

Creating a Dataset of Spatial Parameters of Ground-Mounted Photovoltaic Systems Utilising Orthophotos and the Segment Anything Model

Johannes Albert, Chantal Schymik, Philipp Gärtner, Claudius Wehner, Jan Siegismund, Stephan Klingner

Application Lab for Artificial Intelligence and Big Data at the German Environment Agency, Leipzig, Germany -
(johannes.albert, chantal.schymik, philipp.gaertner, claudius.wehner, jan.siegismund, stephan.klingner)@uba.de

Keywords: Ground-Mounted Photovoltaic, Orthophotos, Segment Anything Model, Row Spacing, Ground Coverage Ratio.

Abstract

The rapid expansion of renewable energy sources poses significant challenges in reconciling energy development with competing interests. This underscores the necessity for precise spatial data to facilitate effective balancing, management, or evaluation of compliance with regulatory frameworks. This paper presents a zero-shot approach for extracting parameters of ground-mounted photovoltaic systems in Germany based on digital orthophotos. This allows for the accurate identification and delineation of essential spatial parameters, including the ground coverage ratio of photovoltaic modules, the row spacing between module rows, and their exact orientation. The results of this study are twofold. First, the developed technical pipeline successfully achieves high-quality segmentation of photovoltaic module rows, with over 71 % of the results demonstrating satisfactory to flawless segmentation. Second, the resulting dataset is made available for further analysis and can serve as a starting point for the development of additional AI models aimed at monitoring the dynamics of photovoltaic systems.

1. Introduction

The energy transition represents a global challenge that requires to navigate contradictions and complexities at multiple levels. Competing interests in land use, nature conservation, environmental protection, and landscape preservation are just a few of the areas where an alignment of conflicting interests is essential (Hilker et al., 2024). However, it is not only dichotomies that exist – synergies are also possible, e.g. regarding biodiversity (Bai et al., 2022; Carvalho et al., 2024), the realisation and maximisation of which require effective planning. The complexity of these issues is further compounded by the need to integrate new technical requirements for grid infrastructure to accommodate the evolving energy generation landscape (Schmietendorf et al., 2017; Smith et al., 2022).

Adding to the complexity, the mainly decentralized structure of renewable energy generation facilities makes the coordination of their expansion significantly more challenging than with traditional large-scale power plants. The urgency of expanding renewable energy sources often leads to ambitious short-term targets, which can complicate the resolution of these opposing interests. In this context, data emerge as a critical resource for addressing these conflicts and challenges. Comprehensive and regularly updated data is essential for continuous reflection and adjustment when weighing up different interests. Such data serve as a basis for various control mechanisms, such as policy consultation processes or planning procedures, and can thus shape legislation, regulatory requirements or the planning of infrastructure projects such as grid expansion.

This paper examines the specific subject matter in context of the expansion of ground-mounted photovoltaic (GMPV) systems in Germany. As stated in § 4 of the Renewable Energy Sources Act (EEG 2023) (BMJV, 2014), the expansion target for the total photovoltaic capacity in Germany is 400 GW by 2040, which represents a significant increase from around 81.7 GW in 2023 and makes it a key element of the energy transition.

A significant challenge is the lack of precise spatial data on photovoltaic installations, which complicates the assessment of ongoing developments and potentially required readjustment of regulatory frameworks. The Marktstammdatenregister (MaStR) (Bundesnetzagentur, n.d.), as the central freely accessible registry of all energy-related installations provides point location data for GMPV systems, but lacks detailed spatial information, such as the area occupied by photovoltaic module rows, the spacing between these rows, and their exact orientation. The absence of these essential information impedes the balancing of opposing interests or the shaping of legal framework conditions. Existing initiatives such as Global Renewables Watch (Robinson et al., 2025) or SATLAS (Bastani, n.d.), while having a global focus, lack sufficient data granularity to effectively address complex issues.

Thus, the study addresses a gap by aiming to extract detailed spatial parameters for all GMPV systems in Germany using digital orthophotos. It employs a zero-shot segmentation model for image segmentation, followed by a classification process, to achieve precise segmentation and classification of components of GMPV plants and derive various parameters based on this analysis.

The aim of this paper is to outline an approach to extract detailed spatial parameters of GMPV systems without training data, create a dataset comprising these parameters extracted for GMPV systems in Germany, and provide a descriptive analysis of the dataset.

Accordingly, the following research questions (RQ) were addressed in this study:

1. RQ1: How can spatial parameters of ground-mounted photovoltaic systems be extracted without ground truth data available?
2. RQ2: Can these detailed spatial parameters of GMPV parks, such as row spacing and covered area, be extracted with sufficient quality?

