Impact, Attention, Influence:  
Early Assessment of Autonomous Driving Datasets

Daniel Bogdoll†‡*, Jonas Hendl†*, Felix Schreyer†, Nishanth Gowda†, Michael Färber† and J. Marius Zöllner†‡  
†FZI Research Center for Information Technology, Germany  
bogdoll@fzi.de  
‡Karlsruhe Institute of Technology, Germany

Abstract—Autonomous Driving (AD), the area of robotics with the greatest potential impact on society, has gained a lot of momentum in the last decade. As a result of this, the number of datasets in AD has increased rapidly. Creators and users of datasets can benefit from a better understanding of developments in the field. While scientometric analysis has been conducted in other fields, it rarely revolves around datasets. Thus, the impact, attention, and influence of datasets on autonomous driving remains a rarely investigated field. In this work, we provide a scientometric analysis for over 200 datasets in AD. We perform a rigorous evaluation of relations between available metadata and citation counts based on linear regression. Subsequently, we propose an Influence Score to assess a dataset already early on without the need for a track-record of citations, which is only available with a certain delay.

Index Terms—Robotics, Autonomous Driving, Datasets, Influence, Impact, Attention, Scientometrics, Bibliometrics

I. INTRODUCTION

Autonomous driving technology does not only affect urban transportation [1] and delivery of goods [2], but also farming [3] or warehouse logistics [4]. With the progress of deep learning and this growing interest in AD in many fields of robotic, the number of related datasets is consistently increasing. The datasets have also increased in size and many have become increasingly specialized [5]. The most extensive collection of datasets known to us, ad-datasets, currently lists 231 datasets in the domain [6]. However, not all of them are being equally used in the robotics community, the distribution of their citations is heavily skewed. As part of the more impactful works, well known datasets for the core tasks perception and prediction dominate [7]–[9]. As part of the long tail, many datasets for niche research areas exist [10]–[12]. Well known datasets tend to bring many advantages with them: They enable comparison between works, have higher quality, advanced tooling, and often community knowledge and support is available. The increasing number of datasets, which are potentially interesting but lack reputation, leads to a lot of untapped potential: Many researchers are hesitant to use such datasets and stick to old, but established ones instead [6]. This is why we asked ourselves the question: Given a novel dataset without a multi-year track record of citations, is there a way to estimate its future development? Datasets with a high potential might be more appealing already in their early days.

* These authors contributed equally

Research Gap. To date, citations are mostly used to assess datasets, which are not available early on. Thus, new datasets can have a hard time gaining traction, which results in untapped potential. It is yet not well understood if and how metadata of datasets relate to future impact or how they can be utilized to assess datasets early on. To the best of our knowledge, such an analysis has not yet been performed.

Contribution. In order to analyze the field of datasets, we first assembled the largest collection of datasets with enriched metadata available, including over 200 datasets with metadata from three different sources. We then applied linear regression to evaluate factors which relate to the future impact of datasets, measured in citations. Finally, we propose the Influence Score (IS), which is a mean to assess datasets early on without the need of a multi-year track record of citations. The IS can be used to assess any dataset at any given year, which also allows for later analysis. Our work aims to help researchers from the robotics community to better understand and assess the performance of datasets. This can lead to the design of better and thus more influential datasets as well as an actionable analysis of new datasets to assess their potential. All data used in this work is as of January 04, 2023. All code is available on GitHub.
II. RELATED WORK

Here, we give an introduction to scientometrics, bibliometrics, and altmetrics, followed by dataset analysis.

A. Scientometrics, Bibliometrics, and Altmetrics

Scientometrics, Bibliometrics, and Altmetrics are highly intertwined fields that focus on the analysis of science and its processes as a whole, written works of science, and online communication of science, respectively [14].

Scientometrics. Ravenscroft et al. [15] examined the impact of research by comparing citation-based metrics, such as citation count or h-index [16], with altmetrics and impact other than citations, e.g., societal and economic impact. However, they found no strong relationship between the fields. Hicks et al. [17] suggest using multiple factors to portray multiple aspects.

