.539	0.527	3.425	14.703	1.633	14.673	94.641	6.480	95.319	383.799	18.185	385.188
	Weitzman	3.613	0.527	3.501	228.601	15.441	225.533	—	—	—	—	—	—
Complexity	Dist. Entropy	3.384	0.578	3.378	14.164	1.810	14.117	91.361	7.622	91.667	373.511	21.987	370.861
Vectors	Summing	3.383	0.578	3.377	14.162	1.810	14.115	91.359	7.622	91.665	373.507	21.987	370.858

Table 6 (continued)

		TS10			TS20			TS50			TS100		
		Mean	Std	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median
	Averaging	3.383	0.578	3.377	14.162	1.810	14.115	91.359	7.622	91.665	373.507	21.987	370.858
	Avg. Maxima	3.392	0.578	3.386	14.181	1.810	14.134	91.406	7.622	91.712	373.604	21.987	370.954
	Weitzman	3.465	0.578	3.460	211.435	8.249	209.964	—	—	—	—	—	—
Iterative	Dist. Entropy	13.237	2.283	13.182	55.422	6.992	55.201	357.503	29.518	356.528	1461.863	82.849	1456.148
Levenshtein	Summing	13.235	2.283	13.180	55.421	6.992	55.199	357.501	29.518	356.526	1461.860	82.849	1456.144
	Averaging	13.235	2.283	13.180	55.421	6.992	55.199	357.501	29.518	356.526	1461.860	82.849	1456.144
	Avg. Maxima	13.244	2.283	13.190	55.439	6.992	55.217	357.548	29.518	356.575	1461.956	82.849	1456.241
	Weitzman	13.318	2.283	13.264	270.098	18.072	268.315	—	—	—	—	—	—
Manhattan	Dist. Entropy	0.736	0.059	0.739	3.083	0.203	3.088	19.857	0.894	19.886	80.632	2.842	80.587
Distance	Summing	0.735	0.059	0.737	3.082	0.203	3.086	19.855	0.894	19.884	80.629	2.842	80.583
	Averaging	0.735	0.059	0.737	3.081	0.203	3.086	19.855	0.894	19.884	80.629	2.842	80.583
	Avg. Maxima	0.744	0.059	0.746	3.100	0.203	3.105	19.902	0.894	19.933	80.725	2.842	80.679
	Weitzman	0.829	0.106	0.822	216.185	14.076	212.040	—	—	—	—	—	—
Jaccard	Dist. Entropy	0.524	0.056	0.526	2.197	0.195	2.189	14.081	1.035	14.195	57.210	4.181	58.177
Similarity	Summing	0.522	0.056	0.524	2.195	0.196	2.186	14.078	1.036	14.192	57.206	4.181	58.173
(coarse)	Averaging	0.522	0.056	0.524	2.195	0.196	2.186	14.078	1.036	14.192	57.206	4.181	58.173
	Avg. Maxima	0.532	0.056	0.534	2.213	0.196	2.205	14.125	1.036	14.240	57.302	4.181	58.271
	Weitzman	0.609	0.060	0.612	205.964	9.919	204.245	—	—	—	—	—	—
Jaccard	Dist. Entropy	0.493	0.045	0.491	2.066	0.151	2.080	13.280	0.666	13.298	53.934	2.275	53.951
Similarity	Summing	0.492	0.045	0.489	2.064	0.151	2.078	13.278	0.666	13.296	53.931	2.275	53.948
(fine)	Averaging	0.492	0.045	0.489	2.064	0.151	2.078	13.278	0.666	13.296	53.931	2.275	53.948
	Avg. Maxima	0.501	0.045	0.498	2.082	0.151	2.097	13.325	0.666	13.342	54.029	2.274	54.041
	Weitzman	0.576	0.046	0.573	209.021	9.708	207.327	—	—	—	—	—	—
Direct	Feature Map	0.074	0.007	0.074	0.147	0.009	0.147	0.365	0.017	0.364	0.736	0.027	0.735
DMs	Test Set Diameter CS	0.002	0.000	0.002	0.005	0.005	0.004	0.010	0.001	0.010	0.019	0.002	0.019
	Convex Hull	0.003	0.000	0.002	0.004	0.001	0.004	0.007	0.001	0.007	0.014	0.001	0.013

(TS_{10}, TS_{20}). For TS_{20}, the computation time is in the order of 3.5 to 4.5 minutes. This is several orders of magnitude larger than the measurements of TS_{10} (see also the table of net aggregation times in Table 10, Appendix C). We also point out at the missing data for Weitzman aggregation for TS_{50} and TS_{100}, as their computation times exceeded the global runtime timeout of 24 hours.

Despite the cost of Weitzman, we notice that also other measures experience significant divergence, ranging from below 0.1 seconds for TS_{10} (norm. Relative Angle Dist.) to below 1 second (Manhattan Dist., $\mathrm{Jaccard}_{\mathrm{coarse}}$, $\mathrm{Jaccard}_{\mathrm{fine}}$), to more than 13 seconds (Iterative Levenshtein). When looking at larger TS sizes, this cost divergence accumulates to a spread of around 8 seconds for norm. Relative Angle Dist., to almost 25 minutes per TS for Iterative Levenshtein for TS_{100}. Nonetheless, we point out that this divergence is mostly due to the pairwise distance measure, rather than the aggregation itself. This is confirmed by measuring exclusively the execution time for aggregation (see Table 10 in Appendix C), which shows that except for Weitzman aggregation, all aggregators execute below 0.1 seconds, independent of TS size.

Direct DMs Considering the Direct DMs, we notice that all of them execute at sub-second runtimes, with Feature Map being the slowest with our implementation. In total, we notice a linear slowdown with TS size for Feature Map, while the others constantly below 0.02 seconds for all TS sizes.

Answer to *RQ1c*: We conclude that for pairwise DMs, the choice of aggregator does not make a significant difference, with the vital exception of Weitzman, which is prohibitively slow for test suites larger than 20 roads. For all other aggregators, the choice of pairwise distance measure has a more significant impact, such that TSs analysed with Iterative Levenshtein require almost 25 minutes of computation time, while norm. Relative Angle Dist. requires less than 8 seconds for TS_{100}. Except for Iterative Levenshtein, all other DMs require less than 7 minutes per TS_{100}.

As for Direct DMs, we notice that their computation time is almost negligible, as they all execute for all TS sizes below 1 second (Feature Map), most even below 0.02 seconds.

4.1.4 *RQ1d*: Additivity

To measure the additivity behaviour of our DMs under study, we proceeded as follows:

1) First we calculate each TSs' DMs.
2) We add *n* new roads to each TS, where *n* corresponds to $n \in \{1 \text{ road}, 2 \text{ roads}, 10\%, 20\%\}$
3) We re-calculate the DMs and record the time.
4) We analyse the results.

Table 7 shows the mean and standard deviation of the relative computation time increase. In the table, blue cells indicate a linear or lower increase of computation time, while orange cells indicate a larger than linear increase. As before, we may immediately observe that adding a single road to the Weitzman-DMs more than doubles the computation time. Adding two roads increases it almost fivefold, and adding four roads ($\mathrm{TS}_{20+20\%}$) increases the

runtime by at least 21 times the base runtime. This underlines the fact that Weitzman aggregation is prohibitively expensive for TSs larger than 20.

As for the other DMs, we notice that most measures on average expose almost linear cost growth behaviour. Clearly, some fluctuation exists for most measures (e.g. some DMs that use Avg. Maxima aggregation), and for large TSs of size 100, there appears to be an additional cost.

For Direct DMs, it seems that Feature Map shows worse additivity behaviour than the other measures, although the difference appears to be insignificant.

Answer to *RQ1d*: In terms of additivity, most DMs show (almost) linear cost increase when adding roads. The exception to this is Weitzman aggregation, which shows exponential increase in runtimes, rendering it unusable for large TS sizes, or use cases where individual roads have to be added to existing TSs.

4.2 *RQ2* – Pairwise Correlation

Figure 6 displays the complete pairwise Spearman correlation plot of all DMs.[8] We can see immediately by the predominant shades of blue that many of the DMs are positively correlated. One exception is the $\text{Jaccard}_{\text{fine}}$ Avg. Maxima DM's empty values. This is because

Table 7 RQ1d – Additivity — Relative change (mean ± standard deviation)