2. Data and materials

To address the research questions, it was first necessary to obtain the footprints within which the parks are located, as well as corresponding high-resolution images for the detection of the module rows. The footprints encompassed the parks and were available nationwide. However, they were based on a different understanding of the park area than what is required for the calculation of precise metrics (see also Section 3). Therefore, three main data sources were used:

1. **GMPV footprints:** The open-access dataset was provided by Manske (2025). The dataset contains manually digitised outlines of 8,789 GMPV footprints across Germany as depicted in Figure 1), and serves as the foundational reference. It is based on the MaStR of the Federal Network Agency (Bundesnetzagentur) in Germany as of January 4, 2024. In addition to the spatial information, the dataset contains information on photovoltaic systems commissioning and decommissioning dates, their cardinal direction (fixed orientation or sun-tracked) and location (ground-mounted, floating-mounted or agrivoltaics).
2. **Digital orthophotos (DOP):** Alongside the photovoltaic footprint dataset, DOP imagery was incorporated into the workflow. The DOP are distortion-free and true-to-scale raster images of the earth's surface. They are derived from georeferenced aerial photographs and a digital elevation model. The dataset covers the whole of Germany as a seamless, non-overlapping mosaic, provided by the surveying authorities of the federal states and published by the Federal Agency for Cartography and Geodesy (BKG, 2025). The orthophotos are available as 3-channel true-colour (RGB) images and single-channel near infrared (NIR) images, with a ground resolution of 0.2 m and size of 1000 m x 1000 m). The image acquisition dates vary by federal state, ranging from 2016 to 2023 (see Figure 1).
3. **Photovoltaic parks:** GMPV systems are typically enclosed by fences or other structural or natural elements, which often correspond to polygon features in OpenStreetMap (OSM). To obtain the relevant OSM data, the Overpass API was queried with a set of photovoltaic-specific tags. The resulting dataset consists of 6,479 photovoltaic parks (see red polygon in Figure 2).

3. Methodology

This study aims to extract spatial parameters for GMPV systems in Germany using DOP. For image segmentation, the Segment Anything Model (SAM) (Kirillov et al., 2023), a zero-shot segmentation model, is applied on DOP. The application of SAM followed by a classification process facilitates the precise segmentation of components within GMPV plants. The following sections describe the extraction of module rows, park area and the approach applied for evaluation. Figure 2 gives a graphical overview of used terminology.

Throughout the remainder of this paper, the designation *module rows* will be adopted, as this arrangement is the prevailing method of installation. Nonetheless, it should be acknowledged that other configurations, including non-row-based setups like sun-tracked modules, do exist.

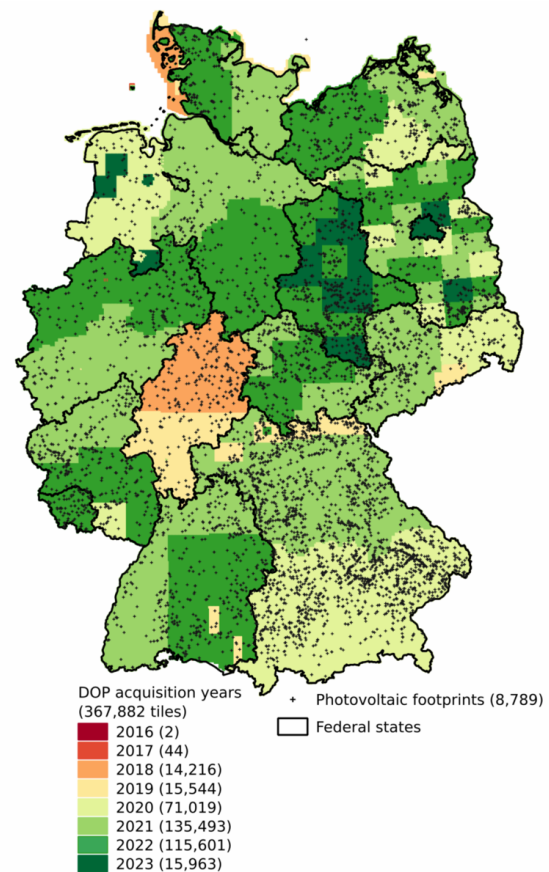


Figure 1. Location of analysed footprints and recency and number of utilised digital orthophoto tiles.

3.1 Photovoltaic module rows

Around every photovoltaic footprint from Manske (2025), a grid of overlapping true-colour DOP image chips was generated, each measuring 640×640 pixels ($128 \text{ m} \times 128 \text{ m}$) with a stride of 320 pixels to ensure 50 % overlap. This grid, comprising over 227,000 image chips, was used as input for Meta's SAM. Since no training data for module row detection were available, SAM was used in zero-shot manner to automatically segment the image chips. The image chips were read using OpenCV (Bradski, 2000) and converted to RGB format. SAM was initialised with the pre-trained ViT-H (Vision Transformer-Huge) backbone with tuned hyperparameters. As SAM is computationally expensive, its execution was performed inside a parallelised array job for batch processing on a high-performance computing cluster with NVIDIA Tesla A30 GPUs.

The federal states were processed individually and the number of GPUs used per state were chosen dependent on the number of image chips per state. For each image, the model generated segmentation masks using the *SamAutomaticMaskGenerator*. These masks were then georeferenced using metadata extracted from the original DOP files.

The SAM segmentation masks comprise visually coherent objects inside and outside the photovoltaic footprints. To separate photovoltaic module row masks from irrelevant background masks, several spectral and geometric properties were calculated. From the RGB and NIR reflectance bands, four complementary spectral metrics, each tailored to highlight different land-surface features were derived.

