Validation and extension of data on solar power plants in the Marktstammdatenregister using aerial images and building data

Master Thesis

presented by
Esther Vogt
Matriculation Number 1722616

submitted to the
Data and Web Science Group
Prof. Dr. Christian Bartelt
University of Mannheim

March 2023

Acknowledgements

I would like to thank Sascha Marton from the Institute for Enterprise Systems (InES) at University of Mannheim and Florian Kotthoff from fortiss GmbH for their continuous support, help and feedback throughout my work on this thesis.

Abstract

The goal of the German government to heavily increase the amount of solar capacity induces the challenge to efficiently monitor an ever more decentralized energy system. In this work, I present a methodology for automated validation and extension of the Marktstammdatenregister (MaStR) for building-mounted solar Photovoltaics (PV) systems. It is based on extraction of information on systems from image and building data and generation of mappings to units in the public registry. It extends previous work by using freely accessible building data from Open Street Map (OSM) and comparing a variety of approaches for creating correspondences to MaStR units. The approach is evaluated on datasets covering the city and district of Munich. My evaluation identifies erroneous entries for several fields: For example, the capacity detected in images exceeds the registered capacity by 13% which indicates an incorrect amount of units registered as in operation. Especially locational information is often inaccurate and not coherent: For 16% of large units no PV system could be detected at the reported place of installation. My approach demonstrates the potential to fill these gaps by automatically generating more precise localizations of systems and corresponding buildings. It allows to complement the MaStR with building properties such as standardized roof shapes or building levels.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement and Contribution

To contribute to climate protection goals, Renewable Energy Sources (RES) are planned to account for 80% of electricity production in Germany by 2030 (Section 1 (2) EEG). To achieve this goal, an expansion of solar capacity from 88 GW in 2024 to 400 GW in 2040 is planned (Section 4 EEG). This transformation requires tools for effectively planning the installation of new power plants and expansion of grids to balance the fluctuating production of solar power. In addition, applications are needed to support the management of an ever more complex, decentralized and flexible energy system.

To provide these efforts with extensive, transparent and reliable information, the Federal Network Agency (BNetzA) introduced the Marktstammdatenregister (MaStR) as the central database of the German energy system in 2019. Section 111e of the Electricity and Gas Supply Act (EnWG) motivates its purpose with regard to different stakeholders and use cases: Technical characteristics such as location or capacity of installations are essential for a Distribution Network Operator (DNO) or Transmission System Operator (TSO) to plan and manage the electricity grid effectively. State authorities like the Federal Ministry for Economic Affairs and Climate Action (BMWK) use the MaStR to monitor the expansion of RES and report corresponding information to the public. It also supports them in the design and adjustment of energy policies.

Additionally, MaStR data can support research in various domains. This includes environmental scientists assessing the potential of RES to reduce greenhouse gas emissions. Moreover, researchers in sociology, anthropology, and political science could use MaStR data to analyze the social and political factors that influence the adoption of RES in Germany [34].

Given its wide-ranging application, it is crucial that data in the MaStR is accurate and up-to-date. Not only could inaccurate data lead to incorrect decisions with negative consequences for the RES sector but also erode public trust. However, there is a growing concern about the accuracy of the MaStR data especially for solar installations. These systems are more difficult to monitor than wind installations due to their decentralization and higher amount. For example, [36] report a significant amount of duplicated entries, erroneous capacities and addresses or installations not listed in the registry for a sample area in North-Rhine Westphalia.

The root cause for this is that MaStR data must be entered manually by the plant operators. This introduces high manual effort not only for the operators but also for the corresponding DNO who is required to ensure sufficient data quality. Automating the data collection process and implementing standardized procedures and formats could help to address these issues and improve the accuracy, timeliness, and efficiency of MaStR data collection.

For solar, automated detection of PV systems with deep learning methods has recently gained more attention in research: While [34] focus on detection of field systems in satellite images, [39] even suggest methods to automatically detect a variety of RES. [36] combine semantic segmentation and 3D-building data to estimate relevant system properties like capacity, tilt or orientation. Not only do these approaches aim to scalably improve data quality of the MaStR, but to also assess the potential of creating such a registry in a fully-automated manner. This is especially relevant in an international context since no other country maintains and publishes a registry like the MaStR to the extent that the BNetzA does in Germany.

This thesis aims to contribute to the ongoing efforts to explore ways for automated validation and complementation of MaStR data for solar installations. The research questions that will be addressed are:

RQ1. How can the MaStR for solar records be validated, i.e. using external data like images and building data?

RQ2. What information from these external sources could provide a useful extension to the MaStR for solar?

The first research question (RQ1) seeks to explore the usefulness of various methods for identifying potential error sources in the MaStR data. The main focus will be on approaches for generating unit-level correspondences between the MaStR and external data sources like aerial imagery and building models. While [36] are the first to assess the potential of highly detailed 3D building data, such information is not yet available on a national level. To fill this gap, this work explores the potential of open-access OSM data. Additionally, previous work applies computationally expensive nearest neighbor matching and proximity heuris-

tics. In contrast, I will evaluate the potential of various types of spatial joins based on address as well as geocoordinates. Not only does this aim to increase the number of mappings but to also assess data quality and limitations of locational information in the MaStR.

From the knowledge gained by answering RQ1, the second research question (RQ2) aims to identify the types of information from external sources that can be integrated into the MaStR database. This expansion should improve data accuracy and provide a more detailed context for solar installations required for different applications.

To evaluate approaches for validation and extension of MaStR information, they are applied to datasets covering the city and district of Munich. In addition, this research prioritizes the evaluation of information on building-mounted systems in the corresponding reference area.

It is structured as follows: First, fundamental concepts for understanding the research methodology as well as the political and legal framework of the MaStR are introduced in Chapter 2. Since the different data sources are the central element in this work, their key characteristics are presented in Chapter 3. Based on the knowledge of their extent and attributes, Chapter 4 proposes methods for processing individual datasets and generating mappings between their entities. The results of applying these approaches to the Munich reference area are presented in Chapter 5. To conclude, Chapter 6 summarizes and discusses the extent to which these findings answer RQ1 and RQ2 by listing relevant limitations and giving an outlook on further research opportunities.

1.2 Related Work

Before detailed introduction to the methods proposed in this work, this section gives a comprehensive overview of related research. Although several authors discuss the automated detection of solar PV systems by deep learning techniques, I focus on those which aim at using these to validate the MaStR and fill gaps in its data.

[36] extend previous research on automated detection of PV systems with deep learning methods by including 3D information of buildings. However, their research involves high-resolution building models with Level of Detail 2 (LoD2) which are not yet available across entire Germany. Their research is limited to roof-mounted installations, as is this work. Not only does their “3D-PV-Locator” detect solar systems but also calculates capacity estimates for four counties in North-Rhine Westphalia (NRW) on an aggregated level and for individual units. Building properties like tilt and orientation are also compared to MaStR values. Addition-

ally, the authors benchmark urban against rural counties and compare their results against the official registry. Instead of a searching for intersecting geometries as described in Section 4.3, they apply a nearest neighbor matching on MaStR addresses to determine correspondences between detected and reported units. However, it remains unclear how they handle groupings of several systems forming a single unit according to the MaStR and how these impact the performance of their results. To fill this gap, this research compares different join types and matching strategies for single as well as groups of adjacent units.

A dataset including multiple RES in Germany is presented by [34]. However, their data only references data until 05/07/2021 and is limited to field systems for solar. For its generation, the authors rely on data fusion from multiple publicly accessible sources like the MaStR as well as restricted sources. To determine correspondences, they propose an allocation heuristic based on spatial proximity and other criteria like matching Official Municipality Code (AGS) code and a reasonable ratio of capacity to plant area. In cases where it fails, they manually generated mappings which limits scalability of their approach i.e. to the high amount of building-mounted solar systems on a national level. For generation and validation of correspondences, the researchers suggest Sentinel-2 satellite and Google Earth Pro imagery. However, they report that these image datasets only yield reliable results for solar units covering more than 3000 m^2 wherefore this approach is not suitable for smaller PV units.

The detection of a multitude of RES and improvement of MaStR data quality is the aim of [39], too. They trained a deep learning based model, the so called "DetEEktor", on images of different resolution covering a selective area of Saxony. Similar to [36], they aim to compare the rural area of Brambach to the urban environment of Chemnitz. For solar, their approach detects building-mounted as well as field systems. Moreover, the authors report that their model can reliably distinguish solar PV and solar thermal collectors (see Section 2.1.1). To explore exact false positive versus false negative ratios, they manually created full-coverage ground-truth masks for Brambach, but not for Chemnitz due to resource constraints. They do not generate exact correspondences to MaStR units but rather compare the total number of installations per area. The researchers report to have found about 2.5 times the amount of roof-mounted PV systems than registered in the MaStR for the urban area of Chemnitz. However, it remains unclear how they adhere to the definition of solar units as given in the MaStR. Another limitation of their research is that their data falls into the introductory phase of the MaStR: their images are from January 2020 and MaStR records from December 2020. Since plant operators were allowed to migrate data for existing units until early 2021 and DNO validations were delayed due to the high number of registrations, data from this time frame might not fully reflect current data quality and extent.