Bibliometrics. Citations can be aggregated on different levels, e.g., for the papers of one author as the h-index does, or on the journal level, like the journal impact factor (JIF), which is the two-year average ratio of citations to articles published. The JIF is ill-suited for evaluating individual papers by means of the journal it was published in [30]. The h-index is widely used to evaluate scientific research. It combines the productivity of an author with the impact of their individual papers. Using the h-index increases robustness compared to simply counting the total number of citations, as few highly cited papers have little effect on the h-index. In addition, there have been efforts to recommend papers and citations [32], [33], predict future citation counts of papers [34]–[37] and the impact of scientists [38]. Such approaches remain challenging and are often domain-specific.

Bornmann and Marx [39] have proposed to expand the bibliometric analysis by not only considering citations but also references. Following this idea, reference analysis has been used to identify influential references [40].

Altmetrics. Online interactions with papers are referred to as altmetrics and are usually available earlier than citations, which gives altmetrics an advantage over bibliometrics [41]. Bornmann and Marx [42] examined if Altmetrics can be used to predict paper quality which was measured through peer assessments and found that both tweets and readers do, with the latter having a stronger relationship. Lamb et al. [43] showed that the Altmetric Attention Score is a predictor of the citations of a paper in ecology and conservation. Zavrel et al. [44] clustered papers released at the International Conference on Machine Learning (ICML) in 2022 and calculated a score for their impact. They used Twitter mentions, citations, the authors’ average h-index, and an award for outstanding papers rewarded by the conference itself. They claim to do a “simple combination of these four scores to calculate an impact score” [44] but do not reveal the formula. Färber analyzed GitHub repositories of papers, mostly from the field of AI, and found a power-law distribution of stars and forks [45]. While Haustein et al. claim that “Altmetrics measures scientific impact based on online references and activity” [46], many disagree with equating altmetrics with impact. For example, Sugimoto states that ”attention is not impact” and calls online interaction with scientific works “attention” [47]. Altmetrics might reflect broader or societal impact [41].

B. Dataset Analysis

Bogdoll et al. [5] gathered metadata of over 200 datasets in the field of autonomous driving. Similarly, Färber and Lamprecht released the data set knowledge graph, which is a collection of over 2,000 datasets with added metadata [48]. D’Ulizia et al. [49] analyzed the metadata of datasets for fake news detection. Utamachant and Anutariya [50] analyzed the datasets of Thailand’s national open data portal, but relied on domain experts to assess impact. Nguyen and Weller proposed FAIRnets, a service to search for neural networks and their related datasets [51] published on GitHub. They build upon the Findable, Accessible, Interoperable, Reusable (FAIR) principles [52], which "put specific emphasis on enhancing the ability of machines to automatically find and use the data” [52]. Khan et al. [53] analyzed datasets from the Global Biodiversity Information Facility (GBIF) which publishes datasets with a DOI and indexes datasets in biodiversity. They promote data standards and the reuse of datasets [54] as well as accompanying publications, which they call "data papers”, that describe a dataset thoroughly [55]. Khan et al. [53] report a strong correlation between dataset download numbers and citation counts, and suggest that downloads and citations signify a similar kind of impact. They also find correlations between altmetrics and citations. Moreover, they question whether every citation means the usage of a dataset and point to differences in citation behavior. Färber et al. proposed an approach to find methods and datasets which authors actually used when citing the related paper [56]. However, unrealistically few dataset usages were identified.
III. REGRESSION ANALYSIS

Here, we first introduce our taxonomy of terms related to the assessment of datasets. Subsequently, we introduce our data sources. Based on these, we describe the regression analysis of citations and metadata. In Section IV, we present the resulting Influence Score. Figure 2 gives an overview over this process.

A. Taxonomy

As became clear in Section II, no common language for specific aspects in the domain has evolved yet. Thus, we introduce a taxonomy to clearly describe different aspects with respect to the development of a dataset or paper. As general terms for this, we utilize success, progress, performance, or potential. For concrete aspects, we establish the following terms, where each one can be applied to any single paper:

Impact: We use the number of citations to measure the scientific impact of a paper, which is common in Scientometrics, but not without criticism [19].