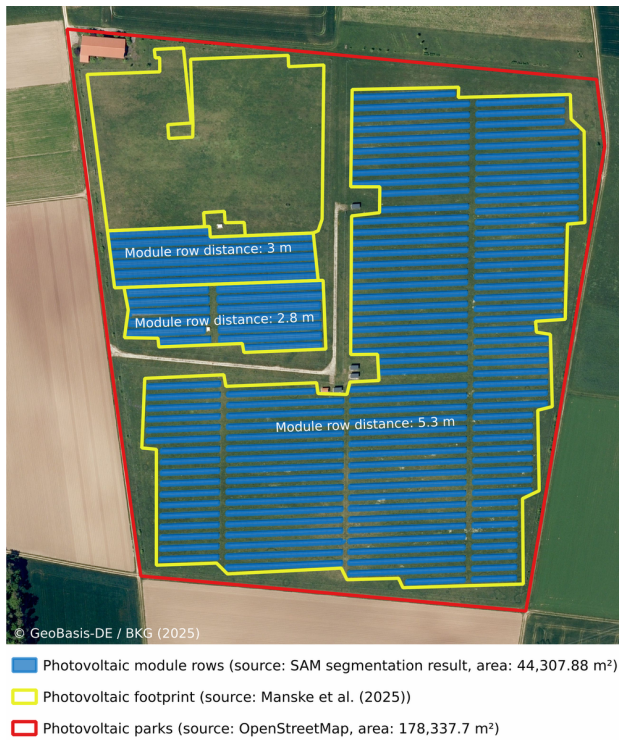


Figure 2. Example of a GMPV park illustrating the terminology used.

The Blue Normalized Difference Vegetation Index (BNDVI), as defined by Wang et al. (2007) leverages the strong contrast between chlorophyll absorption in the blue and high reflectance in the NIR to highlight vegetation.

$$BNDVI = \frac{NIR - Blue}{NIR + Blue} \quad (1)$$

To emphasise build-up and impervious surfaces the Visible to NIR Reflectance Ratio (VNRR), was calculated by the ratio of total visible to NIR reflectance:

$$VNRR = \frac{Blue + Green + Red}{NIR} \quad (2)$$

Taking advantage of the fact that photovoltaic modules absorb more strongly in the blue than in the red portion of the spectrum (Schinle et al., 2015), the Normalized Red Blue Difference (NRBD) was computed. It highlights pixels where red reflectance exceeds blue reflectance, independent of overall visible band brightness.

$$NRBD = \frac{Red - Blue}{Red + Green + Blue} \quad (3)$$

By first inverting red and green reflectance into a “darkness” measure (256-digital number) and then calculating the geometric mean, areas of low illumination in these bands are emphasised, effectively identifying regions of shadow. The Red Green Geometric Mean is given by:

$$RGGM = \sqrt{(256 - Red) \times (256 - Green)} \quad (4)$$

For each segment, the median and standard deviation of the pixel values and spectral metrics were calculated to summarize their spectral characteristics. The calculation of the geometric features for each polygon comprised the

- area,
- footprint overlap percentage,
- oriented bounding box area and
- its relative oriented bounding box area increase compared to the original polygon area, as a proxy for rectangularity.

After a feature value exploration these geometric features together with selected median values of spectral features, were used to filter out single unwanted outlier polygons. In addition, only the segmentation polygons with an intersection with the photovoltaic footprints were kept. The cleaned-up polygon set was then clustered based on the spectral features using the DBSCAN algorithm (Ester et al., 1996).

After normalizing the spectral features, the maximum neighbourhood distance ε of the resulting k -distance curve was automatically detected using the *KneeLocator* implementation from the kneed library (Satopaa et al., 2011). This was performed on a photovoltaic footprint level using a neighbourhood size of $k = 5$ polygons, which served as an input parameter for the DBSCAN clustering. This local fitting and clustering approach was applied to minimise spectral heterogeneity and thus the number of clusters. From the resulting clusters, noise cluster and background clusters – characterized by very small polygon areas or a high boundary touch ratio with other clusters – were discarded, retaining only the clusters representing photovoltaic module rows.

SAM sometimes struggled to detect all photovoltaic module rows and to distinguish them from their shadows or other background elements. However, using multiple overlapping image chips increased the number of times each pixel was segmented, improving the chances of accurately capturing all module rows. This redundancy allowed polygons containing shadows to be filtered through a module row width outlier analysis, without creating gaps in the rows of photovoltaic modules. The many individual segment polygons from the overlapping image chip were then merged into single module row polygons if they were connected. This workflow was applied separately for each federal state, after which all identified module rows were aggregated to create a nationwide dataset for Germany. Finally, the photovoltaic module row data product was enriched with additional attributes, including width, length, area, rectangularity, orientation, nearest neighbour distance and the DOP acquisition date of each module row.

Based on the derived photovoltaic module row product the reference photovoltaic footprints dataset by Manske (2025) was enriched with additional attributes, including row spacing statistics of the photovoltaic module rows within each photovoltaic footprint (median, minimum, maximum and standard deviation). Furthermore the dataset was complemented with the corresponding DOP acquisition dates and the time difference in days between the DOP date and the commissioning date.