		TS10	TS20				TS50					TS100				
		TS10+10%	TS10+20%	TS20+1	TS20+10%	TS20+20%	TS50+1	TS50+2	TS50+10%	TS50+20%	TS100+1	TS100+2	TS100+10%	TS100+20%	
Discrete Fréchet	Dist. Entropy	1.0 ± 0.05	1.01 ± 0.06	1.02 ± 0.06	1.03 ± 0.06	1.04 ± 0.06	1.01 ± 0.04	1.02 ± 0.04	1.06 ± 0.06	1.12 ± 0.1	1.01 ± 0.04	1.02 ± 0.03	1.12 ± 0.04	1.26 ± 0.09	
	Summing	1.01 ± 0.06	1.01 ± 0.06	1.01 ± 0.09	1.0 ± 0.06	1.0 ± 0.07	1.01 ± 0.07	1.01 ± 0.05	1.02 ± 0.1	1.04 ± 0.11	1.01 ± 0.14	1.0 ± 0.11	1.0 ± 0.1	1.01 ± 0.11	
	Averaging	1.0 ± 0.11	0.99 ± 0.09	1.01 ± 0.09	1.01 ± 0.08	1.01 ± 0.07	0.99 ± 0.07	1.01 ± 0.11	1.05 ± 0.19	1.03 ± 0.1	1.02 ± 0.14	1.01 ± 0.12	1.03 ± 0.13	1.05 ± 0.11	
	Avg. Maxima	1.1 ± 0.03	1.19 ± 0.03	1.04 ± 0.03	1.1 ± 0.04	1.2 ± 0.04	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.02	1.21 ± 0.03	1.01 ± 0.02	1.02 ± 0.03	1.11 ± 0.03	1.22 ± 0.04	
	Weitzman	2.16 ± 0.12	5.08 ± 2.9	2.19 ± 0.15	4.86 ± 0.34	23.99 ± 1.69	—	—	—	—	—	—	—	—	
Partial Curve Mapping	Dist. Entropy	1.0 ± 0.11	1.04 ± 0.13	1.02 ± 0.12	1.03 ± 0.12	1.08 ± 0.17	1.0 ± 0.09	1.03 ± 0.12	1.07 ± 0.17	1.13 ± 0.19	1.01 ± 0.07	1.02 ± 0.09	1.08 ± 0.1	1.2 ± 0.14	
	Summing	1.02 ± 0.11	1.03 ± 0.11	1.02 ± 0.1	1.03 ± 0.12	1.07 ± 0.14	1.0 ± 0.09	1.0 ± 0.11	1.02 ± 0.14	1.05 ± 0.24	1.0 ± 0.1	1.0 ± 0.12	0.96 ± 0.13	0.98 ± 0.14	
	Averaging	0.99 ± 0.16	1.0 ± 0.15	1.01 ± 0.16	1.02 ± 0.14	1.03 ± 0.17	1.04 ± 0.22	1.03 ± 0.19	1.05 ± 0.29	1.05 ± 0.28	1.06 ± 0.27	1.05 ± 0.26	1.06 ± 0.28	1.13 ± 0.3	
	Avg. Maxima	1.1 ± 0.07	1.19 ± 0.04	1.05 ± 0.03	1.1 ± 0.03	1.45 ± 2.34	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.03	1.2 ± 0.03	1.01 ± 0.02	1.02 ± 0.02	1.11 ± 0.03	1.23 ± 0.03	
	Weitzman	2.12 ± 0.14	4.7 ± 0.86	2.2 ± 0.12	4.92 ± 0.3	21.24 ± 1.27	—	—	—	—	—	—	—	—	
Dynamic Time Warping	Dist. Entropy	1.01 ± 0.05	1.02 ± 0.06	1.01 ± 0.04	1.02 ± 0.05	1.03 ± 0.04	1.0 ± 0.04	1.02 ± 0.05	1.06 ± 0.13	1.1 ± 0.05	1.01 ± 0.03	1.02 ± 0.03	1.12 ± 0.03	1.26 ± 0.05	
	Summing	1.01 ± 0.09	1.01 ± 0.11	1.01 ± 0.11	1.01 ± 0.13	1.02 ± 0.13	0.99 ± 0.05	1.01 ± 0.1	1.01 ± 0.1	1.03 ± 0.12	1.01 ± 0.09	1.02 ± 0.13	1.03 ± 0.15	1.03 ± 0.13	
	Averaging	1.01 ± 0.07	1.02 ± 0.11	0.99 ± 0.07	1.01 ± 0.13	1.0 ± 0.09	1.04 ± 0.21	1.0 ± 0.06	1.04 ± 0.24	1.03 ± 0.06	1.0 ± 0.13	1.01 ± 0.15	1.03 ± 0.16	1.04 ± 0.13	
	Avg. Maxima	1.1 ± 0.03	1.19 ± 0.03	1.05 ± 0.03	1.1 ± 0.03	1.2 ± 0.03	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.02	1.21 ± 0.03	1.02 ± 0.02	1.02 ± 0.02	1.11 ± 0.03	1.22 ± 0.03	
	Weitzman	2.15 ± 0.1	4.98 ± 1.43	2.2 ± 0.16	4.8 ± 0.34	23.68 ± 1.8	—	—	—	—	—	—	—	—	
norm. Relative Angle Dist.	Dist. Entropy	1.01 ± 0.04	1.01 ± 0.04	1.01 ± 0.05	1.02 ± 0.05	1.03 ± 0.05	1.0 ± 0.06	1.02 ± 0.05	1.05 ± 0.06	1.11 ± 0.06	1.01 ± 0.05	1.02 ± 0.05	1.12 ± 0.06	1.25 ± 0.06	
	Summing	0.99 ± 0.09	0.99 ± 0.08	1.0 ± 0.08	1.0 ± 0.08	1.01 ± 0.12	1.02 ± 0.13	1.02 ± 0.1	1.05 ± 0.16	1.04 ± 0.12	0.98 ± 0.07	1.01 ± 0.12	1.0 ± 0.09	1.02 ± 0.13	
	Averaging	1.02 ± 0.14	1.02 ± 0.16	1.01 ± 0.11	1.0 ± 0.09	1.0 ± 0.09	1.0 ± 0.09	1.02 ± 0.17	1.03 ± 0.13	1.02 ± 0.09	1.0 ± 0.09	1.02 ± 0.11	1.02 ± 0.12	1.06 ± 0.13	
	Avg. Maxima	1.1 ± 0.03	1.2 ± 0.03	1.05 ± 0.04	1.1 ± 0.03	1.2 ± 0.03	1.02 ± 0.02	1.04 ± 0.02	1.11 ± 0.02	1.21 ± 0.02	1.01 ± 0.02	1.02 ± 0.02	1.11 ± 0.03	1.22 ± 0.03	
	Weitzman	2.24 ± 0.8	4.89 ± 1.12	2.18 ± 0.15	4.8 ± 0.37	23.5 ± 1.83	—	—	—	—	—	—	—	—	
Area Between Curves	Dist. Entropy	1.0 ± 0.05	1.01 ± 0.06	1.02 ± 0.05	1.02 ± 0.05	1.03 ± 0.04	1.01 ± 0.06	1.02 ± 0.05	1.05 ± 0.05	1.11 ± 0.08	1.02 ± 0.04	1.03 ± 0.03	1.13 ± 0.03	1.26 ± 0.03	
	Summing	1.01 ± 0.1	1.0 ± 0.06	1.01 ± 0.09	1.02 ± 0.1	1.0 ± 0.07	1.0 ± 0.07	1.01 ± 0.11	1.01 ± 0.12	1.02 ± 0.09	1.03 ± 0.15	1.0 ± 0.06	1.02 ± 0.12	1.04 ± 0.15	
	Averaging	1.01 ± 0.12	1.01 ± 0.15	1.01 ± 0.06	1.03 ± 0.09	1.02 ± 0.17	1.03 ± 0.28	1.0 ± 0.13	1.01 ± 0.09	1.01 ± 0.08	1.03 ± 0.17	1.02 ± 0.11	1.06 ± 0.17	1.07 ± 0.1	
	Avg. Maxima	1.1 ± 0.03	1.2 ± 0.03	1.05 ± 0.02	1.1 ± 0.03	1.2 ± 0.03	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.02	1.2 ± 0.02	1.01 ± 0.02	1.02 ± 0.02	1.11 ± 0.02	1.22 ± 0.03	
	Weitzman	2.49 ± 3.53	4.91 ± 1.45	2.2 ± 0.18	4.85 ± 0.36	23.59 ± 1.76	—	—	—	—	—	—	—	—	
Complexity Vectors	Dist. Entropy	1.01 ± 0.11	1.0 ± 0.08	1.01 ± 0.06	1.02 ± 0.06	1.03 ± 0.07	1.01 ± 0.06	1.02 ± 0.07	1.05 ± 0.1	1.1 ± 0.08	1.02 ± 0.04	1.03 ± 0.04	1.12 ± 0.05	1.26 ± 0.08	
	Summing	1.03 ± 0.23	1.0 ± 0.08	1.01 ± 0.14	0.99 ± 0.09	1.0 ± 0.09	1.0 ± 0.06	1.02 ± 0.07	1.02 ± 0.14	1.01 ± 0.08	1.01 ± 0.08	1.08 ± 0.62	1.02 ± 0.07	1.04 ± 0.1	
	Averaging	1.0 ± 0.07	1.01 ± 0.09	1.01 ± 0.09	1.01 ± 0.09	1.01 ± 0.07	1.0 ± 0.05	1.0 ± 0.06	1.02 ± 0.12	1.02 ± 0.1	1.01 ± 0.05	1.02 ± 0.07	1.04 ± 0.07	1.07 ± 0.08	
	Avg. Maxima	1.1 ± 0.04	1.2 ± 0.05	1.05 ± 0.03	1.1 ± 0.03	1.21 ± 0.03	1.02 ± 0.03	1.04 ± 0.03	1.1 ± 0.03	1.2 ± 0.03	1.01 ± 0.02	1.02 ± 0.02	1.1 ± 0.02	1.22 ± 0.03	
	Weitzman	2.22 ± 0.48	4.96 ± 1.58	2.16 ± 0.1	4.74 ± 0.23	22.86 ± 1.06	—	—	—	—	—	—	—	—	
Iterative Levenshtein	Dist. Entropy	1.0 ± 0.05	1.01 ± 0.06	1.01 ± 0.04	1.02 ± 0.05	1.03 ± 0.05	1.01 ± 0.05	1.02 ± 0.06	1.05 ± 0.06	1.1 ± 0.05	1.02 ± 0.07	1.02 ± 0.03	1.11 ± 0.03	1.25 ± 0.04	
	Summing	1.01 ± 0.1	1.03 ± 0.12	1.0 ± 0.09	1.0 ± 0.05	1.01 ± 0.07	1.03 ± 0.13	1.01 ± 0.09	1.02 ± 0.08	1.02 ± 0.07	1.02 ± 0.15	1.0 ± 0.08	1.02 ± 0.1	1.04 ± 0.13	
	Averaging	0.99 ± 0.08	0.99 ± 0.08	1.02 ± 0.19	1.02 ± 0.1	1.01 ± 0.13	1.0 ± 0.09	1.01 ± 0.1	1.02 ± 0.09	1.04 ± 0.11	1.01 ± 0.09	1.01 ± 0.14	1.03 ± 0.11	1.06 ± 0.13	
	Avg. Maxima	1.1 ± 0.03	1.2 ± 0.05	1.05 ± 0.02	1.1 ± 0.03	1.2 ± 0.03	1.02 ± 0.02	1.04 ± 0.02	1.11 ± 0.02	1.2 ± 0.02	1.02 ± 0.02	1.03 ± 0.02	1.11 ± 0.02	1.22 ± 0.03	
	Weitzman	2.14 ± 0.1	5.94 ± 6.24	2.19 ± 0.16	4.83 ± 0.36	23.48 ± 2.05	—	—	—	—	—	—	—	—	
Manhattan Distance	Dist. Entropy	1.01 ± 0.04	1.03 ± 0.12	1.01 ± 0.05	1.01 ± 0.05	1.03 ± 0.06	1.01 ± 0.13	1.01 ± 0.04	1.04 ± 0.04	1.09 ± 0.07	1.01 ± 0.04	1.02 ± 0.04	1.1 ± 0.04	1.21 ± 0.04	
	Summing	1.02 ± 0.11	1.01 ± 0.07	1.0 ± 0.12	1.01 ± 0.12	1.02 ± 0.19	0.99 ± 0.1	1.02 ± 0.18	1.01 ± 0.12	1.02 ± 0.11	1.03 ± 0.17	1.02 ± 0.1	1.03 ± 0.15	1.04 ± 0.19	
	Averaging	1.0 ± 0.07	1.01 ± 0.13	1.02 ± 0.11	1.02 ± 0.15	1.01 ± 0.1	1.01 ± 0.11	0.99 ± 0.1	1.02 ± 0.13	1.01 ± 0.11	1.02 ± 0.14	1.01 ± 0.11	1.04 ± 0.18	1.07 ± 0.19	
	Avg. Maxima	1.1 ± 0.04	1.19 ± 0.04	1.04 ± 0.03	1.09 ± 0.03	1.2 ± 0.04	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.02	1.2 ± 0.02	1.01 ± 0.02	1.03 ± 0.02	1.11 ± 0.03	1.22 ± 0.03	
	Weitzman	2.26 ± 1.64	4.96 ± 3.0	2.2 ± 0.16	4.88 ± 0.3	23.38 ± 1.58	—	—	—	—	—	—	—	—	
Jaccard Similarity (coarse)	Dist. Entropy	1.04 ± 0.14	1.06 ± 0.16	1.03 ± 0.13	1.06 ± 0.21	1.05 ± 0.18	1.02 ± 0.09	1.06 ± 0.14	1.08 ± 0.19	1.12 ± 0.17	1.0 ± 0.1	1.0 ± 0.13	1.06 ± 0.17	1.14 ± 0.2	
	Summing	1.02 ± 0.09	1.03 ± 0.12	1.01 ± 0.11	1.02 ± 0.12	1.03 ± 0.14	1.0 ± 0.09	1.02 ± 0.13	1.01 ± 0.13	1.04 ± 0.16	0.99 ± 0.11	0.99 ± 0.13	0.97 ± 0.17	0.96 ± 0.19	
	Averaging	1.03 ± 0.14	1.05 ± 0.18	1.07 ± 0.49	1.07 ± 0.34	1.05 ± 0.25	1.04 ± 0.23	1.11 ± 0.32	1.07 ± 0.31	1.1 ± 0.29	1.01 ± 0.3	0.97 ± 0.19	1.0 ± 0.29	1.01 ± 0.29	
	Avg. Maxima	1.1 ± 0.02	1.2 ± 0.03	1.06 ± 0.03	1.1 ± 0.03	1.39 ± 1.61	1.02 ± 0.02	1.04 ± 0.02	1.11 ± 0.03	1.21 ± 0.03	1.01 ± 0.02	1.02 ± 0.02	1.11 ± 0.03	1.23 ± 0.03	
	Weitzman	2.15 ± 0.13	4.76 ± 0.35	2.16 ± 0.08	4.63 ± 0.16	21.94 ± 1.22	—	—	—	—	—	—	—	—	
Jaccard Similarity (fine)	Dist. Entropy	1.01 ± 0.07	1.02 ± 0.07	1.0 ± 0.1	1.02 ± 0.13	1.02 ± 0.12	1.02 ± 0.1	1.05 ± 0.14	1.05 ± 0.11	1.09 ± 0.08	1.0 ± 0.11	1.0 ± 0.08	1.08 ± 0.11	1.17 ± 0.13	
	Summing	1.0 ± 0.07	1.01 ± 0.08	1.0 ± 0.07	1.01 ± 0.07	1.01 ± 0.1	1.01 ± 0.09	1.06 ± 0.18	1.02 ± 0.13	1.03 ± 0.08	0.99 ± 0.09	0.99 ± 0.1	0.99 ± 0.1	0.99 ± 0.11	
	Averaging	1.01 ± 0.11	1.05 ± 0.44	1.01 ± 0.12	1.0 ± 0.06	1.0 ± 0.06	1.04 ± 0.43	1.04 ± 0.27	1.01 ± 0.13	1.02 ± 0.12	1.01 ± 0.09	1.0 ± 0.1	1.04 ± 0.13	1.04 ± 0.08	
	Avg. Maxima	1.1 ± 0.03	1.2 ± 0.04	1.05 ± 0.04	1.1 ± 0.04	1.21 ± 0.04	1.02 ± 0.03	1.04 ± 0.03	1.11 ± 0.03	1.21 ± 0.03	1.0 ± 0.08	1.01 ± 0.08	1.1 ± 0.09	1.22 ± 0.1	
	Weitzman	2.19 ± 0.12	5.44 ± 2.68	2.15 ± 0.09	4.69 ± 0.17	22.51 ± 1.05	—	—	—	—	—	—	—	—	
Direct DMs	Feature Map	1.1 ± 0.04	1.2 ± 0.06	1.05 ± 0.03	1.1 ± 0.03	1.2 ± 0.05	1.02 ± 0.02	1.04 ± 0.02	1.1 ± 0.03	1.21 ± 0.03	1.01 ± 0.01	1.02 ± 0.02	1.1 ± 0.03	1.2 ± 0.03	
	Test Set Diameter CS	1.09 ± 0.06	1.19 ± 0.11	1.04 ± 0.11	1.1 ± 0.16	1.2 ± 0.1	1.02 ± 0.03	1.03 ± 0.04	1.09 ± 0.05	1.2 ± 0.06	1.01 ± 0.03	1.02 ± 0.04	1.09 ± 0.04	1.2 ± 0.07	
	Convex Hull	1.05 ± 0.09	1.12 ± 0.13	1.03 ± 0.08	1.07 ± 0.1	1.12 ± 0.1	1.03 ± 0.09	1.05 ± 0.08	1.1 ± 0.11	1.19 ± 0.12	1.0 ± 0.08	1.01 ± 0.08	1.09 ± 0.1	1.17 ± 0.09	

[8] For reference, we also provide the Pearson correlation plot in Appendix D

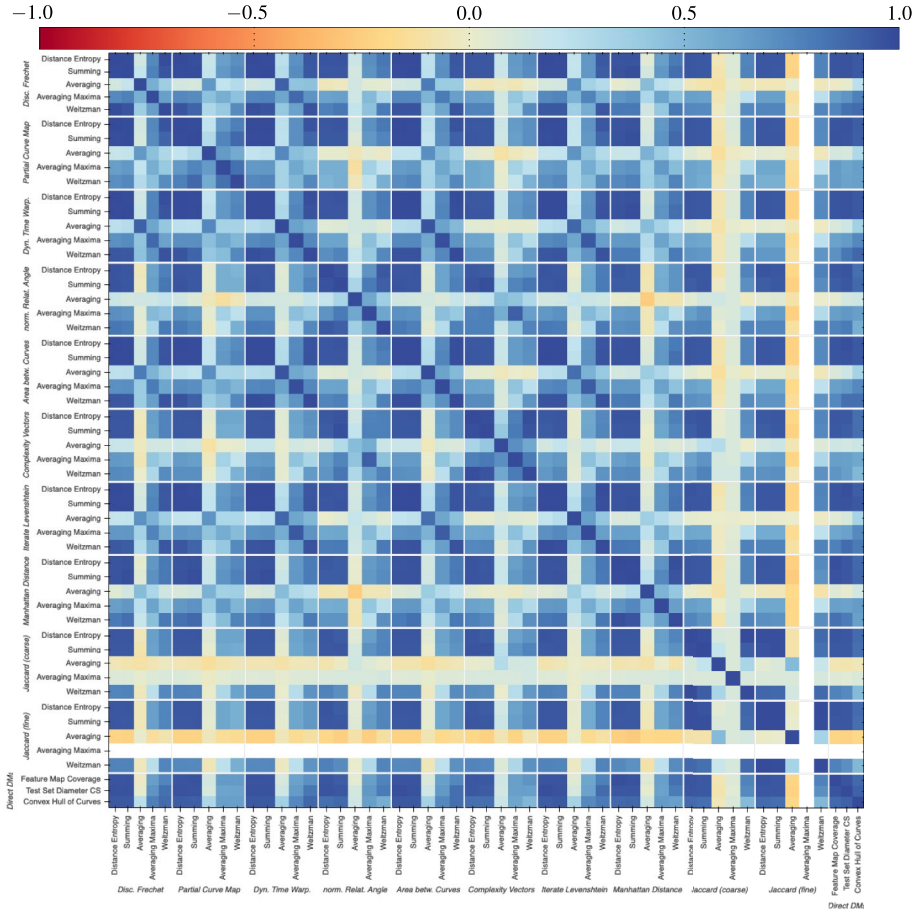


Fig. 6 RQ2 – Correlation heatmap. Each cell indicates the Spearman correlation between the DMs on the row and on the column. Shades of blue indicate positive correlations, while shades of red indicate negative correlation. A larger and interactive version of this figure is available online at http://stklik.github.io/2026-EMSE-Road-Diversity/

all entries here have the value of 1.0, thus no correlation could be computed. We further notice that the Averaging aggregation across all distance measures leads to low correlation.

Interestingly, we may observe that the aggregators appear to have a stronger influence on the correlation than the distance measures. Thus, for instance, for a distance measure such as Partial Curve Mapping, we notice that its Summing and Dist. Entropy aggregators are strongly correlated with Discrete Fréchet's Summing and Dist. Entropy. On the other hand, their correlations with Partial Curve Mapping's Weitzman and Averaging are comparably weaker.

We further note that, even besides Averaging, there are numerous pairs of DMs that expose weaker correlation. This suggests that these DMs capture different aspects of road geometry diversity and may be complementary when used together. For example, DMs based on norm. Relative Angle Dist. show moderate correlation with most other distance measures, indicating that this angle-based approach provides a distinct perspective on diversity. Similarly, the Jaccard-based measures ($\text{Jaccard}_{\text{coarse}}$, $\text{Jaccard}_{\text{fine}}$) show moderate

to weaker correlation with geometric measures like Discrete Fréchet and Area Between Curves, suggesting that segment-based diversity assessment captures different characteristics than curve-based geometric similarity.

The observed correlation patterns have practical implications for selecting DMs in ADS testing. When multiple DMs are strongly correlated (correlation coefficient > 0.8), practitioners may choose to use only one of them, preferably selecting based on the properties identified in ***RQ1*** (i.e. efficiency, monotonicity, insensitivity to duplicates). Conversely, weakly correlated DMs may be used in combination to capture multiple facets of road diversity, potentially leading to more comprehensive test suites that exercise a wider range of driving behaviours.

Answer to *RQ2*: The correlation analysis reveals that aggregation methods have a stronger influence on correlation patterns than the underlying distance functions. Many DMs show strong positive correlation, particularly those using the same aggregation method (Dist. Entropy, Summing, Weitzman) across different distance measures. A notable exception is Averaging aggregation, which shows weak or negative correlations with most other DMs. The presence of both strongly correlated and weakly correlated DM pairs indicates opportunities for both optimisation (by avoiding redundant computations of highly correlated measures) and enhanced diversity assessment (by combining weakly correlated measures to capture different aspects of road geometry).

An interactive version of the correlation heatmap is available online at https://stklik.github.io/2026-EMSE-Road-Diversity/RQ2/heatmap-spearman-full.html, allowing for detailed exploration of individual correlation values and patterns.

4.3 *RQ3* – Correlation with Road Length

Our approach to measure the correlation of DMs with road length follows this procedure:

1) We sample random TSs based on the length of the roads. As we are specifically interested in short (resp. long) roads, we have to adjust our TS sampling algorithm to assemble random TSs from the shortest (resp. longest) quantile of the roads.
2) We then calculate DMs for all TSs and perform a correlation analysis similar to ***RQ2***, i.e.

   (a) we calculate DMs for test suites and
   (b) perform correlation analysis between the DM values and the (TSs' average) road length.

The results of this analysis allow determining whether there is an effect of the road length on the diversity. Combined with the results of ***RQ4.b***, it helps users in selecting appropriate road lengths for ADS testing.

Figure 7 presents the Spearman (left) and Pearson (right) correlation coefficients between each DM and road length across all TS sizes. The results reveal several important patterns. As both Spearman and Pearson measures indicate similar behaviour, we focus our description on Spearman's correlation coefficients (cf. Fig. 7a, but point out that Pearson's correlation analysis describes similar behaviour.