To summarize this literature review, methods for validation of the MaStR with a focus on solar installations have been suggested by various authors. However, they mainly focus on general system detection and capacity estimation. To complement these approaches, this research tries to assess the accuracy of additional reported system properties like address, geocoordinates, place of installation and amount of modules. Additionally, I will explore methods for mapping detected systems to MaStR units based on spatial joins as an alternative to rather computationally expensive assignment heuristics suggested by [36] and [34]. While [36] also use building data to detect and validate building mounted solar units, the building models they use are not openly accessible. Therefore, this work explores the potential of freely available OSM extracts in order to assess scalability of the discussed methods.

Chapter 2

Theoretical Background

This chapter introduces concepts crucial to comprehend the research methodology and results presented in this thesis. After narrowing down the types of energy considered (Section 2.1), relevant concepts for the analysis of geospatial data are presented (Section 2.2). Finally, a general overview of the MaStR and the associated legal framework is given (Section 2.3).

2.1 PV Systems

This section delineates the different types of solar radiant energy use (Section 2.1.1). In addition, conditions and limitations for calculating solar capacity are discussed (Section 2.1.2).

2.1.1 Types of Solar Generation

In general, three types of technology use the power provided by the sun to generate heat or electricity: While solar Photovoltaics (PV) use sunlight to generate electricity, Concentrated Solar Power (CSP) installations convert the sun's energy into heat using mirrors. Solar thermal collectors also use sunlight but to heat water and air [26, p.406].

For cost reasons, CSP parks have so far mainly been deployed in sunny regions such as Morocco. In Germany, only roof- or ground-mounted solar PV and thermal collectors are installed [20]. On aerial imagery, both look similar while the latter usually consists of only 2 to 4 panels compared to approximately 6 to 7 panels for PV installations. Other than this heuristic, it will not be possible to distinguish between both types of installations in this work with the given image data and limited domain expertise. Consequently, the MaStR dataset is limited to PV collectors

only.

2.1.2 Nameplate Capacity

The maximum electric yield achievable by a solar module per hour is referenced in Peak DC Watts (kWp). It is also called nameplate power or nameplate capacity. It is determined given Standard Test Conditions (STC) to ensure comparability of different panels. These reference a setting with a PV cell and air temperature of 25°C, a solar radiation of 1,000 watts per m^2 (F1000) and with an angle of incidence of 48.2°. Thereby, nameplate capacity is a constant metric in contrast to actual yields which would be given in Hourly DC Watts (kWh).

In the MaStR, nameplate power is given as gross and net capacity. The net value is automatically calculated as the minimum of the peak gross capacities of PV modules and inverter. An inverter is used to convert Direct Current (DC) electricity generated by the PV cells of a module into Alternating Current (AC) electricity. This is required for feed-in to the public grid [26, p.423ff].

2.2 Geospatial Data Analysis

This section explains special features to be considered in the processing of georeferenced data. These include the reference framework of the measurements (Section 2.2.1) as well as characteristic data types (Section 2.2.2).

2.2.1 Coordinate Reference Systems

A Coordinate Reference System (CRS) is used to describe geographic objects in terms of 3D spheres. For different countries or regions of the world, codes provided by the European Petroleum Survey Group (EPSG) identify which reference system is used. The most common code on a global scale is EPSG:4326 specifying geographic position as values of latitude and longitude. The images used for this analysis (see Section 3.2) are given in EPSG:25832. Its coordinates are given as values of easting and northing (E,N) which is mostly used in European countries [22]. To align geospatial information requires transformations between different CRS which are called (re)projections [24, p.42].

2.2.2 Geometry Datatypes

The main distinction between geometry datatypes is made between raster and vector data. A raster is a matrix representing characteristics of the real world in the geographical area covered. Images are raster matrices of pixel values that describe

a continuous area. In most cases, the characteristic described by the individual matrix cells is color. It can be represented by a single-band raster producing grayscale images or multi-band rasters i.e. for RGB images.

Special image formats like Tagged Image File Format (TIFF) or GeoTIFF allow extension of raster data with homogeneous metadata. Both can be stored in the same file. So-called tags provide information i.e. on the spatial resolution per pixel, coordinates of the image boundaries and corresponding CRS or image size. Thereby, these formats allow representation of locational characteristics like geocoordinates on a per-pixel level.

A vector represents features in a selection of an area. The description of features is performed by assigning attributes in a textual or numerical format. Vectors have several subtypes: point, polyline, polygon and multipolygon. While a point consists of a single vertex, polylines are multiple connected vertices in an open structure. As such, start and end vertex of the connection differ. On the contrary, a polygon is a closed structure of connected vertices and thereby has an identical start and end vertex. Multipolygons are groupings of these different vector structures.

In the given dataset, the location of MaStR entries is given as points formed by tuples of latitude and longitude. The ground shape of a building however can be represented by a polygon. Moreover, if i.e. a building has an inner courtyard, it can be represented by a multipolygon as a grouping of polygons [37].

2.3 Official German PV registry (MaStR)

The following subsections explain relevant framework conditions for solar units in the MaStR that affect data quality as well as the extent of validations in this research. These include the implementation of the registry (Section 2.3.1) as well as quality assurance measures (Section 2.3.2) and data protection provisions (Section 2.3.3).

2.3.1 Implementation Timeline

The MaStR is operated by the BNetzA since 01/31/2019 as a central online database for data related to the German energy system. Its general framework and motivation was first defined in Sections 111e and 111f of the EnWG in 2016. Since 07/01/2017, the Regulation on the registration of energy industry data (MaStRV) governs its specific legal implementation details [6].

According to Sections 3 and 5 of the MaStRV, operators of plants that generate either electricity or gas are obliged to report master data on themselves and their plants. This also applies to plants that generate electricity but do not supply to

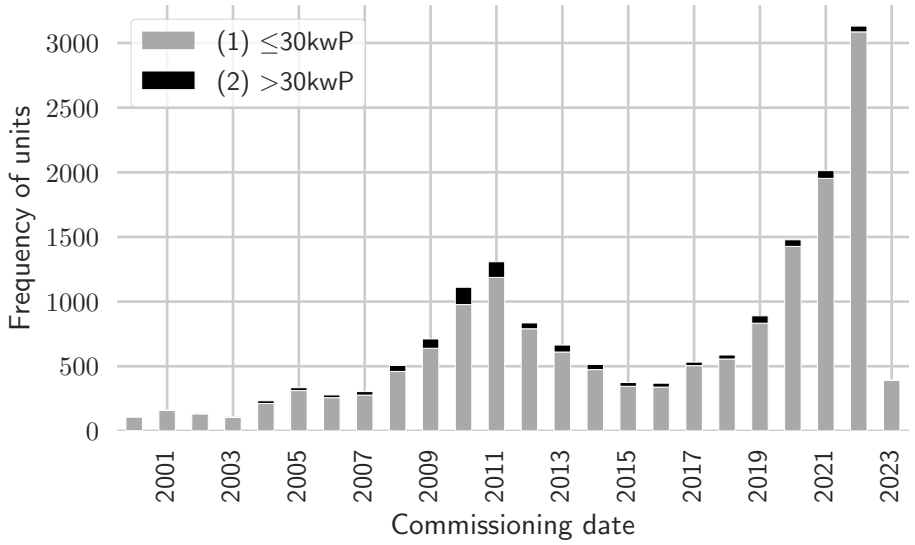


Figure (2.1) Amount of small ($\leq$ 30 kWp) and large (>30 kWp) units commissioned per year since 2000

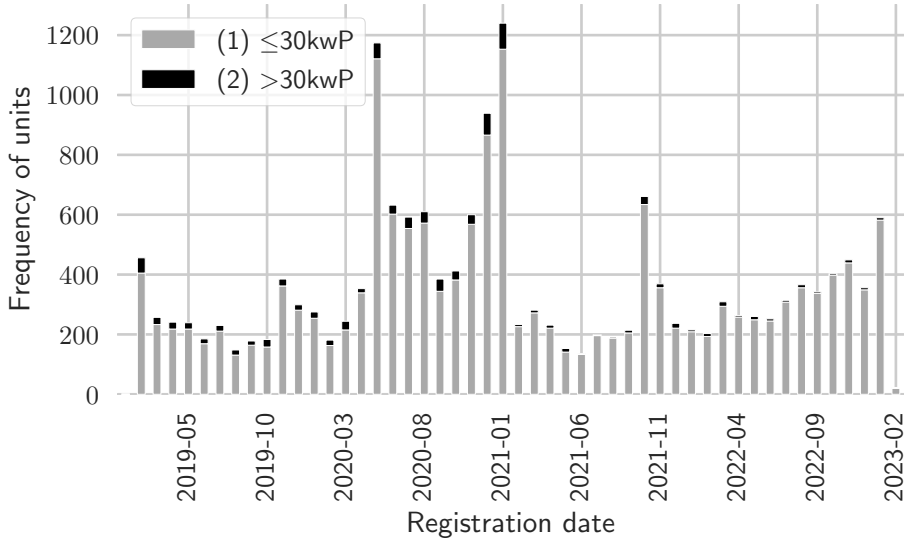


Figure (2.2) Amount of small ($\leq$ 30 kWp) and large (>30 kWp) units registered per month since introduction of the MaStR on 01/31/2019

the public grid or receive subsidy according to the Renewable Energy Act (EEG). Additionally, plants consuming large amounts of electricity have to be registered if they are connected to at least a high-voltage electricity network [2, p.4]. Registration can be carried out by authorized third parties such as family members or the installer [2, p.10].