3.2 Evaluation of photovoltaic module rows

To assess the quality and completeness of the photovoltaic module row product, a visual evaluation was performed. Therefore the extracted photovoltaic module rows were compared with the underlying DOP imagery for each photovoltaic footprint and an evaluation class was assigned based on visual interpretation. The used evaluation classes are illustrated and described in Table 1.




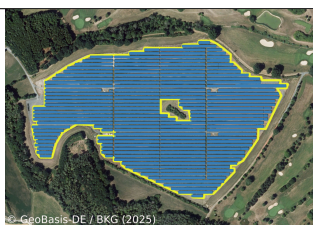
Class	Description	Example
0	True Negative (DOP was recorded before the photovoltaic park was build)	
1	False Negative	
2	True Positive with < 90% of total photovoltaic module area in footprint correct (Missing photovoltaic module areas or incorrect objects)	
3	True Positive with > 90 % of total photovoltaic module row area in footprint correct	

Table 1. Evaluation classes.

3.3 Photovoltaic parks

In order to accurately calculate the ground coverage ratio of the photovoltaic modules (PV-GCR) of each park, it is crucial to obtain the park boundaries. For this purpose, the Overpass API and the OSMPythonTools library (Mocnik, n.d.) were used to download OSM data, since the boundaries of most parks are available as polygons in the database. Accessing relevant data required identifying suitable OSM tags, as the tags specify the data structure. Initial analysis revealed that relying on a single tag was inadequate, since several polygons were linked to a variety of tags. As a result, it was essential to examine all tags linked to “photovoltaic” for PV-GCR calculations. Identifying appropriate tags began with selecting representative OSM polygons using both the photovoltaic footprint data and the OSM basemap as references. Relevant tags were determined through manual inspection of individual OSM objects on the official OSM website¹ and the corresponding wiki

¹ <https://www.openstreetmap.org>

pages for photovoltaic-related tag documentation. Once the appropriate tags were identified, all corresponding polygons were downloaded and merged into a single vector dataset for further analysis. However, the initial dataset included not only the target photovoltaic parks, but also other polygon types, such as rooftop photovoltaic systems, individual photovoltaic module rows, and overlapping geometries that may represent photovoltaic parks. Therefore, a filtering and refinement process was required to ensure data quality and relevance. This process involved removing duplicates and identifying suitable polygons by overlapping OSM and photovoltaic footprint data. The final photovoltaic park dataset consists of polygons that represent the boundaries of GMPV parks. The following list shows the selected tags, with corresponding counts:

- plant:method=photovoltaic (1,957)
- plant:source=solar (301)
- generator:source=solar (4,196)
- plant:output:electricity=yes (17)
- power=generator (4)
- generator:method=photovoltaic (4)

The PV-GCR of each photovoltaic park was calculated by dividing the total area of its photovoltaic module rows, $A_{modules}$, by the total area of the park, A_{park} .

3.4 Evaluation of photovoltaic parks

To determine the optimal polygon outlines, a visual comparison was made between OSM polygons and DOP and footprint data. A sample of 1 % of the photovoltaic footprints was extracted and manually evaluated. The following comparison was made for two purposes. First, it was necessary to determine if an OSM photovoltaic polygon existed for the footprint data. Second, it was necessary to evaluate if the existing polygon matched the outlines of the parks as visible in the DOP imagery. The focus has been on the fences and hedges surrounding the parks. The evaluation consists of the following classes:

- Yes: Corresponding OSM polygon that fits the outlines of the park
- No: No OSM polygon at all
- Intersects: If a corresponding OSM polygon was available but did not depict the outlines properly

4. Results

By employing the described approach (RQ1), three geospatial data products for GMPV in Germany have been created and made freely available on Zenodo (Albert et al., 2025):

1. *Photovoltaic system footprints*, including the GMPV system footprints as defined by Manske (2025), enriched with additional attributes like the evaluation class, DOP metainformation and statistics for the nearest neighbour distance
2. *Photovoltaic module rows*, comprising module rows derived from the DOP, enriched with the spatial parameters width, length, area, rectangularity, orientation and nearest neighbour distance

3. *Photovoltaic parks*, representing the park polygons derived from OSM and the corresponding metadata per park, such as PV-GCR

Utilizing this data, a descriptive analysis of the dataset's characteristics was performed to enhance understanding and contextualize the quality of the results regarding GMPV systems in Germany. The key findings and insights are presented below. It is important to note that, to minimise potential inaccuracies, only parks with evaluation class 3 (5.550 parks in total) were included in the examination.

4.1 Photovoltaic module rows

For the derived photovoltaic module rows product the module row parameters width, length, orientation and row spacing were calculated. Figure 3 and Figure 4 show the value distribution of these parameters for all extracted module rows of high quality in Germany. As the width, length and row spacing (calculated as nearest neighbour distance) values had large outliers the figures show only the 95th percentile for these parameters.

It can be seen that the module rows have a relatively normal distributed width with a median of 4.17 m. Their length, on the other hand, is very heterogeneous with a median of 41.2 m.