First, we observe a striking pattern across aggregation methods. DMs using Averaging aggregation show remarkably strong positive correlations with road length, with correlation coefficients consistently above 0.90 across nearly all distance measures (Discrete Fréchet: 0.94, Partial Curve Mapping: 0.94, Dynamic Time Warping: 0.94, Area Between Curves: 0.94, Iterative Levenshtein: 0.94). Similarly, Avg. Maxima aggregation exhibits very high correlations (typically around 0.87–0.92). This suggests that averaging-based aggregations are highly sensitive to road length, likely because they normalise by the number of elements in ways that systematically scale with longer roads.

Second, in stark contrast, DMs using Summing aggregation show much weaker correlations with road length, typically ranging from -0.21 to 0.41. This indicates that summing-based aggregations are more robust to road length variations, making them preferable when length-independent diversity assessment is desired.

Third, Dist. Entropy aggregation shows moderate positive correlations (ranging from 0.19 to 0.76), with notable variation across different distance measures. The strongest correlation appears with Dynamic Time Warping (0.76) and Area Between Curves (0.75), while the weakest is with norm. Relative Angle Dist. (0.19) and Complexity Vectors (-0.17).

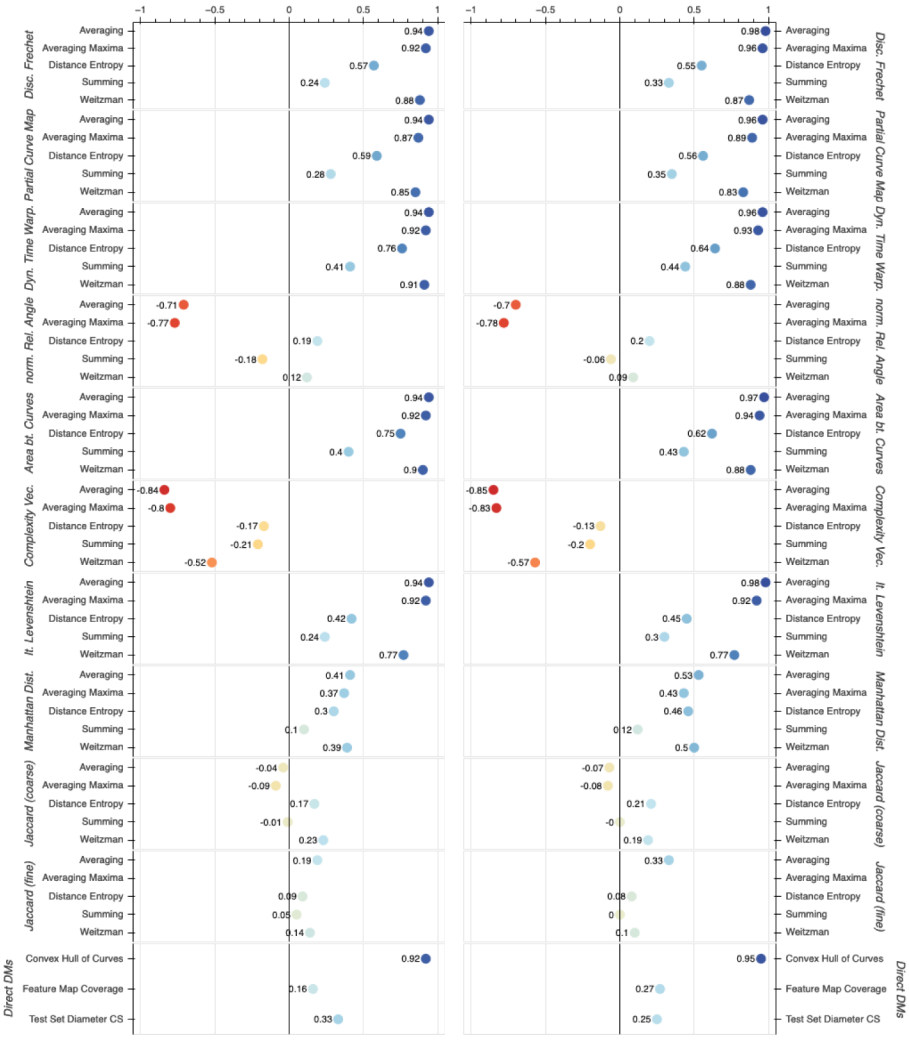
Fourth, certain distance measures show distinctive behaviour. Notably, norm. Relative Angle Dist. and Complexity Vectors exhibit negative correlations with road length when combined with most aggregators (particularly with Averaging: -0.71 and -0.84, respectively). This suggests these measures capture geometric properties that are diluted or inverted as roads become longer.

Among the Direct DMs, we observe highly divergent behaviour: Convex Hull displays very strong positive correlation (0.92) with road length, as expected since longer roads naturally span larger spatial areas. Feature Map shows weak positive correlation (0.16), while Test Set Diameter CS shows moderate positive correlation (0.33).

The Jaccard-based measures ($\mathrm{Jaccard}_{\mathrm{coarse}}$, $\mathrm{Jaccard}_{\mathrm{fine}}$) generally show weak correlations with road length across all aggregators, with values ranging from -0.09 to 0.23, suggesting they are relatively robust to length variations.

These findings have important practical implications for ADS testing practitioners. DMs using Averaging or Avg. Maxima aggregation should be used with caution when test suites contain roads of varying lengths, as their diversity scores may be heavily influenced by road length rather than geometric diversity alone. Conversely, Summing aggregation appears more robust and should be preferred when length-independent diversity assessment is required. For fair comparison across test suites with varying road lengths, practitioners should either normalise DM values by average road length, restrict test suites to roads of similar lengths, or select DMs known to be less sensitive to length effects.

Answer to *RQ3*: The correlation between DMs and road length varies dramatically depending on the aggregation method. Averaging and Avg. Maxima aggregations show very strong positive correlation with road length (typically > 0.9), making them highly sensitive to length variations and potentially problematic for fair diversity assessment. In contrast, Summing aggregation demonstrates weak correlation (ranging from -0.21 to 0.41), making it more robust and preferable for length-independent diversity measurement. Dist. Entropy aggregation shows moderate correlation with substantial variation across distance measures. Among Direct DMs, Convex Hull is highly correlated with road length (0.92), while Feature Map shows weak correlation (0.16 Spearman, 0.27 Pearson). Practitioners should carefully select DMs based on whether they need to isolate geometric diversity from road length effects.



(a) Spearman correlation (b) Pearson correlation

Fig. 7 RQ3 – Correlation between DMs and road length. Spearman correlation on left; Pearson on the right. Positive values indicate that the DM increases with road length, while negative values that it decreases

4.4 *RQ4* – Correlation DMs and BD

Here, we analyse the correlation between the individual DMs and the BD.

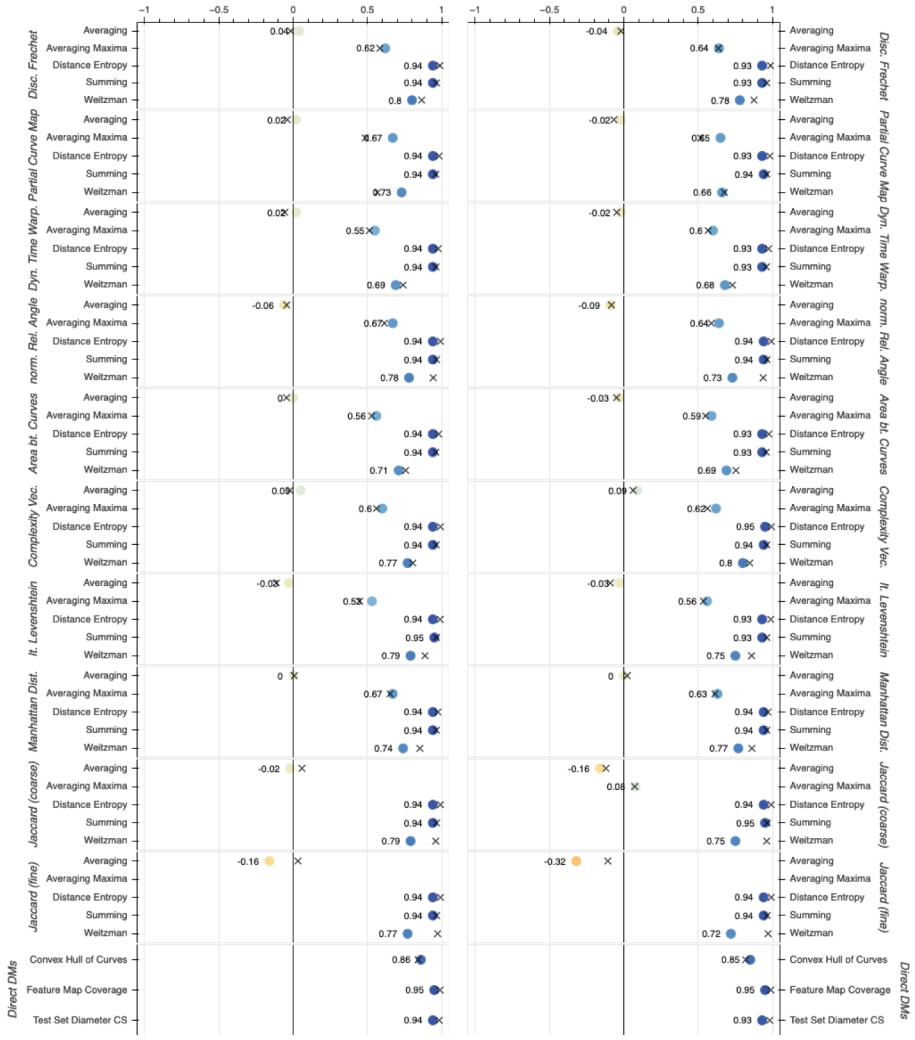
4.4.1 *RQ4a*: Correlation DMs and BD

First, we measure the general correlation between road diversity (i.e. DMs) and BD using Spearman's and Pearson's correlation. Specifically, we proceed as follows: (i) For each TS,

we calculate the DMs and (ii) then, for each TS we calculate the BD (as described above). (iii) Finally, we perform correlation analysis between these values.

The results reveal several important patterns regarding which DMs are effective indicators of behavioural diversity.

Figure 8a and b present both correlation coefficients between each DM and BD for both `BeamNG.AI` and `Dave2`. Given the remarkable similarity between Spearman and Pearson, we will again focus our description on Spearman's correlation measure. Overall, the results



(a) `BeamNG.AI` (b) `Dave2`

Fig. 8 RQ4a – Correlation DM and BD. Each row indicates the Spearman correlation between the DM on the row and BD. Shades of blue indicate positive correlations, while shades of red indicate negative correlation. The Pearson correlation values are indicated as crosses

reveal several important patterns regarding which DMs are effective indicators of behavioural diversity.

First, we observe remarkably strong positive correlations between road DMs and BD for both driving agents. DMs using Dist. Entropy and summing aggregation consistently show very high correlations with BD, with correlation coefficients ranging from 0.93 to 0.95 for both `BeamNG.AI` and `Dave2` across many distance measures. These findings provide strong empirical support for the hypothesis that road diversity is indeed a good predictor of behavioural diversity.

Second, consistent with our findings in ***RQ3***, DMs using Averaging aggregation show essentially no correlation with BD for both agents, with correlation coefficients near zero (ranging from -0.32 to 0.08). This suggests that these measures' sensitivity to road length (as identified in ***RQ3***) severely interferes with their ability to predict behavioural diversity. The Avg. Maxima aggregation shows moderate correlations (0.55–0.67) for both agents, indicating better performance than Averaging but substantially weaker than Dist. Entropy or Summing.

Third, among the distance measures, there is remarkable consistency across different distance functions when combined with Dist. Entropy or Summing aggregation. All distance measures combined with Dist. Entropy show correlations above 0.94, while with Summing they all show correlations of 0.93–0.95. This suggests that the choice of aggregation method is far more critical than the choice of distance function for predicting behavioural diversity. The only notable exception is Weitzman aggregation, which shows moderately strong correlations (0.68–0.89), with some variation across distance measures.

Fourth, the Direct DMs show strong positive correlations. Feature Map displays the strongest correlation among them, with values of 0.95 for `BeamNG.AI` and `Dave2`. Test Set Diameter CS also shows a very strong correlation (0.94 and 0.93), similar to Convex Hull whose positive correlation is slightly below (0.86 and 0.85).

The Jaccard-based measures ($\mathrm{Jaccard}_{\mathrm{coarse}}$, $\mathrm{Jaccard}_{\mathrm{fine}}$) show very strong correlations with BD when combined with Dist. Entropy and Summing (0.94–0.95) aggregation. We point out that, for $\mathrm{Jaccard}_{\mathrm{fine}}$ and Weitzman aggregation, Pearson shows a much stronger correlation than Spearman, which is different than for the other DM- BD correlations.

Interestingly, we observe almost identical correlation patterns between `BeamNG.AI` and `Dave2` , despite their fundamentally different architectures (rule-based vs. learning-based). The correlation coefficients differ marginally across almost all DMs (except $\mathrm{Jaccard}_{\mathrm{fine}}$ Averaging), suggesting that road diversity affects behavioural diversity similarly regardless of the driving agent's internal mechanisms.

These findings have profound implications for ADS testing practitioners. The very strong correlations observed (0.93–0.95 for the best-performing DMs) demonstrate that road diversity is indeed an excellent predictor of behavioural diversity. This validates the use of road DMs as proxies for BD in test generation and selection. The results clearly indicate that DMs using Dist. Entropy or Summing aggregation should be preferred, while Averaging aggregation should be avoided. The strong performance of Feature Map and Test Set Diameter CS suggests that direct DMs may offer the best balance of correlation with BD and computational efficiency.

Answer to *RQ4a*: There is very strong positive correlation between road DMs and BD for both driving agents. DMs using Dist. Entropy aggregation show the strongest correlations (0.93–0.95), followed closely by Summing aggregation that shows similar values, regardless of the underlying distance measure. Among Direct DMs, Feature Map shows the highest correlation with BD (0.95). DMs using Averaging aggregation show essentially no correlation with BD (near zero). Remarkably, correlation patterns are nearly identical between `BeamNG.AI` and `Dave2`, suggesting that road diversity affects behavioural diversity consistently across different types of driving agents. These findings strongly validate the use of road diversity as a proxy for behavioural diversity in ADS testing, particularly when using Dist. Entropy or Summing aggregation, or the Feature Map and Test Set Diameter CS Direct DMs.

4.4.2 *RQ4b*: Correlation of DMs and BD by Road Length

To study the effect of road length on the Spearman correlation between DMs and BD, we evaluate if for TSs with short roads, there is a stronger correlation between DMs and BD.[9] Hence, we proceed as follows: (i) Using the similar-in-length TSs created in ***RQ3***, we calculate for each TS the values of the DMs and the BD (as in ***RQ4.a***) (ii) We then perform a correlation analysis of each DM and BD pair. (iii) Finally, we compare the yielded correlation coefficients and check if the correlation of DMs and BD of TSs with short (resp. long) roads is higher (resp. lower) than the one of "normal" TSs.

Figure 9 and Table 8 present the Spearman correlation coefficients between each DM and BD for test suites grouped by road length (short, medium, and long roads). This analysis reveals how road length affects the relationship between road diversity and behavioural diversity.