Considering the type of information collected, the MaStR only includes master data like optimal performance values or location. It does not provide transaction data like actual production amounts or storage levels.

In Munich, 52.79% of small and 77.23% of large units started operation before introduction of the MaStR on 01/31/2019 (see Figure 2.1). Before 2019, their data was collected in the "Anlagenregister" and "Photovoltaik-Meldeportal". The fact that the reported information across registers was partially redundant caused a lot of organizational overhead for all stakeholders. Following the introduction of the MaStR, all information had to be migrated manually by plant operators due to data protection regulations [2, p.16]. A transition period of two years until 01/31/2021 was granted for existing plants. In contrast, plants commissioned since then have to be registered within one month after commissioning [7, p. 489].

In Munich, the due date for migration of existing units from other platforms to the MaStR triggered a spike of registrations across 2020 (see Figure 2.2). A maximum of 1,153 units were registered in January 2021. After the initial registration effort for existing installations, average monthly registrations over all PV units stabilized at 346 entries in 2021 and 311 entries in 2022.

2.3.2 Quality Assurance Measures

According to Section 13 of the MaStRV, information on units, plants and corresponding operators has to be verified by the DNO. The according Quality Assurance (QA) processes should be performed upon registration of commissioning and changes of data [2, p.11]. In case of discrepancies, the DNO can request the plant operator to update the data. Completion of verification is logged in the corresponding MaStR status field for DNO verifications (Section 13, paragraph 3 of the MaStRV).

The annex to the MaStRV specifies which fields are to be validated by the DNO. An excerpt of fields considered in this thesis is presented in Table 3.2. Corresponding QA procedures are described in more detail in Section 3.1.

At the time of download of the given MaStR data for Munich in February 2023, for 85.46% of solar units in operation the validation by the DNO has been completed. Among larger units, only 8.05% still have to pass the QA process. Overall, this corresponds to a capacity of 23.87 Mega Watts (MW) lacking final approval by the DNO (see Figure 2.3). Planned units are excluded from this analysis since

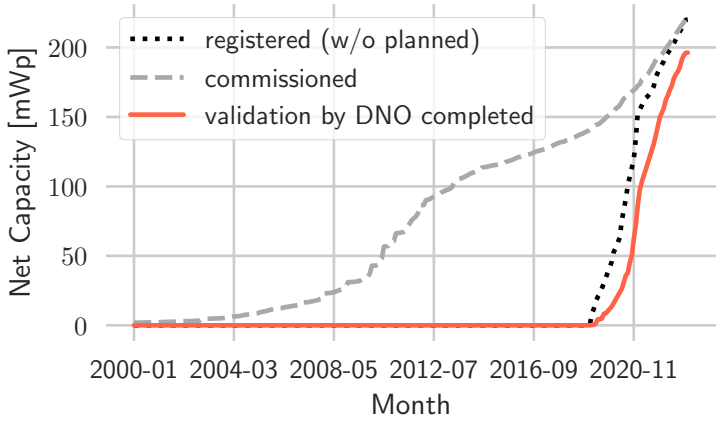


Figure (2.3) Aggregate capacity of units registered and commissioned per month compared to the capacity validated by the responsible DNO. Planned units are excluded since validation is only required after start of operation.

the DNO validation has to be performed only after commissioning.

Independent of the mandatory QA performed by the DNO, the department MaStR-QS of the BNetzA checks the registered data. According to [7, p.17 f.], for solar they recently mainly focused on dimensioning errors of capacities. Also, the BNetzA reports it to be a common problem for solar that units are registered as consumption plants. For units with a net capacity of more than 10 MW, the BNetzA checks all information reported. Obvious errors are fixed immediately and reported to the plant operators who may object. On average, the BNetzA fixes about 600 errors per month for the whole registry [7, p.491].

In general, data on units and plants registered in the MaStR should not be deleted. Even for decommissioned installations, data is persisted to allow historic analysis. After shutdown, the only information deleted is the link to the respective plant operator to ensure data privacy. Nonetheless, operators or the DNO of a unit or plant can request deletion of MaStR entries by the BNetzA. This could be required in case of duplicate registration or erroneous registration of non-existent installations [7, p.490].

Additionally, the legislator emphasizes the importance of data quality in the MaStR by considering registrations performed too late, incorrectly, insufficiently or not performed at all as an administrative offense. According to Section 21 of the MaStRV, those can even be punished with fines.

2.3.3 Data confidentiality and protection

Most information on units and plants in the MaStR is openly accessible. For units with a net capacity of up to 30 kWp, some location information is restricted from publication. This applies to street name, house number, parcel designation and exact geocoordinates of units (Section 15, paragraph 1 of the MaStRV). The most granular location information accessible is the zip code or AGS. The BNetzA introduced this limitation in accordance with Section 111e paragraph 3 of the EnWG. It requires the protection of individuals with regard to processing of personal data in compliance with EU legislation. The same applies to information on plant operators which are natural persons. Subsequently, I will refer to units with a capacity below 30 kWp as small units and large units otherwise.

While locational information is excluded from public access, the BNetzA may process it, as may public authorities and grid operators to the extent required by their legal duties. As for public authorities, those are on the one hand federal institutions and ministries steering and monitoring the transformation of the energy system, such as the BMWK, the German Environment Agency (UBA), the Federal Ministry of Food and Agriculture (BMEL), and the Federal Statistical Office. On the other hand, access is given to the Federal Cartel Office (BKartA), and the Federal Office of Economics and Export Control (BAFA), which deal with the export and pricing of energy. In addition, financial authorities of the federal and state governments use the register for subsidies and taxation. Other authorities can also request access to restricted information from the BNetzA for statistical or scientific purposes according to Section 16 of the MaStRV.

Moreover, DNO as well as TSO are not granted full access, but can only access restricted information on units connected to their grid according to Section 17 of the MaStRV.

Chapter 3

Data Source Description

To define the possible scope of analysis and methodology in this work, this chapter describes key characteristics of the datasets used. Those include MaStR extracts (Section 3.1), aerial images (Section 3.2) as well as OSM building data (Section 3.3). In addition, the dataset used to filter the different sources on the reference area of Munich is described in more detail (Section 3.4). An overview of all datasets is also given in Figure 3.1.

3.1 Solar Master Data

The following section describes accessibility of MaStR data and the corresponding data model with a focus on electricity generation and solar power in particular. After a general delimitation of the concept of a solar unit, information registered for such systems in the MaStR is described. An excerpt of MaStR fields analyzed is given in Table 3.2.

Data Access Section 111d of the EnWG explicitly requires MaStR data to be published in an easily understandable and storable format on a freely accessible platform. The BNetzA adheres to this requirements by provisioning of an online portal for manual access to XML files as well as a web service for automated queries. Data is published under the "Data licence Germany – attribution – Version 2.0" [23] and updated on a daily basis [5].

Overall, the BNetzA reports a consist amount of 50 million requests per month on the MaStR. An increasing amount of requests can be attributed to the web service due to its benefits for analyzing large amounts of data. Between March 2020 and June 2021, the amount of distinct companies using the API increased from 200 to 420 [7, p. 488 f.].

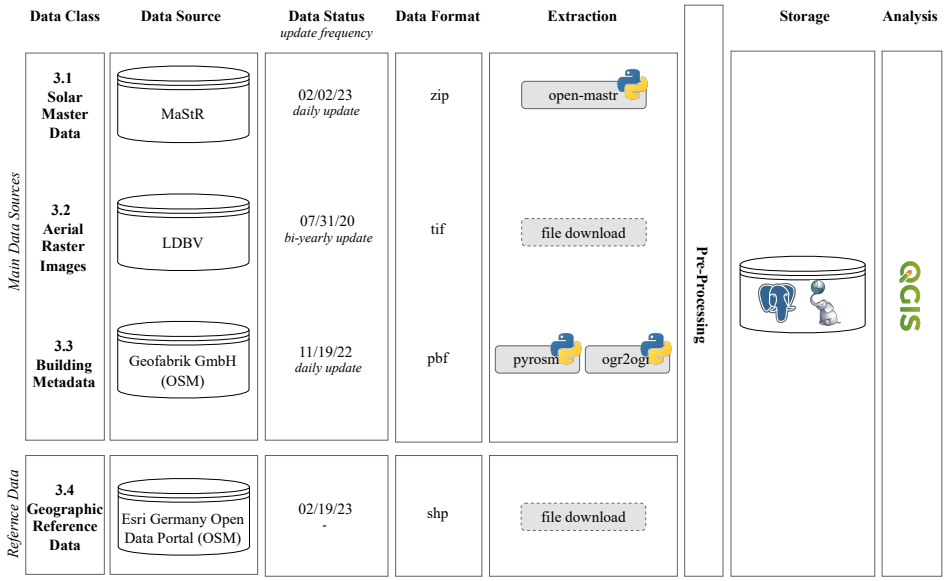


Figure (3.1) Overview Data Sources: After extraction and source-specific pre-processing, data is persisted in a Postgres database with a PostGIS add-on. Raster and vector layers are analyzed in the GIS visualization tool QGIS.