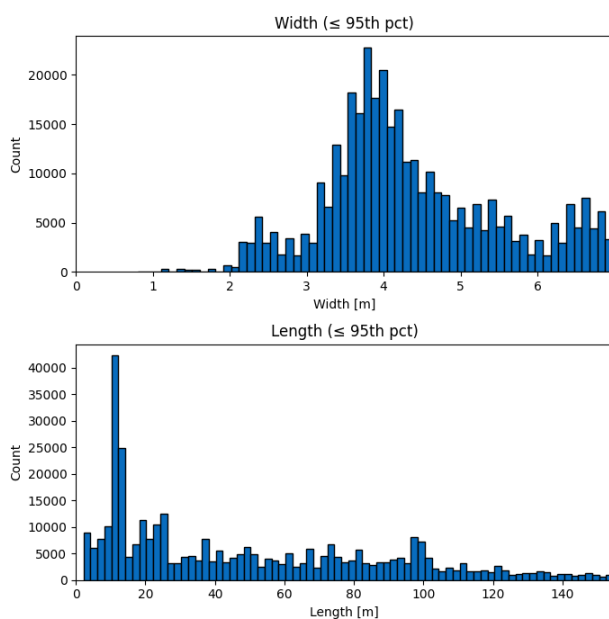


Figure 3. Distribution of width and length of photovoltaic module rows up to the 95th percentile.

The main axis orientation angle, defined as azimuth (0° corresponds to north, angle increases clockwise) indicates that the majority of photovoltaic module rows are oriented in an east-west direction, with a median orientation of 88°, resulting in the modules predominantly facing south. Only a small proportion of 5.0 % of the module rows have a north-south orientation (< 30° or > 150°). The row spacing is relatively normally distributed around a median of 3.26 m.

Figure 5 illustrates the row spacing of newly commissioned photovoltaic systems between the years 2000 and 2023. Due to the prerequisites of being classified as evaluation class 3, having a confirmed commissioning date, and the availability of median row spacing data, this plot encompasses only 59.5 % of the total

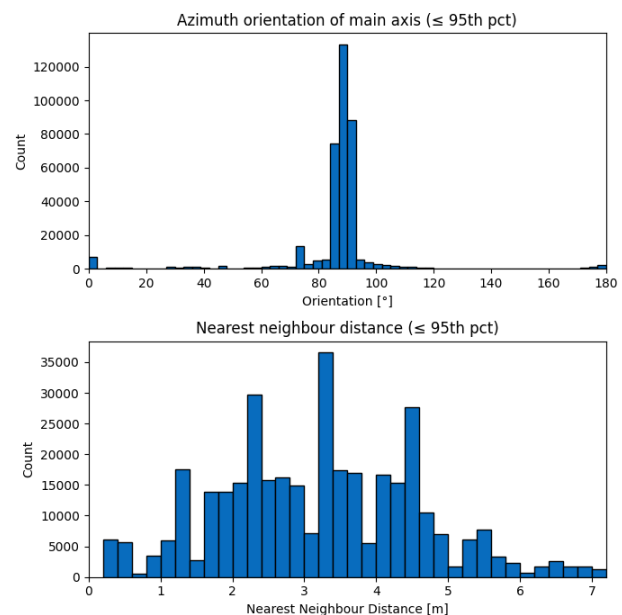


Figure 4. Distribution of orientation and nearest neighbour distance values, up to the 95th percentile (excluding orientation).

photovoltaic footprints. While installations in the first ten years (2000–2009) show wider spacing with higher variability (median mean 4.85 m with a median range of 3.8 m), more recent parks of the last 10 years (2014–2023) tend to have significantly smaller and more consistent row spacing (median mean 2.57 m with a median range of 0.75 m). This trend proved to be significant, with a decrease of 0.18 cm / year from 2003-2023 ($p < 0.001$, $R^2 = 0.081$).

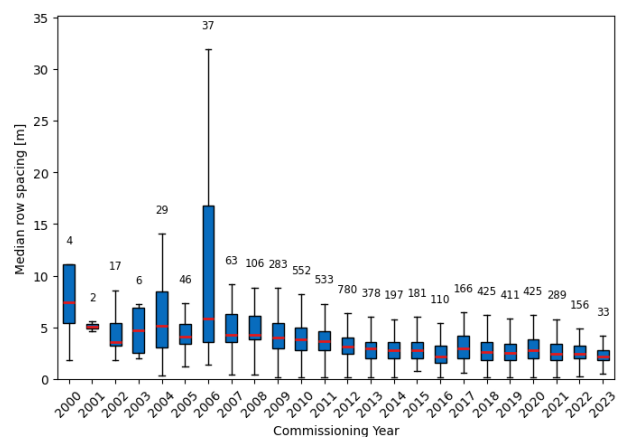


Figure 5. Box-plot of row spacing values for the commissioning years between 2000 and 2023. The red line represents the annual median nearest neighbour distance and the sample size (number of photovoltaic footprints) is annotated above each box.

4.2 Photovoltaic parks

The histogram in Figure 6 displays the distribution of PV-GCR across the evaluated photovoltaic parks. In total, for 4,605 parks OSM-derived outlines were available. The PV-GCR distribution is unimodal and slightly right-skewed, with an interquartile range spanning from 36.02 % to 51.07 %. The median PV-GCR is 43.04 %.