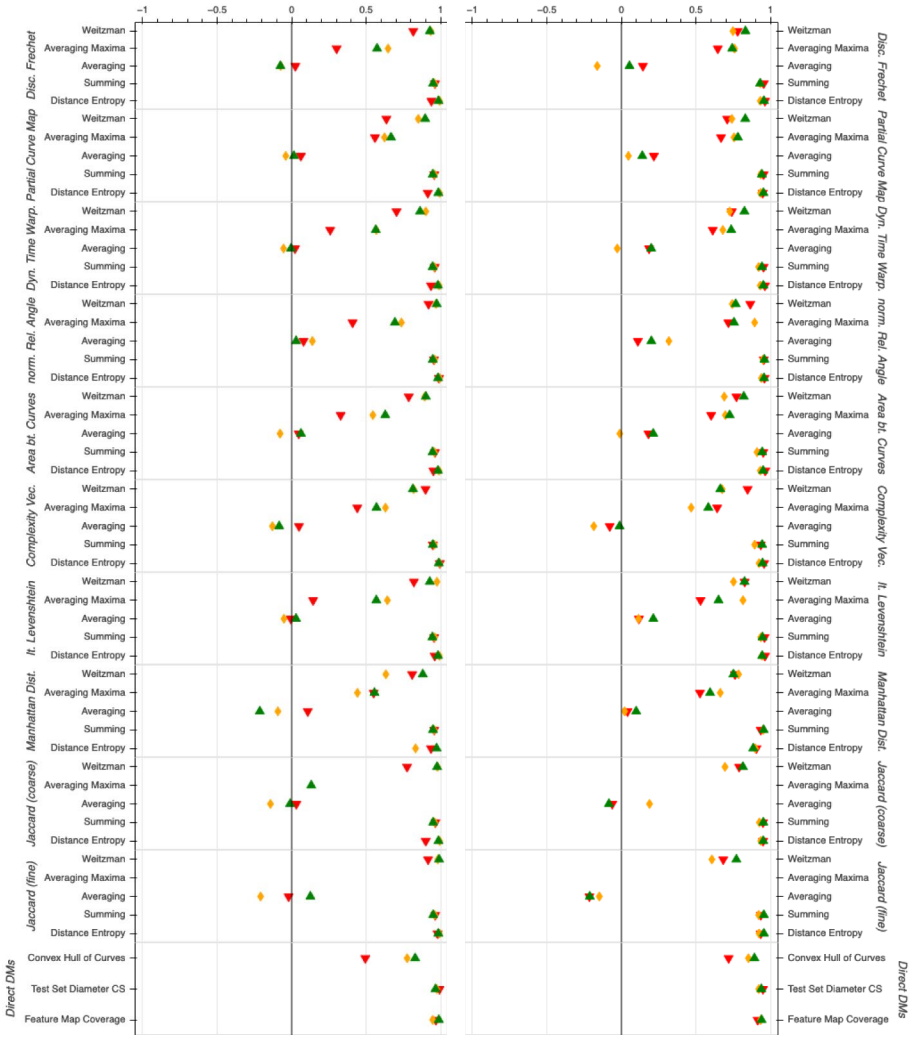
The most striking finding is that DMs using Dist. Entropy and Summing aggregation maintain very strong correlations with BD across all road length categories, though with some variation. For all road lengths, entropy- and summing-based DMs consistently achieve correlations of 0.85–0.97 across all distance measures for both `BeamNG.AI` and `Dave2`. This indicates that Dist. Entropy and Summing aggregation remain highly effective regardless of road length.

The Averaging aggregation continues to show near-zero correlation with BD across all road length categories (ranging from -0.34 to 0.32), confirming our earlier finding from ***RQ4a*** that this aggregation method is ineffective for predicting behavioural diversity regardless of road length. Similarly, Avg. Maxima shows moderate but inconsistent performance across road lengths, with correlations ranging from 0.40 to 0.89, and generally correlates better on medium and long roads than on short roads.

Among the distance measures, norm. Relative Angle Dist. combined with distance entropy shows particularly strong and stable performance, achieving correlations of 0.93–0.96 across all road lengths for `BeamNG.AI`, and 0.94–0.96 for `Dave2`. The Complexity Vectors distance measure also shows strong performance with entropy aggregation (0.92–0.96) across all road lengths.

For the Direct DMs, Feature Map demonstrates consistent strong correlation across all road lengths (0.93–0.96 for `BeamNG.AI`, 0.91–0.94 for `Dave2`), showing minimal sensi-

[9] The Pearson correlation analysis presented in in Appendix E shows similar results.



(a) `BeamNG.AI` (b) `Dave2`

Fig. 9 RQ4b – Spearman Correlation DM and BD by road length. Each row indicates the correlation between the DM on the row and BD. Positive/negative values indicate positive/negative correlation. The test suite's road length is identified as follows: ▼= short TSs, ◆= medium TSs, ▲= long TSs. (Matching Pearson correlation plot in Appendix E)

tivity to road length. Similarly, Test Set Diameter CS maintains strong correlations (0.92–0.95) across all length categories.

Interestingly, Convex Hull shows a clear progression with road length: weak correlation for short roads (0.63 for `BeamNG.AI`, 0.72 for `Dave2`), improving to strong correlation for medium roads (0.83–0.85), and very strong correlation for long roads (0.84 for `BeamNG.AI`, 0.89 for `Dave2`). This pattern is expected given that longer roads naturally span larger spatial areas, making the convex hull a more discriminative measure.

Table 8 RQ4b – Spearman Correlation between DMs and BD by length (`BeamNG.AI` , `Dave2`). Subscripts $_S$, $_M$, and $_L$ indicate short, medium and long roads, respectively. (Matching Pearson correlation table in Appendix E)

		`BeamNG.AI`$_S$	`BeamNG.AI`$_M$	`BeamNG.AI`$_L$	`Dave2`$_S$	`Dave2`$_M$	`Dave2`$_L$
Discrete	Dist. Entropy	0.96	0.95	0.93	0.96	0.93	0.96
Fréchet	Summing	0.95	0.95	0.93	0.95	0.93	0.93
	Averaging	0.06	−0.13	−0.14	0.14	−0.16	0.05
	Avg. Maxima	0.41	0.67	0.55	0.65	0.76	0.74
	Weitzman	0.78	0.77	0.73	0.78	0.75	0.83
Partial	Dist. Entropy	0.95	0.94	0.92	0.95	0.93	0.95
Curve	Summing	0.95	0.95	0.92	0.95	0.93	0.94
Mapping	Averaging	0.07	−0.13	0.01	0.22	0.05	0.14
	Avg. Maxima	0.58	0.66	0.67	0.67	0.76	0.78
	Weitzman	0.63	0.74	0.70	0.71	0.74	0.83
Dynamic	Dist. Entropy	0.96	0.94	0.92	0.96	0.93	0.95
Time	Summing	0.95	0.94	0.92	0.95	0.92	0.94
Warping	Averaging	0.08	−0.05	0.00	0.19	−0.03	0.20
	Avg. Maxima	0.40	0.60	0.49	0.61	0.68	0.74
	Weitzman	0.69	0.77	0.67	0.74	0.73	0.83
norm.	Dist. Entropy	0.96	0.96	0.93	0.96	0.94	0.96
Relative	Summing	0.93	0.95	0.94	0.95	0.95	0.96
Angle	Averaging	0.01	0.14	−0.01	0.11	0.32	0.20
Dist.	Avg. Maxima	0.40	0.77	0.71	0.72	0.89	0.76
	Weitzman	0.82	0.80	0.80	0.86	0.74	0.77
Area	Dist. Entropy	0.96	0.94	0.92	0.96	0.93	0.95
Between	Summing	0.95	0.94	0.92	0.95	0.91	0.94
Curves	Averaging	0.16	−0.05	0.05	0.18	−0.01	0.21
	Avg. Maxima	0.48	0.53	0.61	0.60	0.70	0.73
	Weitzman	0.76	0.75	0.70	0.77	0.69	0.82
Complexity	Dist. Entropy	0.96	0.95	0.95	0.96	0.92	0.95
Vectors	Summing	0.93	0.94	0.94	0.93	0.89	0.94
	Averaging	0.01	−0.17	−0.07	−0.08	−0.19	−0.01

Table 8 (continued)

		BeamNG.AI$_S$	BeamNG.AI$_M$	BeamNG.AI$_L$	Dave2$_S$	Dave2$_M$	Dave2$_L$
	Avg. Maxima	0.42	0.63	0.58	0.64	0.47	0.58
	Weitzman	0.78	0.75	0.73	0.85	0.68	0.66
Iterative	Dist. Entropy	0.96	0.94	0.92	0.96	0.94	0.94
Levenshtein	Summing	0.96	0.94	0.92	0.96	0.93	0.95
	Averaging	0.03	−0.15	0.10	0.12	0.12	0.21
	Avg. Maxima	0.45	0.69	0.53	0.53	0.82	0.65
	Weitzman	0.81	0.80	0.73	0.83	0.75	0.83
Manhattan	Dist. Entropy	0.94	0.85	0.94	0.91	0.90	0.88
Distance	Summing	0.95	0.95	0.95	0.93	0.95	0.95
	Averaging	0.13	−0.01	−0.19	0.04	0.02	0.10
	Avg. Maxima	0.54	0.52	0.58	0.53	0.66	0.59
	Weitzman	0.74	0.59	0.82	0.76	0.79	0.75
Jaccard	Dist. Entropy	0.95	0.95	0.93	0.95	0.93	0.95
Similarity	Summing	0.97	0.95	0.93	0.95	0.93	0.95
(coarse)	Averaging	0.08	−0.34	−0.13	−0.06	0.19	−0.08
	Avg. Maxima	—	—	0.15	—	—	—
	Weitzman	0.77	0.82	0.71	0.79	0.69	0.82
Jaccard	Dist. Entropy	0.97	0.94	0.94	0.93	0.92	0.96
Similarity	Summing	0.97	0.95	0.94	0.93	0.92	0.96
(fine)	Averaging	−0.12	−0.33	0.00	−0.21	−0.15	−0.21
	Avg. Maxima	—	—	—	—	—	—
	Weitzman	0.86	0.77	0.81	0.68	0.61	0.77
Direct	Feature Map	0.96	0.93	0.94	0.91	0.93	0.94
DMs	Test Set Diameter CS	0.94	0.95	0.93	0.95	0.92	0.94
	Convex Hull	0.63	0.83	0.84	0.72	0.85	0.89

Comparing the two agents, correlation patterns remain remarkably similar across all road length categories, with differences typically within 0.03–0.05. This consistency reinforces our finding from ***RQ4a*** that road diversity affects behavioural diversity similarly for both rule-based and learning-based driving agents.

These findings have important practical implications. First, they demonstrate that the choice of appropriate DMs identified in ***RQ4a*** (entropy and summing aggregation) remains valid across test suites containing roads of different lengths. Second, the stability of summing aggregation across road lengths makes it particularly suitable when practitioners cannot control or normalise road lengths in their test suites. Third, for applications specifically focused on either short or long roads, practitioners can confidently use the same DMs without adjustment, though they should be aware that some measures like Convex Hull may be less effective for short roads.

Answer to *RQ4b*: Road length has minimal effect on the strong correlation between DMs and BD observed in ***RQ4a***. DMs using Dist. Entropy aggregation maintain very strong correlations (0.85–0.97) across short, medium, and long roads for both driving agents. Summing aggregation demonstrates exceptional stability, with consistently high correlations regardless of road length. Averaging aggregation remains ineffective across all road lengths (near-zero correlation). Among Direct DMs, Feature Map and Test Set Diameter CS show strong, length-independent correlations (0.93–0.96 and 0.92–0.95, resp.), while Convex Hull performance improves significantly with road length (0.63–0.72 for short roads, 0.84–0.89 for long roads). The similarity of correlation patterns between `BeamNG.AI` and `Dave2` persists across all road length categories. These findings indicate that the DMs recommended in ***RQ4a*** are robust to road length variations, making them suitable for test suites containing roads of diverse lengths.

5 Discussion

5.1 Road Diversity Analysis

Our empirical investigation of 53 road geometry diversity measures evaluated across 97,000 road instances and two architecturally divergent driving agents (`BeamNG.AI` and `Dave2`) constitutes the first systematic and comprehensive evaluation of diversity measures for ADS testing. The convergent evidence across our four research questions establishes clear empirical patterns that advance the theoretical understanding of diversity measure effectiveness in ADS testing contexts.

Dominance of Aggregation Methods Over Distance Functions The most salient finding emerging from our systematic analysis demonstrates that aggregation methods exert substantially greater influence than distance functions in determining both the mathematical properties and empirical effectiveness of diversity measures. This hierarchical relationship manifests consistently across all investigated research questions.

The analysis in ***RQ1*** establishes that aggregation methods constitute the primary determinant of whether a DM satisfies fundamental mathematical properties including monotonicity

and insensitivity to duplicates. Measures employing Dist. Entropy and Summing aggregation consistently exhibited monotonic growth and insensitivity to duplicates across substantially all underlying distance functions, whereas Averaging aggregation systematically failed to satisfy these properties irrespective of the employed distance measure. This pattern indicates that the mathematical structure of aggregation functions—rather than the geometric sophistication of pairwise distance measures—fundamentally determines DM behaviour.

This structural relationship received further empirical support in ***RQ2***, where correlation analysis demonstrated that DMs employing identical aggregation methods exhibit strong positive correlation (typically > 0.8) even when constructed from entirely distinct distance functions. For instance, Dist. Entropy-based DMs constructed from Discrete Fréchet, Dynamic Time Warping, and Area Between Curves distances demonstrate strong mutual correlation despite quantifying fundamentally different geometric properties. Conversely, distinct aggregations of identical distance measures (e.g. Discrete Fréchet with Dist. Entropy versus Discrete Fréchet with Averaging) exhibit weak or negative correlation. These findings indicate that aggregation methods exert dominant influence over the diversity signal encoded by a DM.

The hierarchical dominance of aggregation over distance functions becomes most pronounced in ***RQ3*** and ***RQ4***. In ***RQ3***, Averaging aggregation exhibited uniformly strong correlation with road length (> 0.9) across all distance measures, while Summing demonstrated uniformly weak correlation (-0.21 to 0.41) across identical distance measures. Most significantly, in ***RQ4a***, Dist. Entropy and Summing aggregations achieved very strong correlations with behavioural diversity (0.93–0.95) *independent of the aggregated distance function*, while Averaging exhibited near-zero correlation with BD across all distance functions.

This empirical demonstration of aggregation dominance carries substantial methodological implications. Practitioners can achieve effective diversity measurement by prioritising the selection of appropriate aggregation methods, with distance function selection constituting a secondary consideration. This finding simplifies the diversity measure selection problem and suggests that future research should emphasise the development of novel aggregation approaches rather than new distance metrics.

Dist. Entropy and Summing Aggregations Our systematic evaluation identifies two aggregation methods—Dist. Entropy and Summing —that consistently demonstrate superior performance across all evaluation criteria. Both methods satisfy the desirable property of monotonic growth (***RQ1a***) and exhibit computational efficiency (***RQ1c***), and demonstrate very strong correlation with behavioural diversity (0.93–0.95 in ***RQ4a***).

These methods achieve effectiveness through complementary mechanisms. Dist. Entropy aggregation quantifies the distributional structure of pairwise distances within a test suite, simultaneously rewarding both the presence of dissimilar elements and their frequency distribution. This aligns with fundamental testing objectives of exploring diverse regions of the input space. The strong observed correlation with BD (0.93–0.94) provides empirical support for the hypothesis that distributional diversity in road geometry translates to distributional diversity in driving behaviours.

Summing aggregation, conversely, achieves effectiveness through structural simplicity and robustness. By accumulating all pairwise distances without normalisation, it rewards

test suites where elements exhibit average dissimilarity, while remaining insensitive to the distributional structure of those differences. Notably, Summing demonstrated the most stable performance across road length categories in ***RQ4b***, maintaining correlations of 0.89–0.97 for short, medium, and long roads. This empirically demonstrated stability confers particular value in practical contexts where test suites contain roads of heterogeneous lengths.

Therefore, the selection between Dist. Entropy and Summing depends on the specific testing objectives and constraints. When distributional diversity constitutes the primary objective and computational resources permit marginally longer computation times, Dist. Entropy offers marginally superior correlation with BD. When robustness to road length heterogeneity or computational efficiency constitutes the primary consideration, Summing shows excellent performance with guaranteed monotonicity and enhanced stability.

Systematic Inadequacy of Averaging-Based Aggregations Our empirical findings conclusively demonstrate that Averaging and Avg. Maxima aggregations exhibit systematic inadequacy for diversity measurement in ADS testing contexts. These methods demonstrate multiple fundamental deficiencies, described in the following.

First, they fail to satisfy the monotonicity property (***RQ1a***), such that augmenting a test suite with additional elements can paradoxically *decrease* the measured diversity value. This violation of a fundamental requirement renders them unsuitable for guiding test generation search algorithms that rely on monotonic fitness functions.

Second, they exhibit extreme sensitivity to road length (***RQ3***), with correlations exceeding 0.9 with road length alone. This implies that their diversity scores primarily reflect average road length rather than geometric diversity, rendering them uninformative for comparing test suites with differing road length distributions.

Third, and most critically, they exhibit near-zero correlation with behavioural diversity (***RQ4a***), indicating that they provide no predictive information regarding whether a test suite will exercise diverse driving behaviours. The convergence of these deficiencies renders Averaging-based aggregations not merely ineffective but actively misleading: they can indicate high diversity for test suites that exercise limited behaviours, and conversely.