The MaStR dataset referenced in this thesis was downloaded from the official website of the BNetzA using the open_MaStR package by fortiss [25]. This package provides an extension to both, MaStR web portal and service for simplified download, processing and storage in a local Database Management System (DBMS) [25]. For example, it allows download of extracts for specific technologies like solar or wind and was used in the scope of this thesis for retrieval of MaStR data. All statistics reported thereafter reference the status of the data on the date of extraction, 02/02/2023.

General Data Model for Solar Units In the MaStR, operators have to report data on themselves and their plants or grids. Operators can either be natural persons or organisations, i.e. companies, authorities or institutions. They are registered as market actors which categorize key roles in the energy sector, i.e. “grid” operator or simply “actor in the electricity market”.

For power facilities, the most granular level are units. They are categorized by whether they produce or consume either gas or electricity. For example, a solar unit is considered a Electricity Producing Unit (SEE). Each SEE can then be further defined by the type of technology used, i.e. solar SEE.

	(1) ≤30kwP		(2) >30kwP		Total	
	C	U	C	U	C	U
(1) P	1,433.47	272	16,384.32	19	17,817.79	291
(2) O	117,330.97	16,363	102,448.00	956	219,778.97	17,319
(3) TS	41.97	12	0.00	0	41.97	12
(4) S	282.89	66	0.00	0	282.89	66
Total	119,089.29	16,713	118,832.32	975	237,921.61	17,688

Table (3.1) Amount of MaStR units (U) and sum of gross capacity in kWp (U) by unit size and operating status: P=Planned, O=Operating, TS=Temporarily shut down, S=Shut Down.

Strictly speaking, a single solar module could be assumed to form one unit according to this definition. However, data is registered for the sum of modules as a single SEE for simplification. In the context of this thesis, the terms module and panel are used interchangeably. Another equivalency used in literature is that of solar units and arrays [26, p.419].

Additionally, Section 5 (1) of the MaStRV demands operators to register multiple solar units as a single unit in the MaStR if they are installed at the same geographic location (i.e. same building) and commissioned at the same time. The sum has to be applied to fields that can be aggregated like the amount of modules or the gross and net capacity. In contrast, modules added at a later point in time have to be registered as a separate SEE.

For the city and district of Munich, there are 17,688 solar units registered with a total gross capacity of 270 MW and net capacity of 238 MW. Of these, 291 are units not yet commissioned and 78 are either permanently (66) or temporarily (12) shut down. The remaining 17,319 units are reported to be in operation (see Table 3.1).

Multiple units of the same or different type can be grouped to plants. They are further distinguished by the type of subsidy they are receiving, i.e. according to the EEG. Even if a plant and unit are identical, a grouping will be created in the MaStR. Additionally, units can be linked to other types of groupings like technical locations, i.e.Technical Electricity Producing Location (SEL). Those define the grid access points and give information on the corresponding DNO and billing area.

For data on units and plants, the corresponding operators are responsible for ensuring correctness of information reported in the MaStR. However, for technical locations this responsibility is assigned to the respective DNO [3].

Group	Name	Type	P	R	A	V*1	NP	NP*8
Location	City	str		X			X	
	District	str	?	?			?	
	House Number	str		X		X	X	
	Latitude	float		X		X		
	Longitude	float		X				
	Street Name	str		X		X	X	
	Zip Code	str		X		X	X	
PV System Attributes	Amount of modules	int	X					
	Gross Capacity	float		X			X	
	Inverter Capacity	float		X			X	X
	Net Capacity	float		X	X		X	X
	Orientation (main)	str	X					
	Orientation (other)	str	X					
	System Area	float	X					
	Tilt (main)	str	X					
	Tilt (other)	str	X					
System Installation Environment	Building Usage	str	X					
	Place of installation	str		X			X	

Table (3.2) Description of MaStR fields considered in this work according to Table II of the MaStRV annex: P=Mandatory, R=Required for registration, A=Automated entry, V*1=Restricted access for units $\leq$30kWp, NP=DNO validation, NP*8=DNO validation only for units commissioned after 06/30/17, ?=not specified in the MaStRV annex.

Validatable and extensible fields Due to the discussed data confidentiality measures, address information is publicly accessible for only 975 units which exceed the capacity threshold of at least 30 kWp (see Section 2.3.3). However, information on zip codes are given for all units, including the 94.49% of small systems. To filter MaStR entries on the reference area, search queries on the text fields *"City='München"'* and *"District='Landkreis München"'* are applied. According to Table II of the MaStRV annex, only address or land parcel (II.1.1.2) but not the exact coordinates of a unit (II.1.1.3) have to be verified.

Also, the place of installation, i.e. whether it is a roof-mounted or ground-mounted system will be verified. Specifying the usage of the building is mandatory for roof-mounted systems and the size of the area covered for field systems. Such knowledge can be helpful, i.e. to support the analysis of land cover class distribution. However, both fields are not included in the QA procedures of a DNO.

Other system properties like amount of modules, main orientation and tilt an-

gle have to be specified but will not be checked. The size of a solar unit as well as amount of modules are directly related to the nameplate capacity, wherefore both properties can be used for capacity inference. Knowing the number of modules forming a solar unit is important for equipment procurement and maintenance. It allows project developers and investors to plan for regular maintenance and replacement of the modules over the life of the solar power plant. The orientation and tilt angle of solar PV units affect the actual yield that a system is able to produce. Therefore, both parameters are relevant for modelling of realistic yields.

In contrast, the gross and net capacity will be verified as well as whether the full or partial amount is fed to the grid. As described in Section 2.1.2, net capacity is calculated automatically during MaStR registration as the minimum of system and inverter gross capacity. However, inverter capacity is only validated for units which started operation after 06/30/2017.

Since only subsets of fields in the MaStR are validated by the DNO, this thesis examines the potential of validation and extension using a combination of image and building data. Distributions of corresponding fields and availability in external sources is discussed in more detail in Section 5.2.

3.2 Aerial Raster Images

This section describes key characteristics of the aerial imagery used and discusses limitations of an alternative dataset tested.

Date	Tiles			Images
	City	District	Total	Total
22/04/2020	-	6	6	6
25/04/2020	3	15	16	18
26/04/2020	78	87	130	165
24/06/2020	-	35	35	35
13/07/2020	-	8	8	8
31/07/2020	-	9	9	9
06/12/2020	-	17	17	17
Total	81	177	221	258

Table (3.3) Amount of images captured per day in city and district: The amount of area tiles covered per day is lower since images for city and district partially overlap.

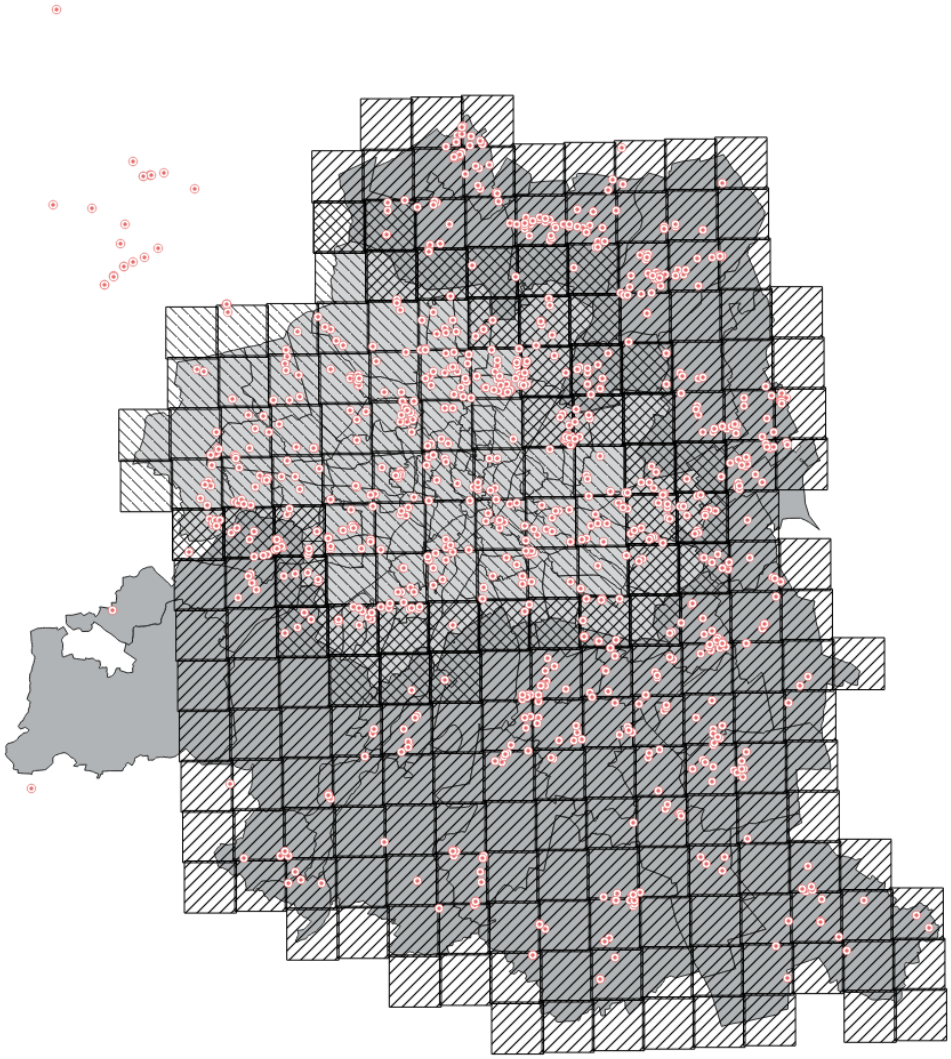


Figure (3.2) Spatial extend of the image dataset across zip code areas in the city (light gray) and district (dark gray) of Munich: Each square corresponds to an image tile. The red points represent (some) MaStR units with a capacity >30 kWp.