About 36.02 % of all photovoltaic parks fall below the first quartile (25 %) PV-GCR threshold. This high proportion can probably be attributed to undetected module rows in the imagery, leading to an underestimation of ground coverage and introducing bias into the overall PV-GCR distribution.

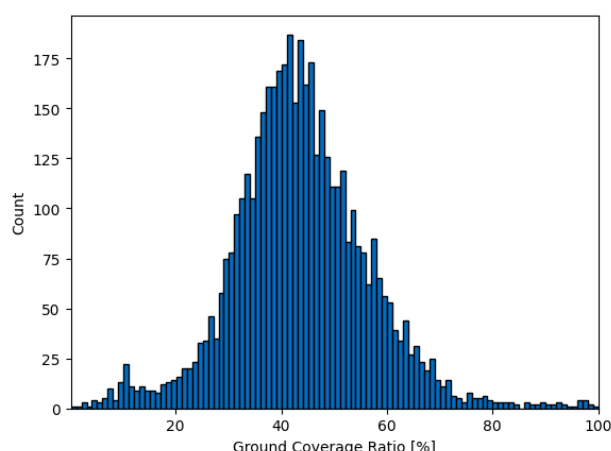


Figure 6. Distribution of PV-GCR.

4.3 Evaluation results

The evaluation of the photovoltaic module rows for the photovoltaic footprint product resulted in the class distribution summarised in Table 2. For 11.9 % of the photovoltaic footprints no module rows existed in the DOP, as the DOP was recorded before the photovoltaic park was built. Approximately 6.5 % of the footprints do not contain any module rows in the product, although they were present in the imagery (false negatives). True positive classes 2 and 3 make up 18.5 % and 63.2 % of the total respectively, reflecting that the majority of the photovoltaic modules were accurately extracted. Taking into account only the footprints, where module row extraction is possible with a current DOP (evaluation classes 1, 2 and 3), the proportion of true positives is 92.7 %, with 71.7 % of these being class 3.

Class	PV footprint count	Share in percent
0	1044	11.89%
1	568	6.47%
2	1622	18.47%
3	5550	63.18%
SUM	8784	100%

Table 2. Evaluation class distribution.

The evaluation of the OSM sample shows that 75.6 % of polygons (*Yes*) are suitable for depicting the park boundaries. For 3.9 % (*Intersects*) of polygons, the parks are incorrectly delineated, but the polygons are still existent. For 20.5 % (*No*) of the sample data no corresponding OSM polygon exists.

5. Discussion

The approach presented in this paper allows for extracting detailed spatial parameters of GMPV systems through segmentation and classification applying SAM respectively DBSCAN

(*RQ1*). This resulted in a comprehensive dataset describing detailed parameters, such as row spacing and the coverage ratio of modules of GMPV parks in Germany (*RQ2*).

The evaluation of the described approach reveals that for 63 % of the GMPV footprints a high-quality extraction of photovoltaic rows was possible. However, the method could not fully segment the rows of every GMPV park in Germany, which limits its applicability for nationwide interpretations. For the purposes of spatial analysis, only Evaluation Class 3 was used. When deriving parameters related to GMPV parks as a first approach, it is important to note that their interpretation is affected by several factors, which will be discussed in the following paragraphs. The generated PV-GCR and row spacing data provide a valuable foundation for further analysis by offering insight into the spatial configuration of photovoltaic parks. For example, the decrease in row spacing suggests an increasing emphasis on energy yield per hectare. This design shift likely reflects economic optimisation strategies, where closer row spacing enables higher installed capacity on limited land.

Further, the PV-GCR analysis suggests that the majority of parks has a relatively low PV-GCR. However, this should be treated with care because interpretations of the PV-GCR are probably limited by the data products. First, the OSM-extracted park polygons must be considered because they contain inconsistencies. The user-generated polygons vary due to individual interpretations of park boundaries. Further, as seen in the filtered OSM dataset, there are cases where larger polygons visually contain more than one GMPV park as well as cases where polygons are smaller and do not match the boundaries of an GMPV park. Second, another limitation involves the underlying method. The data product lacks some photovoltaic rows, as SAM did not detect each individual photovoltaic module row and did not process some of the image chips. Additionally, the automatic filtering process resulted in the exclusion of wider rows from the product. This results in an incomplete representation of some module areas in GMPV parks. These constraints affect the calculation of the PV-GCR, as a lower module area decreases the PV-GCR. In conclusion, the PV-GCR calculation can be used as an orientation, but a thorough individual review is necessary to make reliable statements on a park specific level. Furthermore, it is important to emphasize that the PV-GCR only encompasses the module covered area, whereas legally binding metrics, such as the *Grundflächenzahl* in Germany, also include other structures in addition to the photovoltaic modules.