The systematic failure of averaging-based measures can be attributed to the normalisation by test suite cardinality inherent to the averaging operation. While such normalisation may appear mathematically justified, it conflicts fundamentally with the accumulative nature of diversity. A test suite comprising 100 diverse elements possesses genuinely greater diversity than a test suite of 10 diverse elements, even when the average pairwise distance remains constant. By normalising away this accumulative property, averaging-based measures eliminate the essential signal required for effective testing guidance.

Effectiveness of Direct Diversity Measures Among the three Direct DMs evaluated in our study, Feature Map and Test Set Diameter CS demonstrated exceptional effectiveness, achieving the strongest correlations with behavioural diversity observed in this investigation. This finding merits particular attention given that these measures operate on fundamentally distinct principles compared to pairwise aggregation approaches.

The effectiveness of Feature Map can be attributed to its explicit measurement of high-level road characteristics (curvature, complexity, direction coverage) that directly influence

driving behaviour. Through discretisation of roads into feature-based partitions, it inherently quantifies coverage of behaviourally-relevant road properties. The consistency of its performance across road length categories (***RQ4b***) combined with its computational efficiency (***RQ1c***) establish it as a particularly viable option for practical deployment.

The strong performance of Test Set Diameter CS merits equal attention, particularly given that it quantifies compressibility rather than explicit geometric properties. Its effectiveness provides empirical support for the hypothesis that roads exhibiting low joint compressibility (attributable to dissimilar structural patterns) tend to induce diverse driving behaviours. This finding provides theoretical justification for information-theoretic approaches to diversity measurement.

Conversely, Convex Hull exhibited limited effectiveness for short roads despite strong correlation for long roads. These contrasting results demonstrate that computational simplicity does not ensure measurement effectiveness, and that careful empirical validation remains essential.

Agent-Independence of Correlation Patterns One of the most theoretically significant findings of this investigation concerns the near-identical correlation patterns observed between `BeamNG.AI` (rule-based, complete environmental knowledge) and `Dave2` (learning-based, camera-only perception) across all evaluated diversity measures. In ***RQ4a***, correlation coefficients exhibited maximal deviation of 0.03 between agents, and this agent-independence persisted across all road length categories (***RQ4b***).

This agent-independence carries substantial implications. It provides empirical evidence that road geometry diversity influences behavioural diversity through mechanisms that transcend specific agent architectures. Whether an agent employs classical planning algorithms with complete road knowledge or end-to-end neural networks with visual perception only, geometric diversity in roads translates to behavioural diversity through fundamentally similar mechanisms.

This finding establishes strong empirical support for the generalisability of road diversity measures across distinct ADS implementation paradigms. Practitioners can develop test suites using diversity measures validated on one agent architecture with justified confidence that those measures will demonstrate effectiveness for alternative agent architectures. This substantially enhances the practical utility of our findings and suggests that road geometry captures fundamental aspects of the driving task that influence all agents comparably.

Road Length as a Confounding Variable ***RQ3*** identified road length as a significant confounding variable for numerous diversity measures, particularly those employing Averaging aggregation. However, the analysis in ***RQ4b*** demonstrates that this confound does not compromise the effectiveness of the recommended measures (Dist. Entropy and Summing).

The critical distinction lies in the observation that while Averaging-based measures exhibit high correlation with road length *and* low correlation with BD, Dist. Entropy and Summing exhibit low-to-moderate correlation with road length *and* high correlation with BD. This pattern indicates that effective diversity measures successfully capture geometric diversity that extends beyond simple length-based effects. When roads of differing lengths exhibit geometrically similar structure (scaled versions), entropy and summing appropri-

ately recognise their similarity; when roads of differing lengths exhibit geometrically distinct structure, these measures appropriately recognise their diversity.

The stability of Summing aggregation across road length categories (0.81–0.90 correlation with BD in ***RQ4b***) possesses particular practical value. It indicates that practitioners need not implement length normalisation or restrict test suites to homogeneous-length roads when employing summing-based DMs. The measure intrinsically distinguishes geometric diversity from length variation.

Computational Efficiency and Practical Deployment The efficiency analysis in ***RQ1c*** and additivity analysis in ***RQ1d*** establish crucial practical constraints. For pairwise-distance-based DMs, the selection of distance function dominates computational requirements, with Iterative Levenshtein requiring 25 minutes per 100-road test suite while norm. Relative Angle Dist. requires less than 8 seconds. The aggregation operation itself imposes negligible computational overhead (typically < 0.1 seconds), with the singular exception of Weitzman aggregation, which exceeds 24-hour computational limits for test suites exceeding 20 roads.

This computational characterisation, synthesised with the effectiveness findings from ***RQ4***, enables evidence-based recommendations. Given that Dist. Entropy and Summing aggregations demonstrate comparable correlation with BD (0.88–0.96 versus 0.89–0.97) while both impose negligible computational overhead, distance function selection should prioritise computational efficiency. The combination of norm. Relative Angle Dist. with either Dist. Entropy or Summing aggregation provides optimal balance of effectiveness and efficiency among pairwise-based measures.

However, the Direct DMs (particularly Feature Map) offer a more advantageous alternative. With computation times below 1 second for test suites of 100 roads and the strongest observed correlation with BD (0.95), Feature Map represents the optimal selection when both effectiveness and efficiency constitute priorities. Its satisfaction of monotonicity, insensitivity to duplicates, and robustness to road length variation establish its suitability across diverse testing scenarios.

Evidence-Based Recommendations for Practice Synthesising the empirical findings across all research questions enables the formulation of evidence-based recommendations for practitioners:

- For *maximal correlation with behavioural diversity*, we recommend Feature Map (***RQ4a***: correlation 0.95) or Dist. Entropy aggregation combined with any computationally feasible distance function (***RQ4a***: correlation 0.94–0.96). The difference in effectiveness between these approaches is minimal; the selection should depend on implementation constraints and computational resources.
- For *robustness to heterogeneous road lengths*, we suggest Summing aggregation, which maintains correlation 0.89–0.97 with BD across all road length categories (***RQ4b***), or Feature Map, which demonstrates comparable stability.
- For *computational efficiency*, we suggest using Feature Map (computation time < 1 second for 100 roads, ***RQ1c***) or the combination of norm. Relative Angle Dist. with Dist. Entropy or Summing aggregation (computation time < 8 seconds for 100 roads,

RQ1c). We suggest to avoid Iterative Levenshtein (25 minutes for 100 roads), unless specific geometric properties mandate its application.

- *Guaranteed monotonicity* is achieved using Dist. Entropy or Summing aggregation with most distance functions (***RQ1a***), or Feature Map among direct measures. We explicitly suggest to avoid Averaging and Avg. Maxima .
- *Insensitivity to duplicates* is obtained using Dist. Entropy or Weitzman aggregation with most distance functions (***RQ1b***), or Feature Map and Convex Hull among direct measures; however, the computational cost of Weitzman renders it almost unusable for larger TSs. Note that Summing exhibits intentional sensitivity to duplicates, which may be appropriate in specific testing contexts.
- *Under conditions of uncertainty*: Feature Map provides optimal balance across effectiveness (maximal correlation with BD), efficiency (sub-second computation), and mathematical properties (monotonicity, duplicate insensitivity, length robustness). It constitutes the strongest recommendation in general.

Implications for Automated Test Generation Our findings carry immediate implications for diversity-driven test generation and test suite optimisation methodologies. Search-based test generation algorithms employing diversity as an optimisation objective (such as e.g. Klikovits et al. 2023) should utilise Dist. Entropy or Summing aggregation, as these methods show strong correlation with behavioural diversity, while satisfying the monotonicity requirement essential for gradient-based optimisation. The computational efficiency of these measures (particularly when combined with computationally efficient distance functions such as norm. Relative Angle Dist.) establishes their viability even for generation algorithms that evaluate thousands of candidate test suites during search.

For test suite minimisation, where the objective function seeks to identify a minimal subset of tests maintaining maximal diversity, the strong empirically demonstrated correlation between geometric diversity and behavioural diversity validates the use of DM as a stand-in for BD. A selection algorithm maximising Feature Map or entropy-based diversity will systematically select roads that exercise diverse driving behaviours, despite utilising exclusively geometric information. The demonstrated agent-independence establishes that the same minimised test suite should demonstrate effectiveness across distinct ADS implementations.

Limitations and Future Research Directions While this investigation establishes definitive empirical answers to numerous questions regarding road diversity measurement, it simultaneously identifies important avenues for future investigation. First, our exclusive focus on road geometry necessitates investigation of how road diversity interacts with additional scenario elements (meteorological conditions, traffic density, regulatory infrastructure). Do the same aggregation principles generalise to diversity measurement across multi-dimensional scenario spaces?

Second, while we demonstrated strong correlation between road diversity and behavioural diversity, the correlation remains imperfect (0.94–0.97 for optimal measures). What factors account for the residual variance? Do unquantified geometric properties influence driving behaviour? Or does inherent stochasticity in ADS behaviour impose fundamental limits on achievable correlation?

Third, our study evaluated existing diversity measures without attempting to derive optimal measures from first principles. Could supervised learning approaches discover superior aggregation functions or distance measures through explicit optimisation for correlation with behavioural diversity? The consistent empirical patterns we observed suggest that such optimisation could yield theoretical insights.

Finally, while our agent-independence finding demonstrates robustness across two architecturally different agents. However, since both considered agents represent lane-keeping systems in a simulation-based setting rather than full-scale production ADS stacks, validation across a broader range of ADS implementations—including production systems, hybrid architectures, and agents with heterogeneous sensor modalities—would strengthen confidence in universal generalisability of these findings.

5.2 Threats to Validity

The validity of our study may be affected by different types of threats (Wohlin et al. 2012) that we discuss in the following.

Construct Validity A construct validity threat could be that the approaches we use to investigate our research questions are not appropriate. If this is the case, we may draw wrong conclusions regarding the investigated RQs; for example, we may observe correlation where there is none. To mitigate this threat, we carefully designed the experiments to answer the RQs. To assess monotonicity, we explicitly checked that the diversity does not decrease when increasing the number of considered roads. For monotonic measures, we checked how they grow. We proceeded similarly for "insensitivity to duplicates". In order to check correlation, we used either the Pearson's correlation test or the Spearman's correlation test depending on whether the data is normally distributed or not. Moreover, we also considered the strength of the correlation when reporting the results. With regards to measures of BD, a threat could be that the one we selected is not representative of the behavioural diversity in the ADSs that we use, or that it does not generalise. This could threaten the conclusions we draw in ***RQ4***. To address this threat, instead of coming up with our own BD measures, we looked at literature (Jahangirova et al. 2021) to identify and select a BD measure that have been used and studied before, namely feature maps (Gambi et al. 2022).

Internal Validity An internal validity threat could be that the observed correlation (if any) between road diversity and behaviour diversity is due to other confounding factors. For example, the behaviour of a driving agent depends not only on the road, but also on other elements of the driving scenarios such as road participants, regulatory elements, weather conditions, etc. In complex scenarios in which all these elements are present, it would very difficult (if possible at all) to isolate and measure the influence of the road structure on the driving behaviour. Therefore, there is the risk that we wrongly observe the correlation because the behaviour diversity is triggered by the different elements of the driving scenarios and not the road structure. To mitigate this threat, we use scenarios composed of only one road, without any other elements. The correlation of behaviour diversity with other traffic elements should be part of independent studies and we leave it as future work.

Other confounding factors could be the test suite size and the length of the road. Regarding test suite size, as for some metrics (i.e. those guaranteeing monotonicity) more tests in general

lead to higher values, there is the risk of observing a positive correlation that, however, is only due to the test suite size. To mitigate this threat, we analyse the data by considering different test suite sizes separately, as described in Section 4. Regarding the length of the road, instead, we mitigate this threat by assessing the influence of road length on DMs (***RQ3***) and BD (***RQ4.b***).

External Validity A threat of this type is that the results of our study could be not generalisable to other types of roads, and other driving agents. To mitigate this threat, we sample from a very large set of roads created by different road generators, that can possibly generate roads of different shapes. Moreover, to assess the influence of road length on the diversity and correlation with behaviour diversity, we artificially shortened roads to generate roads of different lengths. Our findings should be interpreted as evidence that road diversity is a strong proxy for behavioural diversity in the considered controlled setting (i.e. a single ego vehicle following a non-intersecting two-lane roads, without modelling additional traffic participants, intersections, pedestrians, or obstacles). In more complex scenarios, behavioural diversity may additionally depend on interactions with other road participants and environmental conditions. Extending the study to such richer scenarios is left for future work. Moreover, different agents could react differently to varying types of roads and so lead to different behaviour diversity values, and so different correlation with road diversity. To mitigate this threat, we consider two very different types of driving agents: `BeamNG.AI`, a rule-based driving agent exploiting the complete knowledge of the road geometry, and `Dave2`, a deep learning-based agent designed at NVIDIA (Bojarski et al. 2016). However, we recognise that other driving agents like Baidu Apollo and Autoware are available and we leave their investigation as future work.

6 Conclusions

Current ADSs testing practices rely on exposing the SUT to suites of diverse scenarios based on the fundamental assumption that scenario diversity induces behavioural diversity. However, this assumption so far lacked rigorous empirical validation, and the diversity measures employed in ADS testing had never been systematically evaluated or compared.

Our study constitutes the first comprehensive empirical investigation of road geometry diversity measures for ADS testing. We evaluated 53 diversity measures across 97,000 individual road geometries and two architecturally divergent driving agents (`BeamNG.AI` and `Dave2`), systematically addressing four research questions concerning mathematical properties (***RQ1***), pairwise correlations (***RQ2***), road length effects (***RQ3***), and correlation with behavioural diversity (***RQ4***).

Our investigation demonstrates very strong positive correlation (0.94–0.97) between road geometry diversity and behavioural diversity for both a rule-based and a learning-based driving agent. This empirically validates the fundamental assumption underlying diversity-driven ADS testing and justifies the use of geometric diversity measures as proxies for behavioural diversity. The strength of this correlation establishes that practitioners can confidently employ geometric diversity as a surrogate metric for behavioural diversity in contexts where direct behavioural measurement is computationally prohibitive.

A particularly significant finding concerns the dominance of aggregation methods over distance functions in determining empirical effectiveness of diversity measures. This hierarchical

relationship manifests across all research questions, demonstrating that the choice of aggregation method is the main factor for effectiveness. This fact simplifies the DM selection problem and suggests that practitioners should carefully choose a pair of beneficial aggregation method and a matching distance measure that is computationally feasible. Moreover, future research should prioritise novel aggregation approaches rather than additional distance functions.

Our evaluation identifies Dist. Entropy and Summing aggregations as consistently superior across all evaluation criteria, achieving very high correlations of 0.93–0.95 with behavioural diversity while satisfying monotonicity and duplicate insensitivity requirements. Among Direct DMs, Feature Map achieves the strongest correlation (0.95), while maintaining sub-second computational efficiency. Conversely, Averaging and Avg. Maxima aggregations exhibit near-zero correlation with behavioural diversity, extreme sensitivity to road length (> 0.9 correlation), and violation of monotonicity requirements, indicating that these measures should be explicitly avoided in ADS testing contexts.

The nearly identical correlation patterns observed across `BeamNG.AI` and `Dave2` (maximum deviation 0.03) demonstrate that road diversity is not affected by the choice of specific agent architectures, at least within the same simulator.

Our empirical findings provide practitioners with evidence-based guidance for diversity measure selection in search-based test generation, test suite minimisation, and diversity-driven quality assurance for ADSs. The comprehensive evaluation framework, detailed results, and interactive visualisations are publicly available to support reproducibility and facilitate adoption in both research and industrial contexts.

6.1 Future Work

While this investigation provides definitive empirical answers regarding road geometry diversity measurement, several important directions for future research emerge from our findings.

Our investigation employed two driving agents (`BeamNG.AI` and `Dave2`) evaluated within the BeamNG.tech simulator using a specific behavioural diversity metric from literature that is based on a 17-dimensional feature map. While this approach demonstrated remarkable consistency across both agents (despite architectural difference), extending the evaluation to additional simulators (such as CARLA, LGSVL, or Apollo), alternative behavioural diversity metrics (based on different feature extractors or distance measures), and a broader range of driving agents would provide stronger evidence for the generalisability of our findings. Investigating whether the identified aggregation principles remain valid across different simulation platforms and behavioural characterisations would establish the universal applicability of our recommendations.