Munich Dataset The image dataset was provided to fortiss by the Bavarian Agency for Digitisation, High-Speed Internet and Surveying (LDBV). However, it is provided under restricted access conditions and not freely available to the public. It includes 81 RGB images covering the city and 177 RGB images for the district of Munich (see Table 3.3). As described in Section 2.2.2, they are stored as GeoTIFF files which allows referencing of the geospatial area displayed.

While the images of the city were taken on 04/25/2020 and 04/26/2020, the district dataset was collected between 04/22/2020 and 07/31/2020. Each image has a resolution of 0.2 m/px, which is below the minimum resolution of 0.3 m/px recommended by [33] for detection of PV systems.

With an extend of 12,060x12,060 px, each tile corresponds to an area of 5.82 km^2. Figure 3.2 displays the total area covered with each square representing an image tile. Overall, 34 tiles are represented redundantly by an image in both, the city and the district dataset.

OpenData Bavaria Dataset Since 01/01/2023, the LDBV started extending their OpenData platform. This initiative is a response to the Act governing the use of public sector data (DUA), which went into effect on 07/16/2021 [4]. On this platform, the institution published an image dataset representing the area of entire Bavaria. However, testing the algorithm described in Section 4.1.1 missed lots of solar units. This is due to the fact that the OpenData images have a resolution of 0.4 m/px while the models were trained on inputs of higher resolution. To scale the method presented in this thesis to Bavaria, a re-training on the new dataset would be required. Thus, the results presented thereafter only refer to the high-resolution dataset for Munich.

3.3 Building Data

The Open Street Map (OSM) project serves as a source of data on building properties. Its accessibility and data model are described in more detail in this section to emphasize the potential to not only validate but extend information given in the MaStR.

Data Access In general, OSM references a large variety of real-world entities like railways, streets or company locations and corresponding attributes. Data is extended by an ongoing global crowdsourcing effort which is supported by single volunteers or local communities focusing on data collection. Open-licensed data from national institutions is used to further increase its extent and quality. The

Figure (3.3) Example of overlapping OSM building polygons: Areas in lighter orange represent building polygons, overlaps are highlighted in darker orange.

combined OSM datasets are available according to the Open Database License (ODbL 1.0) on various platforms and levels of processing [9].

Raw OSM data is commonly distributed as Protocolbuffer Binary Format (PBF) files for efficient serialization and storage. The python library pyrosm is well suited for processing of large regional extracts [40] and was used in this research project for download of OSM extracts. It creates a combined dataset from the providers Geofabrik GmbH [21] for regional and national and BBBike [38] for urban extracts. Both sources update data on a daily basis (see Figure 3.1).

Next to the raw OSM extracts, I used the official OSM Nominatim API which i.e. allows querying of OSM objects by address [10]. This functionality is applied to retrieve geocoordinates of objects assigned to a given address string (see Section 4.3.1). For handling of requests to the API, the python library GeoPy is used [8].

Data Model for buildings According to the OSM Documentation, a building is characterized as "a man-made structure with a roof, standing more or less permanently in one place" [11]. As an entity, it references both, individual and groups of buildings.

On the one hand, building data is provided as multipolygons describing their ground layout (see Section 2.2.2). Geometries in this grouping may represent the outline of all units belonging to this building as well as holes in the layout like courtyards. Additionally, polygons of adjacent buildings might overlap as shown in Figure 3.3. Each building entity is uniquely identifiable by an id assigned by OSM. For Munich, a total of 303,345 ground polygons are provided, 59% spread across the city and 41% across the district.

On the other hand, additional attributes are given as key-value-pairs called

tags. These include for example information on the type of building or usage purpose. With regard to combination with the MaStR, such details could support the analysis of current expansion of PV units, i.e. across public facilities like schools. Moreover, it can support estimation of expansion potentials, i.e. the combination of charging stations for electric vehicles with PV collectors on parking lots.

Tags can be grouped into namespaces, i.e. all fields defining the address, roof or architectural properties of a building. In the scope of this thesis, only a selection of fields from the address and roof namespace with equivalence among MaStR fields or potential for its extension are considered. For future research and extension options, a full overview of available tags and definitions can be found on the corresponding OSM wiki pages [12].

3.4 Geographic Reference Data

To filter all data sources on the same geographic reference area, geospatial information on zip codes is used. After delineation for referencing by AGS, the used reference dataset is described in more detail in this section.

Next to zip codes, the MaStR provides information on the Official Municipality Code (AGS). In general, the AGS serves as another reference for demarcation of areas and is especially used by state authorities and institutions. Zip codes and AGS reference similar but not necessarily identical geographic areas. A zip code boundary might overlap with several AGS boundaries and vice versa. Additionally, the city of Munich is referenced by a single AGS code but multiple zip codes. Since zip codes are the more granular level for the given use case in Munich city and district, I will limit reporting of aggregates to this identifier.

The dataset for zip codes is downloaded from the Esri Germany Open Data Portal [14]. It comprises OSM polygons of zip code area boundaries and corresponding population figures. Overall, there exist 74 zip codes in the area of Munich city and 28 in the district. Their geospatial distribution is shown in Figure 3.2. The Federal Agency for Cartography and Geodesy also provides an official zip code dataset based on records of the German Post Direct GmbH. However, it is only accessible to federal authorities and authorized users according to Section 4 of the Contract on the continuous transmission of official digital geodata of the federal states for use in the federal area (V GeoBund) [19]. Therefore, the validity of the geospatial reference areas given in OSM can not be further verified in the scope of this thesis.

Chapter 4

Methodology

This section presents the methods for extraction and consolidation of information on PV systems. A visualization of the methodological steps is also provided in Figure 4.1. The deep learning and corresponding post-processing techniques for extraction of solar systems from images are described in Section 4.1. Inference of system properties is explained in Section 4.2. Finally, Section 4.3 presents approaches to create correspondences between the different data sources described in Chapter 3.

4.1 Detection of PV Systems in Images

To begin, this section describes characteristics of training and applying models for classification and segmentation of PV systems to the given image dataset (Section 4.1.1). Subsequently, methods for extracting localizable systems are presented (Section 4.1.2).

4.1.1 Classification and Segmentation of PV System Area

Model Training To detect solar units, the Computer Vision (CV) technique of semantic segmentation is applied. Segmentation is defined as the pixel-wise assignment of category labels. Given the task of solar panel detection, the model returns a binary mask with "0" identifying pixels in the background and "1" for pixels of solar panels. In general, segmentation can be carried out as instance or semantic segmentation. In contrast to instance segmentation, semantic segmentation does not differentiate between instances of the same class. Multiple objects are considered as one homogeneous rather than distinct entities. Differentiation of individual PV systems is carried out in a separate postprocessing step described in

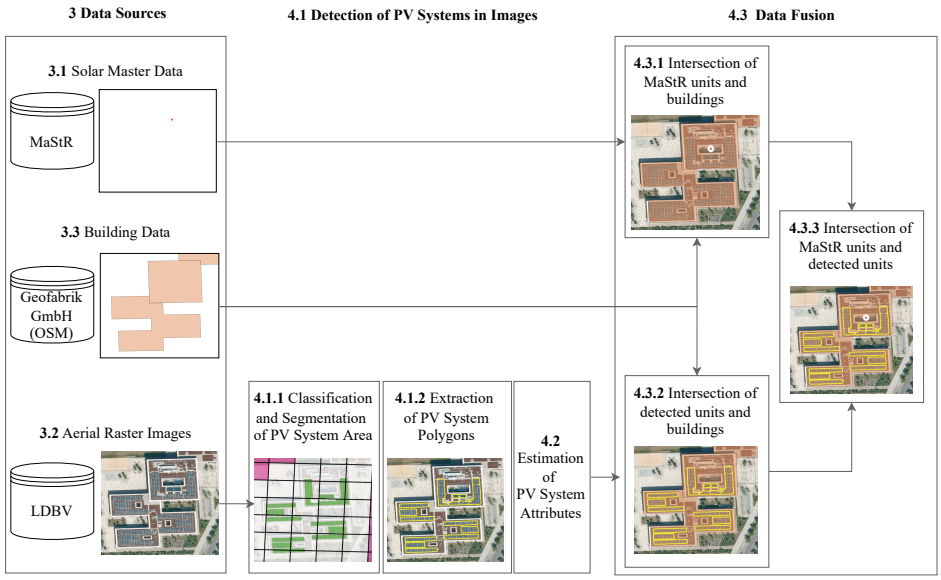


Figure (4.1) Overview of methods and corresponding data sources: 4.1.1 shows the combined results of classification and segmentation for each image crop (black squares): Negatively classified crops (pink areas) are not segmented. Instead, green areas represent high and white areas a low probability of solar panels. After binarization, system polygons (yellow) are extracted (4.1.2) and attributes estimated (4.2). Finally, MaStR units (red points), detected polygons and building polygons (orange areas) are combined (4.3.1 - 4.3.3).

subsequent sections.