Attention: The online reception, such as tweets and Wikipedia articles mentioning a paper, represents the attention by researchers and the public.

Influence: We refer to the resulting score of our proposed method, which combines a multitude of aspects, as the influence, or IS, of a dataset. We deem this term appropriate for any method that goes beyond purely impact-based assessment.

B. Data Sources and Selection

We used three sources for our data: ad-datasets.com [6], the Semantic Scholar Academic Graph API [13], and altmetric.com [57]. Based on the DOI and arXiv-Id from ad-datasets, we automatically extracted the metadata of papers from Semantic Scholar and altmetric.com. Based on these papers, all of which describe datasets, we performed data exploration, regression, and the computation of the IS.

AD-Datasets: This web tool offers an overview of over 200 data sets in AD [5]. It includes a detailed breakdown of most dataset entries by 20 different meta categories, provided by the authors and the research community. This way, relations between datasets, accompanying papers and further metadata are available. The underlying data is stored in the JSON format and can be accessed accordingly. In this work, we utilize the $n_{frames}$ and $n_{sensors}$ metadata, which indicate the size of datasets in different dimensions, which is a potential aspect of the relevance of a dataset.

Altmetric: We used the API by altmetric.com [57], which provides insight into online attention and readership. These properties are provided by the following categories:

Attention Score: The $aa_{curr}$ aggregates different sources into a single score [58]. It is a weighted count of different online sources. For example, the weight for a reference on Wikipedia is 3, while Twitter and Reddit mentions are both weighted with 0.25. Unfortunately, the history of this score is only provided for the most recent year.

Attention Score after three months: The $aa_{3m}$ is the percentile of the papers’ Attention Score three months after publication. The percentile is calculated in comparison to papers that have been released at a similar time.

Readers: The number of people that have saved a paper in their reference management software. Reading a paper is less significant than citing it, but the number of readers might imply interest in a paper early on. The number of readers is provided individually for multiple reference management services, which we sum into a single count for online readers. Altmetric.com cannot verify the number of readers, thus, it is not included in the attention score. This is a relevant attribute, as it decouples the metrics. However, there are no historic data available.

Semantic Scholar: For every accompanying paper of a dataset, we pulled data from Semantic Scholar. Sometimes, multiple datasets are described in the same paper, which will lead to the same information for those datasets. We extracted the following nested data:

- List of referenced papers, including for each a list of all citing papers and the year of citation.
- List of authors and their respective publications, including for each publication a list of citing papers and the year of citation.
- List of citing papers, including for each a list of citing papers and the year of citation.

Wherever possible, we collected associated timestamps, including the publication year $a_{pub}$ of each paper. The first
two categories, while dynamic, are directly available. We use the citations of references as a measure of the impact of references. Having impactful references might indicate that a paper is covering popular topics within AD or that the authors are knowledgeable in the field.

The performance of authors can be estimated by evaluating their paper count and how many citations they have received, which becomes only meaningful over time. As discussed earlier, not every citation means usage of a dataset. While it would have been interesting to take into account, in which section a paper has been cited, this data was not available for most papers. Based on the $n_{cit3}$ citations from the previous three years, citations of works that cited a dataset signify the value created by working with the dataset, which is why we included those.

A critical aspect of the collected data is that oftentimes, no historic information was available. Also, oftentimes, data was not available due to limitations, e.g., Altmetric is incompatible with DOIs from IEEE publications, which are common in the fields of robotics, autonomous driving, and machine learning. Similarly, Ravenscroft et al. [15] expressed concerns about Altmetric, as they were unable to find 40 % of the papers they analyzed, all from the field of computer science.

C. Data Aggregation

To further utilize the raw data we collected, we aggregated some of it with the aim to assemble a finite list of features that describe a dataset. We aggregated some of our data sources using the concept of the h-index formula, as it is widely known, transparent, and easy to reproduce. In order to analyze smaller timespans, we deviated from the typical 5-year duration and calculated multiple 3-year indexes ourselves.