Validating our findings across a broader range of ADS architectures—including production implementations, hybrid architectures, heterogeneous sensor modalities, and different control paradigms—would strengthen confidence in universal generalisability. Additionally, investigating whether the sequence in which diverse roads are executed influences test suite effectiveness, and examining the relationship between diversity measures and actual fault detection across different fault types, would provide valuable practical insights. Finally, algorithmic optimisations and approximation techniques could enable application of diversity measures to industrial-scale test repositories, facilitating the transition from controlled experimental contexts to comprehensive industrial ADS validation.

Our exclusive focus on road geometry represents a deliberate simplification necessary for rigorous empirical investigation. Future research should investigate whether the aggregation prin-

ciples identified in this study—particularly the superiority of Dist. Entropy and Summing over Averaging—generalise to diversity measurement across multi-dimensional scenario spaces incorporating meteorological conditions, traffic participants, and regulatory infrastructure. Understanding these interactions would enable more comprehensive diversity-driven testing strategies.

More specifically, an important next step is to extend our analysis of the geometry of the roads to scenarios identified with other elements, such as road signs and signals, as well as pedestrians, other vehicles and their interactions, and to study how the diversity of such scenarios affects behavioural diversity. In this paper, we considered a variety of diversity metrics for characterising road-shape diversity, but they rely on geometric properties of roads. When traffic elements, pedestrians, or other vehicles are also included, diversity will require definitions that are necessarily beyond what we considered in this paper. For multi-dimensional scenario spaces, it may be possible to define diversity for each element and then combine these definitions to define an overall scenario diversity. One possible approach is to compute diversity of each scenario element separately and then combine the resulting values using summation after normalisation. This would ensure that each scenario element is taken into account. Another approach is to define the scenario diversity as the minimum of the diversity values for the individual scenario elements. This may create a new notion of diversity, where a set of scenarios is considered diverse if all scenario elements are sufficiently diverse. We believe that investigating whether such combined definitions are appropriate and useful can be an interesting research direction. Another important direction for future work is to study whether the stronger behavioural diversity observed in more road-diverse test suites also leads to higher fault-revealing effectiveness.

In this paper, we used Spearman's correlation to identify rank-based associations between different DMs. A potential extension is to also analyse top-k overlap. In our problem setting, this means whether two road-diversity measures identify the same top-k most diverse test suites in terms of their road geometry. Such an analysis can be practically important for the test-case selection problem (Birchler et al. 2023b, c, a; Arcaini and Cetinkaya 2026), where only a limited number of test cases can be selected for execution, as it is essential to identify whether the choice of the DM matters significantly for this problem.

Appendix A: Data Description and Pre-Evaluation

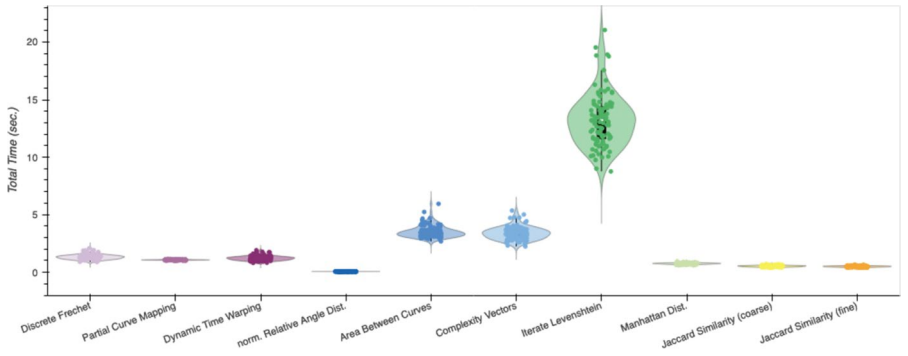


Fig. 10 TS10 – Total distance computation time (in sec.)

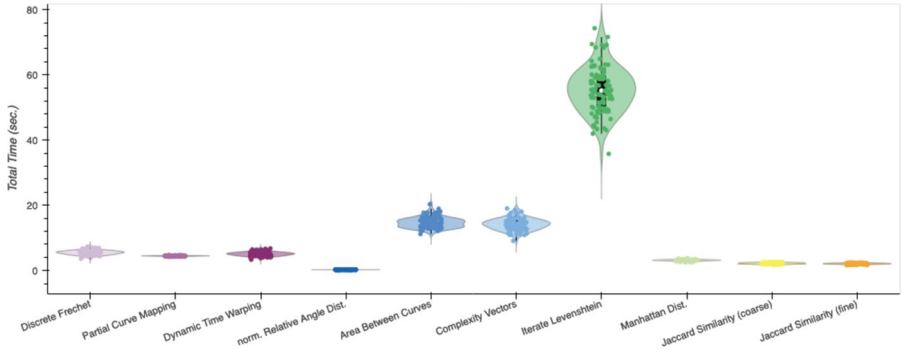


Fig. 11 TS20 – Total distance computation time (in sec.)

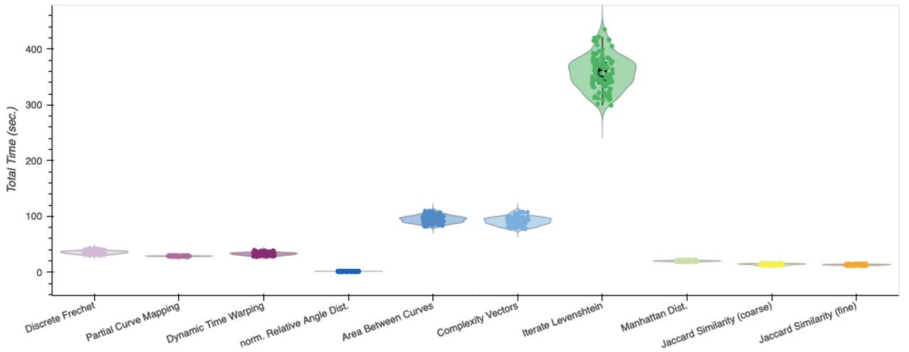


Fig. 12 TS50 – Total distance computation time (in sec.)

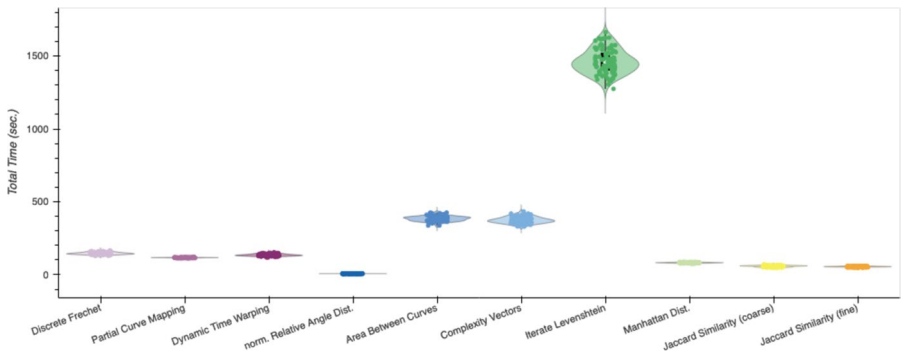


Fig. 13 TS100 – Total distance computation time (in sec.)

Appendix B: RQ1a: Growth and Monotonicity

Table 9 RQ1a – Relative growth (25%-quantile values)

		TS10				TS20				TS50				TS100			
	0.25	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%	+10%	+20%	+50%	+100%
Discrete Fréchet	Dist. Entropy	1.06	1.16	1.38	1.84	1.07	1.14	1.37	1.79	1.08	1.15	1.39	1.81	1.07	1.16	1.41	1.82
	Summing	1.18	1.39	2.17	3.89	1.19	1.4	2.19	3.89	1.2	1.43	2.2	3.92	1.2	1.43	2.23	3.91
	Averaging	0.96	0.95	0.93	0.92	0.98	0.97	0.96	0.95	0.99	0.99	0.97	0.97	0.99	0.99	0.99	0.97
	Avg. Maxima	0.99	0.98	0.98	1.0	0.99	0.99	1.0	1.01	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.01
	Weitzman	1.04	1.09	1.22	1.49	1.05	1.09	—	—	—	—	—	—	—	—	—	—
Partial Curve Mapping	Dist. Entropy	1.05	1.13	1.36	1.78	1.06	1.14	1.36	1.75	1.06	1.14	1.36	1.77	1.07	1.16	1.4	1.81
	Summing	1.16	1.37	2.08	3.7	1.17	1.38	2.13	3.78	1.19	1.42	2.19	3.88	1.2	1.42	2.21	3.91
	Averaging	0.95	0.93	0.89	0.88	0.97	0.95	0.93	0.92	0.99	0.98	0.97	0.96	0.99	0.99	0.98	0.97
	Avg. Maxima	0.97	0.97	0.96	1.01	0.98	0.98	0.99	1.02	0.99	1.0	1.01	1.03	0.99	1.0	1.0	1.06
	Weitzman	1.02	1.06	1.15	1.4	1.03	1.06	—	—	—	—	—	—	—	—	—	—
Dynamic Time Warping	Dist. Entropy	1.03	1.13	1.35	1.79	1.05	1.12	1.35	1.8	1.06	1.15	1.37	1.8	1.07	1.16	1.41	1.83
	Summing	1.16	1.37	2.09	3.71	1.17	1.39	2.15	3.79	1.19	1.42	2.17	3.82	1.19	1.42	2.21	3.88
	Averaging	0.95	0.93	0.9	0.88	0.96	0.96	0.94	0.92	0.98	0.98	0.96	0.95	0.99	0.98	0.98	0.97
	Avg. Maxima	0.99	0.98	0.98	0.99	0.99	0.99	1.0	1.0	1.0	0.99	1.0	1.01	1.0	1.0	1.0	1.01
	Weitzman	1.02	1.06	1.2	1.44	1.04	1.07	—	—	—	—	—	—	—	—	—	—
norm. Relative Angle Dist.	Dist. Entropy	1.1	1.21	1.56	2.17	1.1	1.18	1.49	2.03	1.09	1.18	1.47	1.97	1.09	1.19	1.48	1.96
	Summing	1.19	1.42	2.22	4.0	1.2	1.42	2.22	3.95	1.2	1.43	2.22	3.98	1.2	1.43	2.23	3.96
	Averaging	0.97	0.97	0.95	0.95	0.99	0.98	0.97	0.96	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.98
	Avg. Maxima	0.98	0.99	0.98	1.01	0.99	0.99	1.0	1.02	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.01
	Weitzman	1.06	1.13	1.35	1.72	1.07	1.13	—	—	—	—	—	—	—	—	—	—
Area Between Curves	Dist. Entropy	1.04	1.15	1.42	1.94	1.06	1.15	1.4	1.92	1.07	1.16	1.4	1.85	1.08	1.16	1.43	1.86
	Summing	1.15	1.36	2.1	3.7	1.17	1.37	2.18	3.85	1.19	1.41	2.17	3.87	1.19	1.42	2.22	3.88
	Averaging	0.94	0.93	0.9	0.88	0.97	0.95	0.95	0.94	0.98	0.98	0.96	0.96	0.98	0.98	0.98	0.97
	Avg. Maxima	0.98	0.98	0.98	1.0	0.99	0.99	1.0	1.0	1.0	0.99	1.0	1.01	1.0	1.0	1.0	1.02
	Weitzman	1.03	1.09	1.23	1.5	1.04	1.09	—	—	—	—	—	—	—	—	—	—
Complexity Vectors	Dist. Entropy	1.06	1.16	1.43	1.87	1.06	1.14	1.37	1.77	1.06	1.14	1.37	1.75	1.08	1.16	1.4	1.79
	Summing	1.18	1.39	2.2	3.92	1.17	1.4	2.22	3.97	1.19	1.41	2.2	3.93	1.19	1.41	2.18	3.89
	Averaging	0.97	0.95	0.94	0.93	0.96	0.96	0.97	0.97	0.98	0.98	0.97	0.97	0.98	0.98	0.96	0.97
	Avg. Maxima	0.98	0.98	1.0	1.03	0.99	0.99	1.0	1.02	1.0	0.99	1.0	1.02	1.0	0.99	1.0	1.0
	Weitzman	1.03	1.09	1.25	1.51	1.04	1.08	—	—	—	—	—	—	—	—	—	—
Iterative Levenshtein	Dist. Entropy	1.09	1.22	1.55	2.2	1.1	1.19	1.53	2.1	1.09	1.2	1.52	2.08	1.1	1.21	1.52	2.05
	Summing	1.19	1.41	2.21	3.97	1.19	1.41	2.21	3.94	1.2	1.43	2.21	3.91	1.2	1.43	2.22	3.94
	Averaging	0.97	0.96	0.95	0.94	0.98	0.97	0.97	0.96	0.99	0.99	0.98	0.97	0.99	0.99	0.98	0.98
	Avg. Maxima	0.99	0.99	1.0	0.99	0.99	0.99	0.99	1.0	0.99	0.99	0.99	1.0	1.0	0.99	1.0	1.0
	Weitzman	1.04	1.12	1.31	1.63	1.06	1.12	—	—	—	—	—	—	—	—	—	—
Manhattan Distance	Dist. Entropy	1.03	1.13	1.38	1.8	1.03	1.11	1.32	1.62	1.05	1.09	1.25	1.5	1.03	1.08	1.2	1.38
	Summing	1.17	1.39	2.19	3.82	1.18	1.4	2.18	3.89	1.19	1.41	2.18	3.88	1.19	1.42	2.21	3.9
	Averaging	0.96	0.95	0.94	0.9	0.97	0.96	0.95	0.95	0.98	0.97	0.96	0.96	0.99	0.99	0.98	0.97
	Avg. Maxima	0.99	0.98	0.99	1.0	0.99	0.99	1.0	1.02	1.0	0.99	0.99	1.0	1.0	1.0	1.0	1.02
	Weitzman	1.03	1.09	1.26	1.51	1.04	1.09	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (coarse)	Dist. Entropy	1.15	1.29	1.7	2.46	1.11	1.22	1.58	2.17	1.09	1.2	1.5	2.0	1.09	1.19	1.48	1.96
	Summing	1.22	1.46	2.31	4.17	1.21	1.45	2.27	4.08	1.21	1.44	2.26	4.02	1.21	1.44	2.25	4.01
	Averaging	1.0	1.0	0.99	0.99	1.0	1.0	0.99	0.99	1.0	1.0	1.0	0.99	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.1	1.2	1.46	1.92	1.08	1.16	—	—	—	—	—	—	—	—	—	—
Jaccard Similarity (fine)	Dist. Entropy	1.16	1.31	1.81	2.67	1.13	1.26	1.65	2.36	1.11	1.23	1.59	2.19	1.11	1.22	1.57	2.14
	Summing	1.22	1.46	2.33	4.2	1.21	1.45	2.28	4.09	1.21	1.44	2.26	4.03	1.21	1.44	2.26	4.02
	Averaging	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Avg. Maxima	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Weitzman	1.11	1.21	1.52	2.03	1.1	1.19	—	—	—	—	—	—	—	—	—	—
Direct DMs	Feature Map	1.1	1.2	1.4	1.87	1.06	1.16	1.35	1.72	1.05	1.12	1.31	1.6	1.05	1.1	1.26	1.49
Test	Set Diameter CS	1.05	1.13	1.32	1.64	1.07	1.14	1.39	1.72	1.09	1.17	1.43	1.87	1.09	1.18	1.43	1.85
	Convex Hull	1.0	1.0	1.04	1.16	1.0	1.0	1.06	1.13	1.0	1.01	1.03	1.08	1.0	1.0	1.03	1.06

Appendix C: RQ1c: Efficiency

Table 10 RQ1c – Efficiency – Mean, standard deviation and median of aggregation time (in seconds)