The models for classification and segmentation are provided by another thesis student at fortiss GmbH [16]. In this part of the project, only inference is performed and adjustments are made to results only. Parameters are fixed and no further training is carried out. Therefore, presentation of model architecture and training is limited to relevant aspects. Further details can be found in [16].

The model was trained in a fully-supervised manner and comprises two subsystems for classification and segmentation. The classifier performs pre-selection of images such that only those including solar panels are segmented. It is based on the architecture of a ResNet34, which was pre-trained on Imagenet. The segmenter relies on the same network architecture as backbone to a U-Net.

Both models are trained on a combination of two labeled datasets: An extract of 18 images of Munich as described in Section 3.2 and 601 images of four cities in the US [1]. While the US dataset already contained labels, ground truth segmentation masks for Munich were created at fortiss GmbH.

The classification and segmentation network expect input tensors of size 224x224 px. Since images in the Munich dataset are of size 12,060x12,060 px and of size 5,000x5,000 px for the US, inputs have to be cropped. Positive samples are created by extracting windows centered at solar panels. Additionally, negative samples are sampled randomly and may or may not contain panels. Note that [16] does not specify the ratio of positive to negative samples. In total, 5,724 crops of Munich and 53,762 of the US were generated.

Since the last layer of the segmentation network contains a sigmoid activation function, the output masks yield the pixel-wise probability of a solar panel. A sample probability distribution for a single image is shown in Figure 4.2. In contrast, the ground truth masks are binary arrays with a value of "0" identifying background pixels and "1" for pixels of PV systems. To compare both masks, the predictions are binarized with a treshold of 0.5: If the predicted probability is equal or less than 0.5, the pixel is assigned the class value of "0" and "1" otherwise.

For testing, only samples from the Munich dataset were used. According to the 80-10-10 train/test-split, this resulted in 561 image crops for testing. The remaining images were divided into training and validation sets accordingly. Performance of the segmentation model is measured with the dice coefficient:

$$Dice = \frac{2 * |X \cap Y|}{|X| + |Y|} \tag{4.1}$$

It evaluates the pixel-wise agreement between predicted and ground-truth segmentation masks for background and panel class. On the test set, the given segmentation model achieved a dice coefficient of 0.952 when trained on the Munich dataset and of 0.75 if trained on the US dataset.

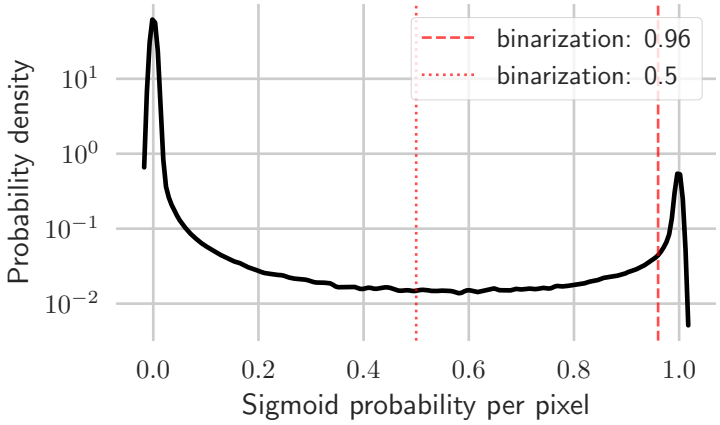


Figure (4.2) Probability density in output masks returned by the segmenter for image NO00403 (black curve). The vertical red lines represent low (0.5) and high (0.96) thresholds for binary classification of pixels as PV panel (1) or background (0).

Inference To compare detection results to entries in the MaStR, classification and segmentation models are run on all images in the Munich dataset. Image tiles covered redundantly by the subsets for city and district are only considered once (see Section 3.2). In total, 221 distinct images are passed to the segmentation pipeline.

The centering on positive detections during training causes the models to favor panels in the center and miss panels closer to the image borders more often. Therefore, the crops extracted from the source images need to overlap to achieve sufficient coverage. This is achieved by using a sliding window with stride 112, so offsetting the next window to be cropped by only half the input dimension. For a single image, this results in 11,664 crops which is four times the amount compared to using a stride of 224.

To reduce segmentation efforts, all images are passed to the classifier, but only those positively categorized as containing solar panels are forwarded to the segmenter. During development of the detection pipeline, results generated by the segmenter have been found to include far more noise if pre-selection is eliminated.

During model training, the sigmoid probability masks returned for each image crop are binarized with a threshold of 0.5. However, after running the models on the entire Munich dataset, a manual validation of the binary class distribution and source images showed lots of false positives for solar panels. A false positive in this context defines an area that is positively classified but visually identified as not

(a) source image (b) binary segmentation masks

Figure (4.3) Comparison of segmentation masks created with different binarization thresholds: Black lines represent boundaries of image crops, pink areas are not segmented. With a lower binarization threshold of 0.5, a larger (blue) area is considered as PV systems than with a threshold of 0.96 (green area).

being part of a solar PV system. However, the actual false positive rates cannot be determined computationally due to the lack of extensive ground truth labels.

After testing various cut-off values, a higher probability of 0.96 has shown to yield more reliable results (Figure 4.2). As shown in an example extract in Figure 4.3, requiring a higher confidence eliminates erroneous classifications of other roof structures shown in blue (4.3b). However, it might also slightly decrease the panel area at its boundaries.

For comparison with the source image, the binary masks of image crops are recombined into a raster of size 12,060x12,060 px and converted to a GeoTIFF file. For overlapping crop areas, the element-wise maxima are selected.

4.1.2 Extraction of PV System Polygons

Localization of panels is performed by extracting their boundary polygons from the recombined binary masks. The python package rasterio [35] provides necessary functionality and identifies panels as continuous areas of positive detections. To reduce false positives, very small polygons are filtered out. Only detections larger than the current average module size of 1.7 m^2 are considered for further analysis. The same postprocessing was applied by [28] and [36]. While the former also set 1.7 m^2 as a threshold, the latter suggest a value of 5 m^2.

Additionally, intersecting detections in areas of overlap between crops of dif-

ferent images are unified into a single polygon. This prevents them from being considered multiple times i.e. for area and capacity estimation.

4.2 Estimation of PV System Attributes

To validate reported MaStR data, its values must be compared to the attributes of detected systems. This section explains the calculation of central characteristics like system area size (Section 4.2.1), number of modules (Section 4.2.2) and capacity (Section 4.2.3) of detected systems.

4.2.1 System Area Size

Knowledge of the 2D area size $\hat{A}$ of a PV system is the basis for inference of system properties like amount of modules (Section 4.2.2) or capacity (Section 4.2.3). It can be inferred from the detected polygons using the PostGIS method *ST_AREA*. Due to tilting, the panel collector area is usually different from the ground area covered by the installation [26, p.430]. Deriving an area estimate from orthorectified images is likely to underestimate the true size and fall somewhere between both area parameters. However, this inaccuracy is assumed to be negligible for an approximation of PV system attributes in this research.

4.2.2 Amount of modules

Currently, a single module typically has a size of 1700x1000 mm and yields about 0.35 kWp [17]. To achieve a yield of 1 kWp thus requires about three to four modules. However, module sizes varied a lot over time, i.e. in the 1990s dimensions the standard dimension was 1200x600 mm while yielding only about 90 W [17]. Since then, modules became larger and their yield to area ratio increased. Current guidelines of the German Institute for Building Technology (DIBT) only allow solar PV panels of up to 2 m^2 on roofs with a maximum tilt of 75°. In contrast, solar thermal collectors can have a maximum size of 3 m^2 (see Section 2.1.1).

An analysis of the temporal expansion of solar units in Munich in Section 2.3.1 showed that a large fraction are older installations and thus most likely consist of smaller panels. The exact panel type, area per module (A^p) and yield per detection cannot be determined with the methods discussed in this thesis. To still cover a variety of panel types, the amount of modules per detection $\hat{P}$ are estimated for standardized smaller and larger panels according to Formula 4.2. It is directly inferred from the area covered by the detected polygon $\hat{A}$.

$$\hat{P} = \frac{\hat{A}}{A^p} \tag{4.2}$$

The lower panel count estimate $\hat{P}_{min}$ is determined according to [36]. For NRW, this study assumes an average area to capacity ratio of 6 m^2/kWp. They referenced standard panel sizes discussed on the webpage [17], but with a status from 2020. According to the current documentation from 2022, this corresponds to panels covering an area of $A^p_{min} = 1.6\ m^2$ and yielding about 250 Wp.

The amount of larger panels $\hat{P}_{max}$ are calculated based on current standard panel sizes of $A^p_{max} = 1.7\ m^2$.

4.2.3 Gross Capacity

In alignment with the estimation of the amount of modules, the capacity $\hat{C}$ per detection is inferred from the corresponding polygon area $\hat{A}$ according to:

$$\hat{C} = \hat{A} * C^p \tag{4.3}$$

C^p specifies the average nameplate capacity per square meter of a module. Since inverter capacities cannot be estimated using image and building data, the estimated nameplate power corresponds to the gross capacity. In alignment with the inference of module counts, a lower and upper capacity estimate are calculated based on: $C^p_{min} = \frac{0.25}{1.6} = 0.156\ kWp/m^2$ and $C^p_{max} = \frac{0.35}{1.7} = 0.206\ kWp/m^2$.