Although the results reveal interesting details, there are important limitations to discuss regarding the selection of data sources. First, the photovoltaic footprints of Manske (2025) are not complete. While the dataset is expected to demonstrate a high degree of completeness due to its reliance on the MaStR, there are a few instances where certain GMPV systems remain unrepresented. Additionally the geometric accuracy of the outlines is partially insufficient and the actuality is limited until 4 January 2024. This has implications for all derived products, as they have been the foundational reference for segmenting and filtering. Second, using DOP is essential due to its very high spatial resolution of 0.2 m, which allows for an accurate delineation of photovoltaic module rows and precise assessment of row spacing. However, the temporal inhomogeneity of DOP imagery available for Germany poses a considerable limitation. As the most recent DOP available for this research dates back to 2023, GMPV parks installed in 2024 are absent in the imagery. This leads to 11.9 % of true negatives in our generated dataset. A theoretical workaround could be a prior photovoltaic footprints fil-

tering based on the official commissioning date. This is impractical as many parks are frequently installed and appear in the DOP imagery long before they are formally registered. Furthermore, as previously noted, the inconsistent acquisition dates of the DOP imagery within Germany complicate a temporally harmonised interpretation of the PV-GCR and the row spacing on a national scale. Last, working with OSM data presents both opportunities and challenges. As an open platform it provides access to extensive pre-labelled geospatial data on a global scale, but it also presents difficulties related to data quality and consistency, as the dataset is user-generated and may vary in accuracy and reliability. In order to extract the required polygon features, it was necessary to identify the relevant tags, as the tag structure lacked consistency and enforced standards. For example, some tags contained mixed polygons representing both single photovoltaic modules as well as GMPV parks. Although the OSM community is generally very responsive and effective at updating information about newly constructed installations, 2310 photovoltaic parks (26 % of the footprints available) are still missing in OSM, indicating a substantial underrepresentation.

Besides the data sources that were used, the methodology also faces some limiting factors. Due to the unavailability of training data on photovoltaic modules, SAM was applied for the photovoltaic footprints as an unsupervised foundation model for image segmentation. This dependence on the GMPV location information is the main limitation for an independent regular repetition of the analysis. Combining this information with DOP imagery forms a suitable data basis for implementing a zero-shot approach with SAM, which tends to perform better on high-resolution images (Osco et al., 2023). Promising results could be expected due to the availability of extensive hyperparameters. However, finding an appropriate configuration that would generate optimal results across the diverse range of photovoltaic systems was challenging. Even within a single GMPV park, it was difficult to accurately identify individual rows as different segments, often leading to under- and over-segmentation of module rows, e.g. in the presence of shadows. Ren et al. (2023) had similar findings, in which they found that SAM has difficulties in recognising seemingly clearly defined objects, such as photovoltaic modules. As a result, the process of determining the optimal configurations to achieve the best results, while keeping processing costs reasonable proved to be more complex than initially anticipated. To increase the stability of the segmentation results and to ensure an accurate representation of each photovoltaic module row, overlapping image chips were used. However, this approach required a large amount of computational resources. Furthermore, due to the presence of multiple objects within the parks, such as buildings and transformer stations, it was not always feasible to isolate only the rows of photovoltaic modules.

Thus, the generated dataset itself has some notable characteristics. For instance, having a closer look at the polygons representing the photovoltaic module rows reveals that the edges are not entirely straight. Addressing this limitation by refining the polygon edges could improve the quality of the dataset. For example the implementation of an edge enhancement approach such as the edge-enhanced SAM introduced by Chen et al. (2025) is a potential solution for this problem. Another aspect is that there are cases in which multiple rows are merged within a single polygon. As a result, this lack of clear delineation sometimes impedes the accurate quantification of the number of rows within a specific photovoltaic park.

One possible next step would be to integrate additional remote sensing data sources, such as LiDAR data and derived elevation models, as they would provide benefits in two main aspects. Initially the integration could improve the dataset by addressing the issue of an incomplete segmentation of module rows and improving the accuracy of polygon edges, as elevation models could provide a more accurate delineation of objects. Another advantage of incorporating elevation models is the potential to identify additional photovoltaic-related features such as module height, tilt and park-related elements such as fences. If the detection of fences is feasible, this capability would enhance the delineation of park areas and allow for the definition of more precise park boundaries, independent of OSM data. To be completely independent from manually mapped data, the integration of an automated object detection model could be a solution.

6. Conclusion

The presented study supplements existing products for remote sensing analyses of GMPV systems with an approach to identify additional parameters such as row spacing and PV-GCR, accompanied by a descriptive analysis of the derived data. Additionally, the findings allow for potential applications that extend beyond this immediate use case, as the results open numerous possibilities for extension or further studies at various levels.

First, the methodological approach can be adapted and utilised in other remote sensing applications that similarly lack training data. In the context of the presented study, this could include other components of GMPV systems, such as storage transformers or biotope areas. However, this approach is not limited to GMPV, as it is generally applicable.

Second, the employed technical approach is adaptable and can be expanded for application in diverse geographical regions, thereby enhancing its broader applicability as well as contributing to the development of a more comprehensive dataset.

Third, the dataset can be utilized for the derivation of additional products. The creation of the dataset involved significant processing steps and computational effort, resulting in a representation of the current state of knowledge. To facilitate future updates efficiently, the next planned step involves using the created dataset to train a model for the detection of module areas. This approach will enable the regular and efficient collection of system parameters, allowing for improved monitoring of developments.