		TS10			TS20			TS50			TS100		
		Mean	Std	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median
Discrete	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Fréchet	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.001	0.019	0.047	0.001	0.047	0.097	0.002	0.097
	Weitzman	0.083	0.006	0.080	215.540	14.079	214.460	—	—	—	—	—	—
Partial	Dist. Entropy	0.002	0.000	0.002	0.002	0.001	0.002	0.003	0.001	0.003	0.005	0.001	0.005
Curve	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mapping	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.010	0.000	0.010	0.019	0.001	0.019	0.048	0.001	0.048	0.098	0.002	0.098
	Weitzman	0.086	0.011	0.083	195.153	8.387	193.241	—	—	—	—	—	—
Dynamic	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Time	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Warping	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.001	0.018	0.047	0.001	0.047	0.097	0.002	0.097
	Weitzman	0.082	0.006	0.080	215.565	14.578	213.783	—	—	—	—	—	—
norm.	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Relative	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Angle	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Dist.	Avg. Maxima	0.009	0.000	0.009	0.019	0.000	0.018	0.047	0.001	0.047	0.097	0.002	0.097
	Weitzman	0.083	0.006	0.080	215.320	15.923	211.720	—	—	—	—	—	—
Area	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Between	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Curves	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.000	0.018	0.047	0.001	0.047	0.097	0.002	0.097
	Weitzman	0.083	0.006	0.081	213.916	15.229	211.635	—	—	—	—	—	—

Table 10 (continued)

		TS10			TS20			TS50			TS100		
		Mean	Std	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median
Complexity	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Vectors	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.000	0.019	0.047	0.001	0.048	0.097	0.002	0.097
	Weitzman	0.083	0.008	0.081	197.273	8.007	196.082	—	—	—	—	—	—
Iterative	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.004	0.000	0.004
Levenshtein	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.000	0.018	0.047	0.001	0.047	0.097	0.002	0.097
	Weitzman	0.083	0.006	0.081	214.677	16.843	211.605	—	—	—	—	—	—
Manhattan	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.003	0.000	0.003
Distance	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.018	0.001	0.018	0.047	0.001	0.047	0.096	0.002	0.096
	Weitzman	0.095	0.086	0.081	213.104	14.082	209.246	—	—	—	—	—	—
Jaccard	Dist. Entropy	0.002	0.001	0.002	0.003	0.001	0.002	0.003	0.001	0.003	0.004	0.001	0.004
Similarity	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(coarse)	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.010	0.000	0.009	0.019	0.001	0.019	0.047	0.001	0.048	0.097	0.003	0.097
	Weitzman	0.087	0.010	0.085	203.770	9.840	201.966	—	—	—	—	—	—
Jaccard	Dist. Entropy	0.002	0.000	0.002	0.002	0.000	0.002	0.002	0.000	0.002	0.003	0.001	0.003
Similarity	Summing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(fine)	Averaging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Avg. Maxima	0.009	0.000	0.009	0.019	0.001	0.019	0.047	0.001	0.047	0.099	0.029	0.096
	Weitzman	0.084	0.009	0.081	206.958	9.682	205.186	—	—	—	—	—	—
Direct	Feature Map	0.074	0.007	0.074	0.147	0.009	0.147	0.365	0.017	0.364	0.736	0.027	0.735
DMs	Test Set Diameter CS	0.002	0.000	0.002	0.005	0.005	0.004	0.010	0.001	0.010	0.019	0.002	0.019
	Convex Hull	0.003	0.000	0.002	0.004	0.001	0.004	0.007	0.001	0.007	0.014	0.001	0.013

Appendix D: RQ2: Pearson Correlation

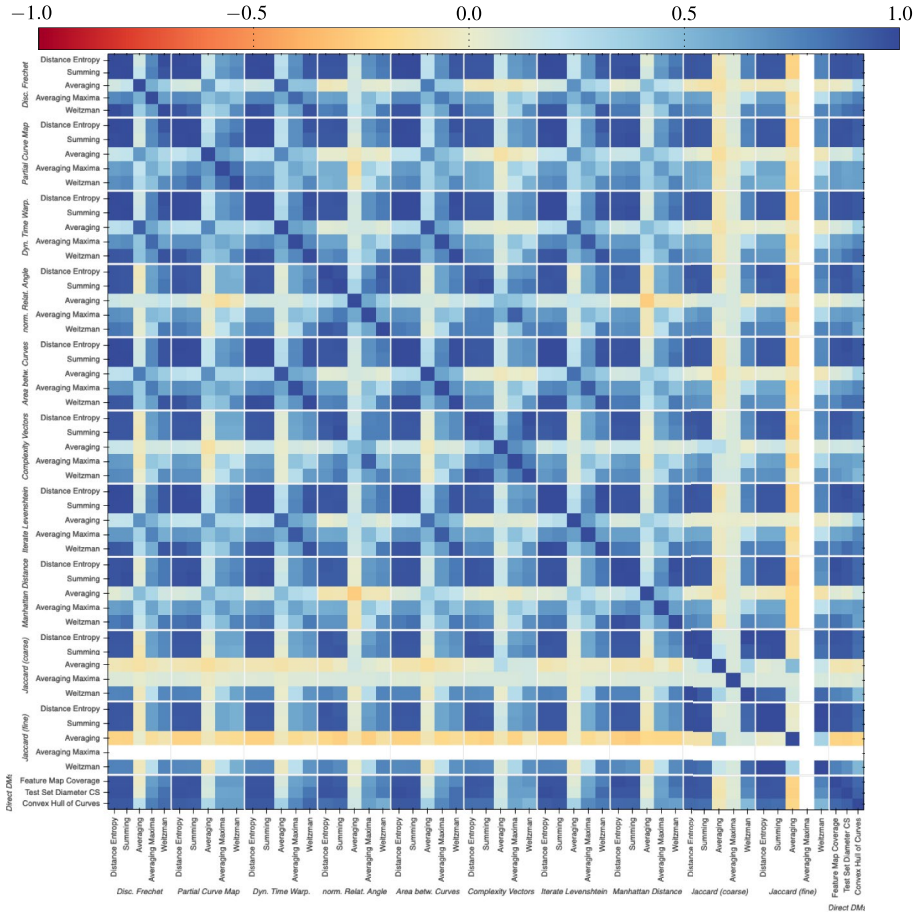
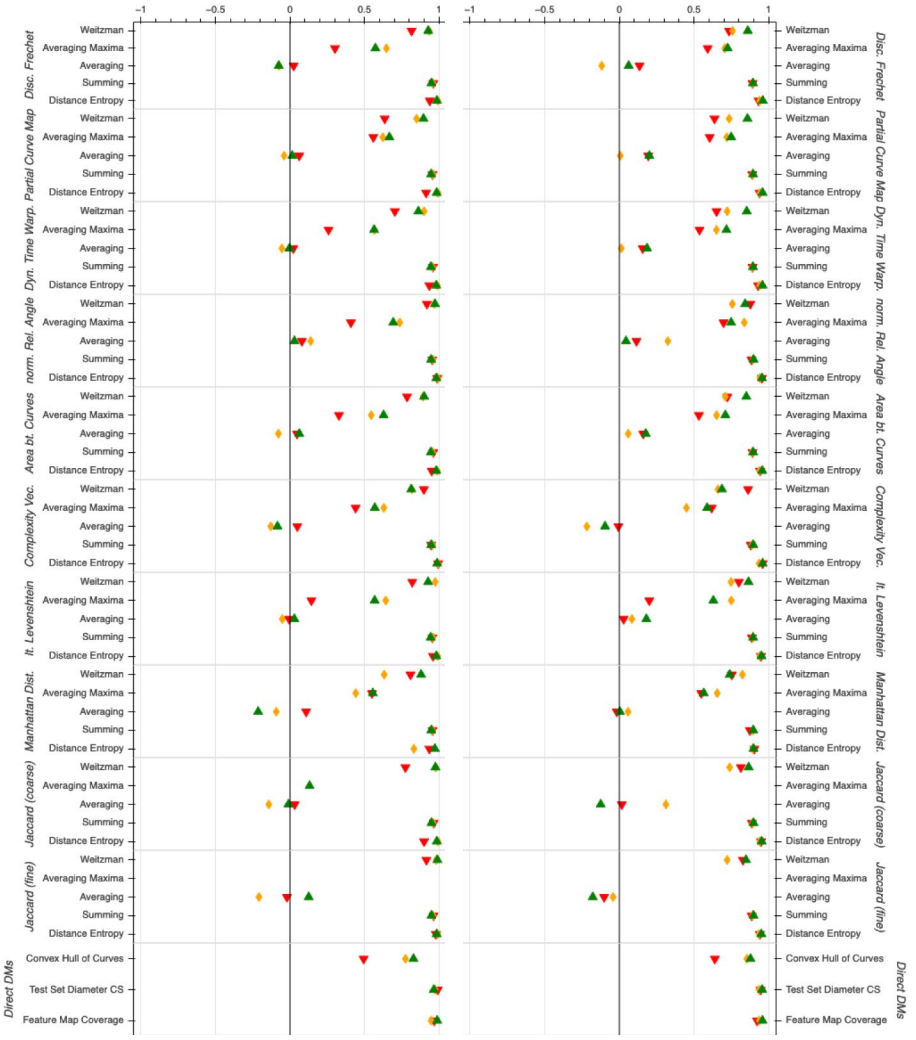


Fig. 14 RQ2 – Correlation heatmap. Each cell indicates the Pearson correlation between the DMs on the row and on the column. Shades of blue indicate positive correlations, while shades of red indicate negative correlation. A larger and interactive version of this figure is available online at http://stklik.github.io/2026-EMSE-Road-Diversity/

Appendix E: RQ4b: Pearson Correlation DM and BD by Road Length



(a) BeamNG.AI (b) Dave2

Fig. 15 RQ4b – Pearson Correlation DM and BD by road length. Each row indicates the correlation between the DM on the row and BD. Positive/negative values indicate positive/negative correlation. The test suite's road length is identified as follows: ▼= short TSs, ◆= medium TSs, ▲= long TSs

Author Contributions All authors contributed to the study conception and design. Raw Data was provided by A. Gambi and V. Riccio, who organised the 2022 SBST CPS Competition. E. Castellano and A. Cetinkaya wrote initial data analysis scripts. S. Klikovits implemented the full experimentation workflow. Evaluation

Table 11 RQ4b – Pearson Correlation between DMs and BD by length (BeamNG.AI, Dave2). Subscripts $_S$, $_M$, and $_L$ indicate short, medium and long roads, respectively

		BeamNG.AI$_S$	BeamNG.AI$_M$	BeamNG.AI$_L$	Dave2$_S$	Dave2$_M$	Dave2$_L$
Discrete	Dist. Entropy	0.94	0.99	0.99	0.93	0.94	0.96
Fréchet	Summing	0.96	0.96	0.95	0.89	0.89	0.90
	Averaging	0.02	−0.07	−0.08	0.14	−0.12	0.06
	Avg. Maxima	0.30	0.65	0.57	0.59	0.71	0.73
	Weitzman	0.82	0.93	0.93	0.73	0.76	0.86
Partial	Dist. Entropy	0.91	0.99	0.98	0.94	0.94	0.96
Curve	Summing	0.96	0.96	0.95	0.89	0.89	0.90
Mapping	Averaging	0.06	−0.04	0.02	0.19	0.01	0.20
	Avg. Maxima	0.56	0.62	0.67	0.61	0.72	0.75
	Weitzman	0.64	0.85	0.90	0.64	0.74	0.86
Dy-	Dist. Entropy	0.94	0.99	0.98	0.93	0.94	0.96
Time	Summing	0.96	0.96	0.95	0.89	0.89	0.90
Warping	Averaging	0.02	−0.05	0.00	0.16	0.01	0.19
	Avg. Maxima	0.26	0.57	0.56	0.54	0.65	0.72
	Weitzman	0.70	0.90	0.86	0.65	0.72	0.85
norm.	Dist. Entropy	0.99	0.99	0.98	0.96	0.94	0.96
Relative	Summing	0.96	0.96	0.95	0.89	0.89	0.90
Angle	Averaging	0.08	0.14	0.03	0.12	0.32	0.04
Dist.	Avg. Maxima	0.41	0.74	0.69	0.70	0.84	0.75
	Weitzman	0.92	0.97	0.97	0.88	0.76	0.84
Area	Dist. Entropy	0.95	0.99	0.98	0.94	0.94	0.96
Between	Summing	0.96	0.96	0.95	0.89	0.89	0.90
Curves	Averaging	0.05	−0.08	0.06	0.16	0.06	0.18
	Avg. Maxima	0.33	0.55	0.63	0.53	0.65	0.71
	Weitzman	0.79	0.89	0.90	0.72	0.71	0.85
Com-	Dist. Entropy	0.99	0.99	0.99	0.96	0.94	0.96
Vectors	Summing	0.95	0.96	0.95	0.88	0.89	0.90
	Averaging	0.05	−0.13	−0.08	−0.01	−0.22	−0.10
	Avg. Maxima	0.44	0.63	0.57	0.62	0.45	0.59
	Weitzman	0.90	0.82	0.81	0.86	0.66	0.69
Iterative	Dist. Entropy	0.96	0.99	0.98	0.95	0.94	0.95
Leven-	Summing	0.96	0.96	0.94	0.89	0.89	0.90
	Averaging	-0.01	-0.05	0.03	0.03	0.08	0.18
	Avg. Maxima	0.14	0.64	0.57	0.20	0.75	0.63

Table 11 (continued)

		BeamNG.AI$_S$	BeamNG.AI$_M$	BeamNG.AI$_L$	Dave2$_S$	Dave2$_M$	Dave2$_L$
	Weitzman	0.82	0.98	0.93	0.80	0.75	0.87
Man-	Dist. Entropy	0.94	0.83	0.97	0.90	0.89	0.90
Distance	Summing	0.96	0.96	0.95	0.87	0.90	0.90
	Averaging	0.11	−0.09	−0.21	−0.02	0.06	0.00
	Avg. Maxima	0.55	0.44	0.56	0.55	0.66	0.57
	Weitzman	0.81	0.63	0.88	0.75	0.82	0.74
Jaccard	Dist. Entropy	0.90	0.99	0.99	0.95	0.94	0.95
Similar-	Summing	0.96	0.96	0.95	0.89	0.89	0.90
(coarse)	Averaging	0.03	−0.14	−0.01	0.02	0.31	−0.12
	Avg. Maxima	—	—	0.13	—	—	—
	Weitzman	0.77	0.98	0.98	0.81	0.74	0.87
Jaccard	Dist. Entropy	0.98	0.99	0.98	0.94	0.94	0.95
Similar-	Summing	0.96	0.96	0.95	0.89	0.89	0.90
(fine)	Averaging	−0.02	−0.21	0.13	−0.10	−0.04	−0.18
	Avg. Maxima	—	—	—	—	—	—
	Weitzman	0.92	0.98	0.99	0.83	0.72	0.85
Direct	Feature Map	0.97	0.95	0.99	0.92	0.94	0.96
DMs	Test Set Diameter CS	0.99	0.97	0.97	0.95	0.94	0.96
	Convex Hull	0.49	0.78	0.83	0.64	0.85	0.88

experiments were run by S. Klikovits. The first draft of the manuscript was written by all authors. S. Klikovits performed data analysis and data visualisation (tables, figures), and wrote the evaluation section; P. Arcaini provided feedback on the evaluation section. All authors read and approved the final manuscript.

Funding Open access funding provided by Johannes Kepler University Linz. Ahmet Cetinkaya is supported by JSPS KAKENHI Grant No. JP23K03913 and JP26K07561. Paolo Arcaini is supported by the ASPIRE grant No. JPMJAP2301, JST.

Data and Code Availability Evidently, the results of our analysis are specific to the implementation, the evaluation environment and data set used. In the spirit of open science, we make our analysis code publicly available, so other users and researchers can (a) reproduce the results of our study, (b) validate the algorithms and implementation of our DMs, and (c) extend this study by applying our correlation analyses to other diversity measures, behavioural metrics and data.− The source code of our analyses can be viewed at https://github.com/stklik/2026-EMSE-Road-Diversity. − The data sets are available on Zenodo: https://doi.org/10.5281/zenodo.17555435 − An online website providing additional plots, figures and interactive data exploration is available at https://stklik.github.io/2026-EMSE-Road-Diversity.

Declarations

Ethical approval Not applicable.

Informed consent Not applicable.

Clinical Trial Number Not applicable.