4.3 Data Fusion

The following section describes the methods for finding correspondences between the different data sources. First, MaStR units and detection polygons are mapped to OSM buildings separately (Section 4.3.1 and 4.3.2). Secondly, correspondences between MaStR units and detections are determined if both are located on the same OSM building (Section 4.3.3).

4.3.1 Intersection of MaStR units and buildings

As described in Section 2.3.3, geocoordinates and address information in the MaStR are only accessible for units with a capacity of at least 30 kWp. Therefore, the methods described in this Section only apply to larger units aimed at finding entity-to-entity correspondences between MaStR and OSM. Although only roof-mounted

systems should have correspondences with buildings, the MaStR dataset is not prefiltered to validate information reported on the system installation environment (see Section 5.2.1).

Four different join types on either the geocoordinates or address are implemented. This allows to analyze the accuracy of both types of geospatial information. Additionally, it helps to find additional correspondences. On the one hand, coordinates might have been located on the correct property but not on the respective building. On the other hand, geocoordinates can be easily misconfigured during the MaStR registration process which is harder to identify than an incorrect address text: The MaStR registration interface provides the option to calculate the geocoordinates automatically from a given address. Alternatively, they can be assigned by choosing a point in a provided map. When using a third party tool, they need to be entered manually in a decimal format. In general, entering an incorrect digit in one of the higher decimal positions might already induce a difference of a couple of 100 meters.

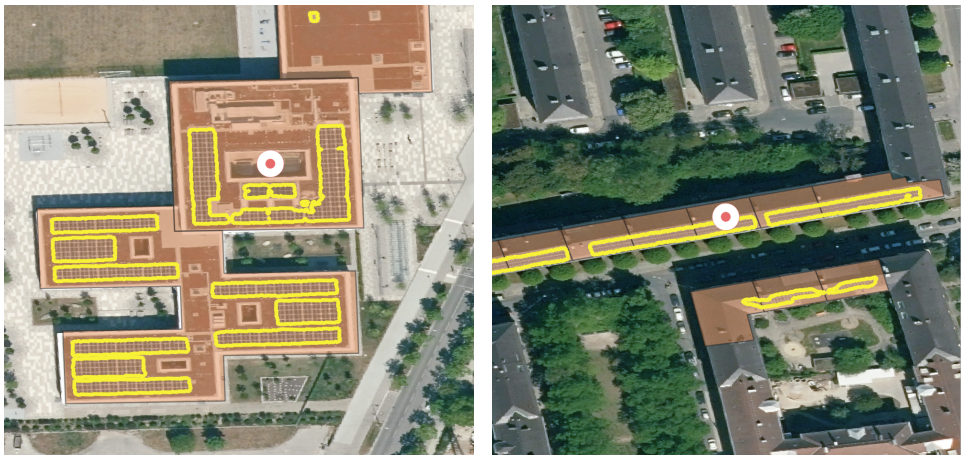
In the optimal case, a correspondence between a MaStR unit and a building is detected by multiple join types. This case confirms that geocoordinates and address given in the MaStR reference the same entity.

Independent of the join type, an n:n relationship can exist between entities in the MaStR and OSM. This is for two reasons: Polygons of different buildings can overlap as discussed in Section 3.3 and shown in Figure 3.3. Additionally, units mounted to the same building might have been installed at different points in time leading to separate registrations. Subsequent paragraphs describe the join types compared in this thesis in more detail.

Spatial join by given geocoordinates The geocoordinates given as tuples of latitude and longitude in the MaStR are converted to point geometries. A correspondence between a solar unit and a building is identified if these points are contained by the OSM polygon.

Join by address text As described in Section 3.3, OSM provides an address namespace next to building polygons. This allows joining both datasets on the text fields street name, house number and zip code. Since it does not consider coordinates, it has the advantage of not depending on the exact positioning of MaStR points in relation to building polygons.

Spatial join by geocoded address OSM offers the Nominatim API for geocoding which is the encoding of addresses into geographic locations (see Section 3.3). This provides the option of retrieving a point geometry corresponding to a given



(a) separate detections on adjacent roofs (b) detection over multiple adjacent roofs

Figure (4.4) Examples of different types of adjacent buildings and detections: While in (a) and (b) the red MaStR points are located on only one building, the unit attributes correspond to a grouping of detected systems (yellow) on adjacent roofs (orange areas).

street name, house number and zip code which can be tested for correspondence with building polygons. Correspondence is given if the address point is positioned within the building polygon.

Spatial join by reverse geocoded coordinates Reverse geocoding describes the conversion of geocoordinates to addresses. On the one hand, this allows direct comparison of the given and retrieved address text. On the other hand, it generates additional correspondences if the MaStR geometries are not positioned within building polygons. Overall, it should yield results similar to the direct join by MaStR geometries.

4.3.2 Intersection of detected units and buildings

To identify roof-mounted units, all detections intersecting with a building polygon are unified into one multipolygon per building. On the one hand, this approach is recommended by [36]. On the other hand, it is based on the MaStR guideline described in Section 3.1: All panels on the same roof should be considered as the same solar unit if commissioned together. Units starting their operation at a later point in time have to be registered separately. Since the aerial images only reflect a snapshot from 2020, exact commissioning of units cannot be determined. Thus,

this analysis is limited to a 1:1 relation between roofs and units. Additionally, the difference between the ground area of a building and its roof area is assumed to be negligible.

However, a panel polygon can reach across multiple roofs of adjacent buildings as shown in Figure 4.4b. Adjacency in this context refers to buildings with touching or even intersecting boundary polygons. In contrast, the terminology of neighboring buildings is used in this thesis for buildings located in spatial proximity i.e. on the same property but without such polygon contacts.

Adjacent buildings might contain distinct polygon units (see Figure 4.4a). According to the MaStR definition, these could still be reported as one solar unit. Therefore, two types of assignments are considered: A direct assignment of a detection to a single building as well as an indirect assignment to a grouping of adjacent buildings.

Overall, the grouping of detections by buildings helps to remove noise among the detections. On the one hand, it enables a direct correspondence search among large MaStR units which are located on the same building or adjacent building group. This approach is discussed in Section 4.3.3. On the other hand, it allows aggregation of estimates like capacity across all detected units per zip code area no matter the capacity limit. These estimations can be compared to the aggregates reported for building-mounted systems in the MaStR without requiring unit-to-unit correspondences.

4.3.3 Intersection of MaStR units and detected units

Correspondences between MaStR units and detected units are determined by overlaps in single buildings and adjacent building groups as defined in Section 4.3.2. This two-way map reduces the dependency of the correspondence search on the exact positioning of the MaStR points for units grouping detections on multiple adjacent buildings.

Although images were captured between April and July 2020, all MaStR entries registered until the latest extraction in February 2023 are considered for identifying correspondences with detections. Thereby, the operating status of units which have been planned but not yet commissioned at the time of image capture can be identified, too.

Chapter 5

Experiments and Results

This chapter describes the performance of the methods presented in Chapter 4 applied to the datasets described in Chapter 3. After evaluating the results of detecting PV systems in images (Section 5.1), Section 5.2 presents an assessment of data quality for a selection of MaStR fields. Based on these findings, Section 5.3 evaluates the potential of extending the MaStR with information from external sources.

5.1 Evaluation of PV System Detection

In addition to evaluating the effect of model architecture and hyperparameters on the detection of PV systems, Section 5.1.1 performs a sensitivity analysis of results with respect to visual characteristics. Results and limitations of PV system polygon extraction are discussed subsequently in Section 5.1.2.

5.1.1 Evaluation of PV System Area Classification and Segmentation

As described in Section 4.1.1, 221 distinct images across city and district of Munich are passed to the segmentation pipeline. No detections could be derived for 10 images, all of them covering areas of the Munich district. A visual analysis confirmed that these show mostly forests, lakes and farmland. Additionally, no building references can be found in the OSM dataset for 4 of the corresponding area tiles.