For authors, we applied the h3-index for each individual. We then aggregated the h-indices of all authors of a paper via the arithmetic mean in $aut_{ph3}$. Respectively, we applied the h-index formula to references and citations. For the references of a paper, the $ref_{h3}$ is calculated identically to the way it is utilized for authors. Just like an author has papers with citations, a paper has references with citations. A high h-index for references would signify that several of the referenced papers gained lots of attraction. We also applied the h3-index formula to the citations and their citations to get the h3-index of citations $cit_{h3}$, following Schubert et al. [59]. The final list of all extracted and calculated features can be found in Table I.

D. Cluster Analysis and Regression Setup

We evaluated our computed features with respect to their ability to predict future citations. Therefore, we performed linear regression. For this, we first computed clusters of the datasets to determine a meaningful time horizon. Subsequently, we defined our regression setup.

Cluster Analysis: To show that there are meaningful variations between clusters, we looked at the impact of papers for up to 2 years after publication in a journal or conference proceedings. As visualized in Fig. 3, clear clusters are visible, where line-thickness indicates cluster size. For k-means clustering, we used six clusters based on the elbow plot, which shows which additional cluster provides a non-marginal reduction of the total variation within clusters. The growth of citation counts behaves exponentially for the top performing works. A clear differentiation between all clusters becomes apparent already after one year, which we chose as the time horizon for the regression. This allowed us to include more recent papers, which would have been excluded otherwise due to their missing track record of citations.

Regression Setup: As independent variables, we included the features $n_{sensors}$, $a_{pub}$, $ref_{h3}$, $aut_{ph3}$, $n_{cit3}$, and $aas_{3m}$, as shown in Table I, in order to estimate the citation count after one year. Preliminary data exploration suggested a curvilinear relationship between $aas_{3m}$ and the number of citations. Therefore, a quadratic term was added. All predictors were standardized by subtracting the mean and dividing by the standard deviation prior to the analysis. The feature $n_{cit3}$ was log(x+1)-transformed to ensure a normal distribution of the residuals, which are the error terms of the regression.

For the regression, we were able to utilize 111 datasets, as values for all included features were available, and they had been released at least one year prior. Residuals and collinearity, the ability to linearly predict one independent variable with other independent variables, were checked. The collinearity was quantified through the variance inflation factor of each regressor which all were lower than three. We performed the Breusch-Pagan and White test for heteroskedasticity, which is the inconsistency of the variance of residuals at different levels of the dependent variable. Both tests indicated that we do not have sufficient evidence for the presence of heteroskedasticity. Still, robust standard errors were used to ensure the standard errors are calculated correctly in the presence of heteroskedasticity which at worst leads to standard errors being estimated larger.

![Fig. 3. Development of the number of citations for publication-clusters over a dynamic 3-year window. Papers are clustered based on similar performance. Semantic Scholar also tracks citations of pre-prints, which leads to citations prior to the publication date of the final work.](image-url)
We chose not to include \( n_{frames} \) for the regression because numerous of the datasets did not contain this meta-information. However, we examined a model in which the feature was included, which did not lead to new findings.

E. Regression Analysis

With the explained regression setup, we were now interested in finding statistically significant predictor variables for the citation count at the end of the year after publication. The \( aas_{3m} \) and \( aas_{2m} \) were positively related to the number of citations and both relationships were significant at <0.0001. Both coefficients were positive. All other features were not significantly related to the number of citations. The results are reported in Table II.

Since only one feature showed a relationship with the number of citations, we do not consider a stable prediction of citations possible with the available data. In order to still perform an early evaluation of datasets, in the following we present our Influence Score (IS).

IV. INFLUENCE SCORE

We propose the Influence Score (IS), which includes a variety of features that are available early on. These are weighted dynamically in order to receive an indication of the relative performance of any given dataset at any given time. The calculation compares each dataset with all existing ones from the domain, so that relative differences and trends are immediately recognizable.