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Agarwal PK, Avraham RB, Kaplan H, Sharir M (2014) Computing the discrete Fréchet distance in subquadratic time. SIAM J Comput 43(2):429–449

Alt H, Godau M (1995) Computing the Fréchet distance between two polygonal curves. Int J Comput Geometry Appl 5:75–91. https://doi.org/10.1142/S0218195995000064

Arcaini P, Cetinkaya A (2026) DETOUR: a tool for regression testing of autonomous driving systems. Sci Comput Program 253:103490. https://doi.org/10.1016/j.scico.2026.103490

Aronov B, Har-Peled S, Knauer C, Wang Y, Wenk C (2006) Fréchet distance for curves, revisited. European symposium on algorithms, pp 52–63

Bachman E (2025) Tesla deaths. Regularly updated at tesladeaths.com; version hosted on Zenodo will be updated periodically. https://doi.org/10.5281/zenodo.17810101

Ben Abdessalem R, Nejati S, Briand LC, Stifter T (2018) Testing vision-based control systems using learnable evolutionary algorithms. In: Proceedings of the 40th international conference on software engineering, Association for Computing Machinery, New York, NY, USA, ICSE '18, pp 1016–1026. https://doi.org/10.1145/3180155.3180160,

Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. KDD Workshop Seattle WA USA 10:359–370

Birchler C, Khatiri S, Derakhshanfar P, Panichella S, Panichella A (2023a) Single and multi-objective test cases prioritization for self-driving cars in virtual environments. ACM Trans Softw Eng Methodol 32(2). https://doi.org/10.1145/3533818

Birchler C, Ganz N, Khatiri S, Gambi A, Panichella S (2023b) Cost-effective simulation-based test selection in self-driving cars software. Sci Comput Program 226(C). https://doi.org/10.1016/j.scico.2023.102926,

Birchler C, Khatiri S, Bosshard B, Gambi A, Panichella S (2023c) Machine learning-based test selection for simulation-based testing of self-driving cars software. Empirical Softw Engg 28(3). https://doi.org/10.1007/s10664-023-10286-y,

Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J, Zieba K (2016) End to end learning for self-driving cars. arxiv:1604.07316

Bueno PM, Jino M, Wong WE (2014) Diversity oriented test data generation using metaheuristic search techniques. Inf Sci 259:490–509. https://doi.org/10.1016/j.ins.2011.01.025

Calò A, Arcaini P, Ali S, Hauer F, Ishikawa F (2020a) Generating avoidable collision scenarios for testing autonomous driving systems. In: IEEE 13th int conf on software testing, validation and verification (ICST), pp 375–386. https://doi.org/10.1109/ICST46399.2020.00045

Calò A, Arcaini P, Ali S, Hauer F, Ishikawa F (2020b) Simultaneously searching and solving multiple avoidable collisions for testing autonomous driving systems. In: Proceedings of the 2020 genetic and evolutionary computation conference, Association for Computing Machinery, New York, NY, USA, GECCO '20, pp 1055–1063. https://doi.org/10.1145/3377930.3389827,

Castellano E, Cetinkaya A, Arcaini P (2021) Analysis of road representations in search-based testing of autonomous driving systems. In: 2021 IEEE 21st int conf on software quality, reliability and security (QRS), pp 167–178. https://doi.org/10.1109/QRS54544.2021.00028

Castellano E, Klikovits S, Cetinkaya A, Arcaini P (2022) FreneticV at the SBST 2022 tool competition. In: 2022 IEEE/ACM 15th international workshop on search-based software testing (SBST), pp 47–48. https://doi.org/10.1145/3526072.3527532

Chen TY, Leung H, Mak I (2004) Adaptive random testing. In: Advances in computer science - ASIAN 2004, higher-level decision making, 9th Asian Computing Science Conference, Dedicated to Jean-Louis Lassez on the Occasion of His 5th Cycle Birthday, Chiang Mai, Thailand, December 8-10, 2004, Proceedings, Springer, Lecture Notes in Computer Science, pp 320–329. https://doi.org/10.1007/978-3-540-30502-6_23

Cheng M, Zhou Y, Xie X (2023) BehAVExplor: behavior diversity guided testing for autonomous driving systems. In: Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis, Association for Computing Machinery, New York, NY, USA, ISSTA 2023, pp 488–500. https://doi.org/10.1145/3597926.3598072,

Chiu CH, Chao A (2014) Distance-based functional diversity measures and their decomposition: a framework based on hill numbers. PLoS ONE 9(7):e100014

DeBenedictis PA (1973) On the correlations between certain diversity indices. Am Nat 107(954):295–302

DMV California (2022) Autonomous vehicle collision reports. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports, last accessed: 18 Apr 2026

Dryden IL, Mardia KV (2016) Statistical shape analysis: with applications in R. John Wiley & Sons

Efrat A, Fan Q, Venkatasubramanian S (2007) Curve matching, time warping, and light fields: new algorithms for computing similarity between curves. J Math Imaging Vis 27(3):203–216

Elgendy IT, Hierons RM, McMinn P (2025) A systematic mapping study of the metrics, uses and subjects of diversity-based testing techniques. Softw Test Verif Reliab 35(2):e1914. https://doi.org/10.1002/stvr.1914

Fan C, Luo J, Zhu B (2010) Fréchet-distance on road networks. In: Akiyama J, Jiang B, Kano M, Tan X (eds) Computational geometry, graphs and applications - 9th international conference, CGGA 2010, Dalian, China, November 3-6, 2010, Revised Selected Papers, Springer, Lecture Notes in Computer Science, pp 61–72. https://doi.org/10.1007/978-3-642-24983-9_7,

Feldt R, Poulding S (2017) Searching for test data with feature diversity. arxiv:1709.06017

Feldt R, Poulding S, Clark D, Yoo S (2016) Test set diameter: quantifying the diversity of sets of test cases. In: 2016 IEEE international conference on software testing, verification and validation (ICST), pp 223–233. https://doi.org/10.1109/ICST.2016.33

Fraser G, Arcuri A (2011) EvoSuite: automatic test suite generation for object-oriented software. In: Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on foundations of software engineering, Association for Computing Machinery, New York, NY, USA, ESEC/FSE '11, pp 416–419. https://doi.org/10.1145/2025113.2025179

Fraser G, Rojas JM (2019) Software testing, Springer International Publishing, Cham, pp 123–192. https://doi.org/10.1007/978-3-030-00262-6_4,

Gailly J, Adler M (2026) Zlib: a massively spiffy yet delicately unobtrusive compression library. https://zlib.net, last accessed: 18 Apr 2026

Gambi A, Mueller M, Fraser G (2019) Automatically testing self-driving cars with search-based procedural content generation. In: Proceedings of the 28th ACM SIGSOFT international symposium on software testing and analysis, Association for Computing Machinery, New York, NY, USA, ISSTA 2019, pp 318–328. https://doi.org/10.1145/3293882.3330566,

Gambi A, Jahangirova G, Riccio V, Zampetti F (2022) SBST tool competition 2022. In: 15th IEEE/ACM international workshop on search-based software testing, SBST 2022, Pittsburgh, PA, USA, May 9, 2022

Huai Y, Chen Y, Almanee S, Ngo T, Liao X, Wan Z, Chen QA, Garcia J (2023) Doppelgänger test generation for revealing bugs in autonomous driving software. In: 2023 IEEE/ACM 45th international conference on software engineering (ICSE), pp 2591–2603. https://doi.org/10.1109/ICSE48619.2023.00216

Humeniuk D, Khomh F (2025) Representation improvement in latent space for search-based testing of autonomous robotic systems. https://doi.org/10.48550/ARXIV.2503.20642

Jahangirova G, Stocco A, Tonella P (2021) Quality metrics and oracles for autonomous vehicles testing. In: 2021 14th IEEE conference on software testing, verification and validation (ICST), pp 194–204. https://doi.org/10.1109/ICST49551.2021.00030

Jekel CF, Venter G, Venter MP, Stander N, Haftka RT (2019) Similarity measures for identifying material parameters from hysteresis loops using inverse analysis. International Journal of Material Forming. https://doi.org/10.1007/s12289-018-1421-8

Klikovits S, Castellano E, Cetinkaya A, Arcaini P (2023) Frenetic-lib: an extensible framework for search-based generation of road structures for ADS testing. Sci Comput Program 230:102996. https://doi.org/10.1016/j.scico.2023.102996

Kong Z, Guo J, Li A, Liu C (2020) PhysGAN: generating physical-world-resilient adversarial examples for autonomous driving. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 14242–14251. https://doi.org/10.1109/CVPR42600.2020.01426

Laurent T, Klikovits S, Arcaini P, Ishikawa F, Ventresque A (2023) Parameter coverage for testing of autonomous driving systems under uncertainty. ACM Trans Softw Eng Methodol 32(3). https://doi.org/10.1145/3550270,

Lehman J, Stanley KO (2011) Abandoning objectives: evolution through the search for novelty alone. Evol Comput 19(2):189–223

Leinster T, Cobbold CA (2012) Measuring diversity: the importance of species similarity. Ecology 93(3):477–489. https://doi.org/10.1890/10-2402.1

Levenshtein VI et al (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady Soviet Union 10:707–710

Luo Y, Zhang XY, Arcaini P, Jin Z, Zhao H, Ishikawa F, Wu R, Xie T (2021) Targeting requirements violations of autonomous driving systems by dynamic evolutionary search. In: 2021 36th IEEE/ACM international conference on automated software engineering (ASE), pp 279–291. https://doi.org/10.1109/ASE51524.2021.9678883

Majumdar R, Mathur A, Pirron M, Stegner L, Zufferey D (2021) Paracosm: a test framework for autonomous driving simulations. In: Guerra E, Stoelinga M (eds) Fundamental approaches to software engineering. Springer International Publishing, Cham, pp 172–195

McNaughton M, Urmson C, Dolan JM, Lee JW (2011) Motion planning for autonomous driving with a conformal spatiotemporal lattice. In: 2011 IEEE int conf on robotics and automation, IEEE, pp 4889–4895

Mosig A, Clausen M (2005) Approximately matching polygonal curves with respect to the Fréchet distance. Comput Geom 30(2):113–127

Mouchet MA, Villéger S, Mason NW, Mouillot D (2010) Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. Funct Ecol 24(4):867–876

Munich ME, Perona P (1999) Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In: Proceedings of the seventh IEEE international conference on computer vision, vol 1, pp 108–115. https://doi.org/10.1109/ICCV.1999.791205

Nguyen V, Huber S, Gambi A (2021) SALVO: automated generation of diversified tests for self-driving cars from existing maps. In: 2021 IEEE international conference on artificial intelligence testing (AITest), pp 128–135. https://doi.org/10.1109/AITEST52744.2021.00033

Riccio V, Tonella P (2020) Model-based exploration of the frontier of behaviours for deep learning system testing. In: Proceedings of the 28th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, Association for Computing Machinery, New York, NY, USA, ESEC/FSE 2020, pp 876–888. https://doi.org/10.1145/3368089.3409730,

Sadat A, Segal S, Casas S, Tu J, Yang B, Urtasun R, Yumer E (2021) Diverse complexity measures for dataset curation in self-driving. In: 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE Press, pp 8609–8616. https://doi.org/10.1109/IROS51168.2021.9636869

Shi Q, Chen Z, Fang C, Feng Y, Xu B (2016) Measuring the diversity of a test set with distance entropy. IEEE Trans Reliab 65(1):19–27. https://doi.org/10.1109/TR.2015.2434953

Tang Y, Zhou Y, Wu F, Liu Y, Sun J, Huang W, Wang G (2021) Route coverage testing for autonomous vehicles via map modeling. In: 2021 IEEE international conference on robotics and automation (ICRA), IEEE Press, pp 11450–11456. https://doi.org/10.1109/ICRA48506.2021.9560890

Tang S, Zhang Z, Zhang Y, Zhou J, Guo Y, Liu S, Guo S, Li YF, Ma L, Xue Y, Liu Y (2023) A survey on automated driving system testing: landscapes and trends. ACM Trans Softw Eng Methodol 32(5). https://doi.org/10.1145/3579642,

Tuncali CE, Fainekos G, Ito H, Kapinski J (2018) Sim-ATAV: simulation-based adversarial testing framework for autonomous vehicles. In: Proceedings of the 21st international conference on hybrid systems: computation and control (Part of CPS Week), Association for Computing Machinery, New York, NY, USA, HSCC '18, pp 283–284. https://doi.org/10.1145/3178126.3187004,

US Department of Transportation, National Highway Traffic Safety Association (2022) Summary report: Standing general order on crash reporting for automated driving systems (DOT HS 813 324). Tech. rep

Vlachos M, Gunopulos D, Das G (2004) Rotation invariant distance measures for trajectories. In: Proceedings of the Tenth ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery, New York, NY, USA, KDD '04, pp 707–712. https://doi.org/10.1145/1014052.1014144,

Weise J, Mostaghim S (2021) Many-objective pathfinding based on Fréchet similarity metric. In: Evolutionary multi-criterion optimization: 11th international conference, EMO 2021, Shenzhen, China, March 28–31, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, pp 375–386. https://doi.org/10.1007/978-3-030-72062-9_30

Weitzman ML (1992) On diversity. Q J Econ 107(2):363–405

Werling M, Ziegler J, Kammel S, Thrun S (2010) Optimal trajectory generation for dynamic street scenarios in a Frenét frame. In: 2010 IEEE international conference on robotics and automation, pp 987–993. https://doi.org/10.1109/ROBOT.2010.5509799

Weyuker EJ (1986) Axiomatizing software test data adequacy. IEEE Trans Softw Eng 12(12):1128–1138
Witowski K, Stander N (2012) Parameter identification of hysteretic models using partial curve mapping. In: 12th AIAA aviation technology, integration, and operations (ATIO) conference and 14th AIAA/ISSMO multidisciplinary analysis and optimization conference, p 5580
Wohlin C, Runeson P, Hst M, Ohlsson MC, Regnell B, Wessln A (2012) Experimentation in software engineering. Springer Publishing Company, Incorporated
Xie Q, Memon AM (2006) Studying the characteristics of a "good" GUI test suite. In: Proceedings of the 17th international symposium on software reliability engineering, IEEE Computer Society, USA, ISSRE '06, pp 159–168. https://doi.org/10.1109/ISSRE.2006.45,
Zhong Z, Tang Y, Zhou Y, Neves VdO, Liu Y, Ray B (2021) A survey on scenario-based testing for automated driving systems in high-fidelity simulation. arxiv:2112.00964
Zhong Z, Kaiser G, Ray B (2023) Neural network guided evolutionary fuzzing for finding traffic violations of autonomous vehicles. IEEE Trans Softw Eng 49(4):1860–1875. https://doi.org/10.1109/TSE.2022.3195640
Zhou H, Li W, Kong Z, Guo J, Zhang Y, Yu B, Zhang L, Liu C (2020) DeepBillboard: systematic physical-world testing of autonomous driving systems. In: Proceedings of the ACM/IEEE 42nd international conference on software engineering, Association for Computing Machinery, New York, NY, USA, ICSE '20, pp 347–358. https://doi.org/10.1145/3377811.3380422,
Zhu S, Aksun-Guvenc B (2020) Trajectory planning of autonomous vehicles based on parameterized control optimization in dynamic on-road environments. J Intel Robot Syst 100(3):1055–1067
Zhu B, Zhang P, Zhao J, Deng W (2022) Hazardous scenario enhanced generation for automated vehicle testing based on optimization searching method. Trans Intell Transport Sys 23(7):7321–7331. https://doi.org/10.1109/TITS.2021.3068784
Zohdinasab T, Riccio V, Gambi A, Tonella P (2021) DeepHyperion: exploring the feature space of deep learning-based systems through illumination search. In: Proceedings of the 30th ACM SIGSOFT international symposium on software testing and analysis, Association for Computing Machinery, New York, NY, USA, ISSTA 2021, pp 79–90. https://doi.org/10.1145/3460319.3464811

Authors and Affiliations

Stefan Klikovits[1] · Vincenzo Riccio[2] · Ezequiel Castellano[3] · Ahmet Cetinkaya[4] · Alessio Gambi[5] · Paolo Arcaini[3]

✉ Stefan Klikovits
stefan.klikovits@jku.at

Vincenzo Riccio
vincenzo.riccio@uniud.it

Ahmet Cetinkaya
ahmet@shibaura-it.ac.jp

Alessio Gambi
alessio.gambi@ait.ac.at

Paolo Arcaini
arcaini@nii.ac.jp

[1] Johannes Kepler University Linz, Linz, Austria

[2] University of Udine, Udine, Italy

[3] National Institute of Informatics, Tokyo, Japan

[4] Shibaura Institute of Technology, Tokyo, Japan

[5] Austrian Institute of Technology, Vienna, Austria