Authorship contribution statement

All authors contributed equally.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The results are derived from a joint use case between the authors and the German Federal Agency for Nature Conservation (BfN). The authors would like to thank the BfN team members who actively contributed nature conservation-related research questions, as well as regulatory needs and requirements, to the use case in their role as Use Case Owners.

References

- Albert, J., Gärtner, P., Schymik, C., Wehner, C., Siegismund, J., Klingner, S., 2025. Ground-Mounted Photovoltaic Systems in Germany. <https://doi.org/10.5281/zenodo.15387100>.
- Bai, Z., Jia, A., Bai, Z., Qu, S., Zhang, M., Kong, L., Sun, R., Wang, M., 2022. Photovoltaic panels have altered grassland plant biodiversity and soil microbial diversity. *Frontiers in Microbiology*, 13, 1065899.
- Bastani, F., n.d. Satlas: Open AI-Generated Geospatial Data. <https://github.com/allenai/satlas> (10 June 2025).
- BKG, 2025. Digital Orthophotos with 20 cm Ground Resolution (DOP20). Federal Agency for Cartography and Geodesy. <https://gdz.bkg.bund.de> (10 June 2025).
- BMJV, 2014. Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG 2023). Federal Ministry of Justice.
- Bradski, G., 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Bundesnetzagentur, n.d. Marktstammdatenregister (MaStR). <https://www.marktstammdatenregister.de/MaStR> (10 June 2025).
- Carvalho, F., Lee, H. K., Blaydes, H., Treasure, L., Harrison, L. J., Montag, H., Vucic, K., Scurlock, J., White, P. C. L., Sharp, S. P., Clarkson, T., Armstrong, A., 2024. Integrated policymaking is needed to deliver climate and ecological benefits from solar farms. *Journal of Applied Ecology*, 1365–2664.14745. <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.14745>.
- Chen, Y., Zhou, J., Chen, Y., Wang, J., Zhang, X., Ge, Y., Ma, H., 2025. Edge-enhanced SAM for extracting photovoltaic power plants from remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 140, 104580. <https://linkinghub.elsevier.com/retrieve/pii/S1569843225002274>. Publisher: Elsevier BV.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining*.
- Hilker, J. M., Busse, M., Müller, K., Zscheischler, J., 2024. Photovoltaics in agricultural landscapes: “Industrial land use” or a “real compromise” between renewable energy and biodiversity? Perspectives of German nature conservation associations. *Energy, Sustainability and Society*, 14(1), 6.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. <https://arxiv.org/abs/2304.02643>.
- Manske, D., 2025. Geo-locations and system data of renewable energy installations in germany. <https://doi.org/10.5281/zenodo.14627853>.
- Mocnik, F.-B., n.d. OSMPythonTools. <https://github.com/mocnik-science/osm-python-tools> (10 June 2025).
- Osco, L. P., Wu, Q., De Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., Marcato, J., 2023. The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124, 103540. <https://linkinghub.elsevier.com/retrieve/pii/S1569843223003643>. Publisher: Elsevier BV.
- Ren, S., Luzi, F., Lahrichi, S., Kassaw, K., Collins, L. M., Bradbury, K., Malof, J. M., 2023. Segment anything, from space? <https://arxiv.org/abs/2304.13000>.
- Robinson, C., Ortiz, A., Kim, A., Dodhia, R., Zolli, A., Nagaraju, S. K., Oakleaf, J., Kiesecker, J., Ferres, J. M. L., 2025. Global Renewables Watch: A Temporal Dataset of Solar and Wind Energy Derived from Satellite Imagery. Version Number: 1.
- Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B., 2011. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, IEEE, Minneapolis, MN, USA.
- Schinke, C., Christian Peest, P., Schmidt, J., Brendel, R., Bothe, K., Vogt, M. R., Kröger, I., Winter, S., Schirmacher, A., Lim, S., Nguyen, H. T., MacDonald, D., 2015. Uncertainty analysis for the coefficient of band-to-band absorption of crystalline silicon. *AIP Advances*, 5(6). <https://pubs.aip.org/adv/article/5/6/067168/650/Uncertainty-analysis-for-the-coefficient-of-band>. Publisher: AIP Publishing.
- Schmietendorf, K., Peinke, J., Kamps, O., 2017. The impact of turbulent renewable energy production on power grid stability and quality. *The European Physical Journal B*, 90(11), 222. <http://link.springer.com/10.1140/epjb/e2017-80352-8>.
- Smith, O., Cattell, O., Farcot, E., O’Dea, R. D., Hopcraft, K. I., 2022. The effect of renewable energy incorporation on power grid stability and resilience. *Science Advances*, 8(9), eabj6734. <https://www.science.org/doi/10.1126/sciadv.abj6734>.
- Wang, F.-m., Huang, J.-f., Tang, Y.-l., Wang, X.-z., 2007. New Vegetation Index and Its Application in Estimating Leaf Area Index of Rice. *Rice Science*, 14(3), 195-203. [https://doi.org/10.1016/S1672-6308\(07\)60027-4](https://doi.org/10.1016/S1672-6308(07)60027-4).