Percentiles are used to allow relative scoring within the surrounding group of datasets. The datasets roughly follow a normal distribution in their IS scores. As shown in Table I, we utilize eight different features for the IS: \( n_{frames}, n_{sensors}, ref_{h3}, aut_{h3}, n_{cit3}, cit_{h3}, aas_{curr} \) and \( n_{readers} \). This way, we consider more than just the citations, but do not exclude them: If early citations are already available, they become a meaningful part of the score, as the relation to other datasets of the peer group is relevant. This way, citation velocity is included. The IS is defined as follows:

\[
IS(paper) = 1/n * \sum_{i=0}^{n} \text{percentile}(feature_i)
\]

where:
\( i = \text{Feature Index} \)
\( n = \text{Number of available features} \)

Only features, which are available, are dynamically included in the IS. As we used percentiles of each feature to facilitate the understanding of the features, common issues are mitigated. E.g., typical feature values change over time: For example, with the growth of AD, the \( n_{cit} \) value of a paper today is likely higher than a decade ago, which becomes clearly visible in Figure 1. Furthermore, commonly observed values for features might differ between different fields. This helps people who are not familiar with AD or the features to easily assess if the score a dataset achieved is high or low.

A. Qualitative Demonstration

To showcase the IS, we compare exemplary the development of the five most and least cited papers with a latest publication in 2019, by their IS and visualize the results in Fig. 4. It becomes clearly visible, that the two groups are easily distinguishable by their IS, but also that differences within the groups are visible.

The individual features show different pictures: For \( ref_{h3} \), also papers with only a few citations can have meaningful references in their works. \( n_{cit3} \) and \( cit_{h3} \) only confirm what was known by our data selection, as we selected the datasets by citation count. \( aut_{h3} \) shows, how successful datasets can also boost personal careers, as some authors became professors and remained active in their field. \( n_{sensors} \) and \( n_{frames} \) show
of datasets, given a flood of new datasets published every year. We analyzed impact measured by citations and evaluated relations of metadata and features which we extracted from multiple online sources. Our regression analysis showed no strong relation between future citations and our selected features. Subsequently, we presented our developed Influence Score (IS). This score utilizes a set of eight features to assess any dataset also early on. This is based on an analysis within the peer group of all datasets, which allows for the early detection of relative trends.

Our work contributes to a better understanding of datasets, which enables researchers to find and assess published datasets in the domain of autonomous driving without the need of waiting for a track record of citations.

**Limitations and Outlook:** For our work, we evaluated the paper accompanying the dataset assuming that the paper is a good representation of the dataset. When measuring scientific impact through citations, we think this holds because the paper is actually the cited scientific work. However, not every citation might be meaningful, positive, or indicate the usage of a dataset. Ideally, large language models could evaluate if a dataset is actually used, if cited. Khan et al. [53], who analyzed datasets in biodiversity, suggested that the correlation between the number of downloads and citations signifies that these two measures are comparable representations of impact. However, in the domain of AD, download numbers are typically not available, but this might change. As some datasets are presented in the same paper, a further decoupling of accompanying papers and the respective datasets would be helpful. We found, that the quality and availability of metadata in AD provided by the creators of datasets varies strongly. Thus, standards should be established [90]. While we focussed on dataset and paper specific features for this work, we are also interested in the venue or journal of publication, which can be considered as an additional feature in future work.

**V. CONCLUSION**

In this paper, we addressed the lack of knowledge with respect to the scientific impact, attention, and influence of datasets in robotics. Our focus was on an early assessment rather static results, with a trend towards larger datasets being more successful.

**B. Quantitative Demonstration**

In order to show the quantitative performance of the IS, we showcase all datasets released in 2022 in a detailed overview in Table III. Such a pre-filtering process is useful in order to explore novel datasets. Here, it becomes clear that the IS captures a wide variety of different aspects of a dataset. Of particular interest is the fact that even low-performing datasets can lead in certain features. Thus, if a researcher is interested in certain aspects of a dataset, they can simply focus on the features they are interested in and omit the others, which enables less-known datasets to be discovered and used. Figure 5 shows an overview of the IS distributions.

![Figure 5](image.png)

*Fig. 5. Distribution of the Influence Score (IS) of all datasets from 2022.*
VI. ACKNOWLEDGMENT

This work results partly from the KIGLIS project supported by the German Federal Ministry of Education and Research (BMBF), grant number 16KIS1231. We want to thank both Altmetric and Semantic Scholar, who have provided us with the necessary API accesses for this work.

REFERENCES