To reduce segmentation efforts, only image crops positively classified as containing solar panels are passed to the segmentation model. Figure 5.1 shows an example of an image with segmented crops highlighted in grey. The classifier correctly excluded the building and panel-free area of the English Garden in the center top as background. The majority of forwarded crops come from areas with lots of

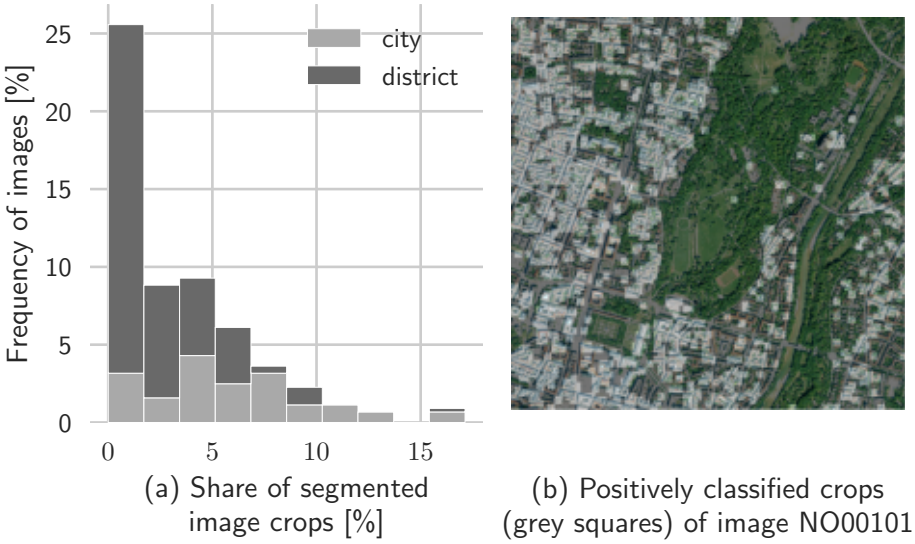
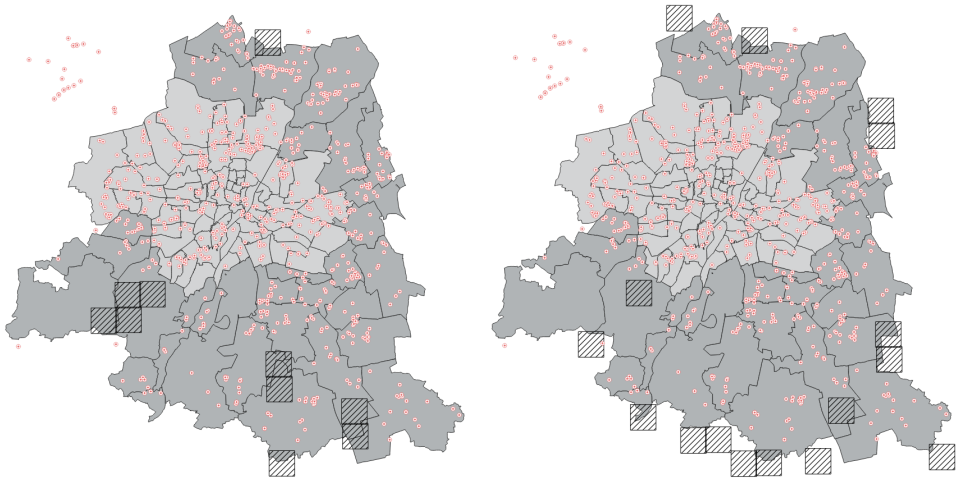


Figure (5.1) Effect of the classifier on the segmentation process: Only a small fraction of image crops is forwarded to the segmenter (a) as shown for sample image NO00101 (b)



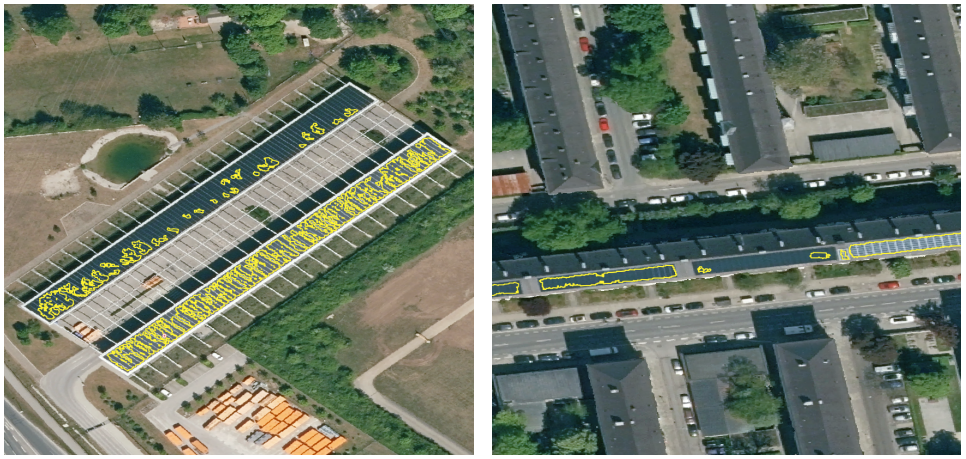
(a) w/o detections (b) w/o OSM building references

Figure (5.2) Location of image tiles (black squares) without detections (a) and/or OSM buildings (b) in regard to zip codes of the city (light gray) and district (dark gray) of Munich. Solar MaStR units are represented as red points.

(a) bright roof w/o special structure (b) dark roof w/o special structure (c) dormer

(d) patterned roof (e) patterned glass roof (f) greenhouse

Figure (5.3) Examples of False Positive Detections (yellow polygons): The segmentation model shows a sensitivity to roof structures with similar rectangular patterns or colours as PV Systems.



(a) dark panels w/ visible delimiters (b) new panels w/o visible delimiters

Figure (5.4) Examples of False Negative Detections: The segmentation model is less confident about predicting areas with few visible delimiters between panels as PV Systems (yellow polygons).

buildings. Considering the entire dataset, there is no image where more than 17.5% of crops are forwarded by the classifier (see Figure 5.1). Especially in the district, a high fraction of images has only few windows with positive classifications. On the contrary, in the city subset a slightly higher share of image crops is considered to include solar panels.

During training and testing, the fully-supervised model trained on the US dataset performed extremely well and achieved a precision of 0.989 and recall of 0.941 when tested on extracts of the Munich dataset [16, p.29]. Since the Munich dataset is only partially labeled, this work cannot provide exact performance metrics for classification and segmentation. Instead, it focuses on extending performance results presented in [16] by manual spot checks in order to highlight strengths and limitations of the given models in regard to identification of certain structures.

For example, Figure 5.3 presents examples of manually identified false positives. Classifier and segmenter show sensitivity to roof structures of colours typical for solar panels: a dark blue-grey (5.3b, 5.3c) as well as white (5.3a, 5.3d). Additionally, it struggles to distinguish between solar panels and rectangular roof structures of similar size like windows or glass roofs of squared tiles i.e. as found on greenhouses (5.3f) or the Olympiahalle (5.3e). A similar problem was reported by [39] who emphasize the importance of high quality training datasets covering such complex distinction cases.

Another type of false positive would be a confusion of solar PV panels and solar thermal collectors. However, based on images at the given resolution, these panel types are almost impossible to distinguish even for human annotators. Therefore, based on the given data sources, no statement can be made about the possible extent of this confusion. However, [39] reported that by adding respective samples to the input dataset for their detection model, they were able to automatically distinguish both system types reliably.

While false positives can cause overestimation of properties like system area and thus capacity, false negatives can lead to underestimation. As shown in Figure 5.4, it seems to be more complex for the segmenter to identify dark solar systems compared to brighter ones with clear delimiters. “Full-black” modules are relatively new on the market and use a black instead of a white protective foil which eliminates the typical grid pattern between the individual cells and on the frame. Overall, this results in a uniform dark appearance of the entire solar unit [18]. As shown during discussion of false positives, they might be easily be mistaken for roof structures with similar appearance like dormers (see Figure 5.3c).

During post-processing of detections, the main parameter affecting the occurrence of false positives and false negatives is the binarization threshold: Keeping areas the model is less confident to have classified correctly as a solar system is likely to increase the false positive rate while decreasing false negatives. Since

labels are given for a subset of images only, exact numbers for both groups cannot be determined for the full image dataset. However, the overall effect on the total extend of detections is quite significant: The detected polygons cover an area of 1.21 km^2 with a cut-off at 0.96 compared to 2.28 km^2 for a value of 0.5. In the scope of this thesis, the threshold of 0.96 is determined by manual spot checks and analysis of confidence distributions for sample images (see Figures 4.2 and 4.3). In contrast, [36] included optimized for the optimal threshold during hyperparameter tuning of both, classifier and segmenter.

5.1.2 Evaluation of PV System Polygon Extraction

As discussed in Section 4.3.2, detected panel polygons enclosed by a building outline are unified into a multipolygon. Additionally, a panel polygon might reach across multiple buildings and will then be assigned to each them. The extraction of polygons yields a total amount of 21,530 distinct polygons for solar systems and 22,410 unique buildings these can be assigned to. As mentioned in Section 5.1.1, this corresponds to a detected system area of 1.21 km^2. In contrast, an area of about 1.0 km^2 cannot be mapped to a building.

Given the definition of solar units in the MaStR (see Section 3.1), the amount of detected polygons or buildings cannot be expected to be equivalent to the amount of solar units reported in the MaStR. On the one hand, detections would have to be separated into units which started operation at different points in time. On the other hand, it is unclear which units of adjacent or neighboring buildings need to be grouped to a single unit because they were commissioned together (see Figure 5.11).

Given the grouping of polygons by buildings, it should be mentioned that no OSM buildings are given for 16 image tiles in the district. Four of them additionally do not yield detections. However, those are mostly tiles at the district boundaries that only partially overlap with the district zip code areas (see Figure 5.2b). Since buildings are pre-filtered on these boundaries, no correspondences can be rightfully determined for image tiles exceeding these.

5.2 Validation of selected MaStR fields

The following section discusses the results of validating the MaStR for a selection of fields individually: Initially, an exploratory data analysis is applied to identify data quality issues without using external information. It is referred to as internal validation. Secondly, the results of validation using image and building data are presented which is performed in two steps: The first approach examines mappings

of either detections or MaStR entities to OSM data and only considers aggregates either by city and district or more granular by zip codes areas. The term mapping is used to describe a correspondence of entities from two data sources created by one of the discussed join types. The second approach checks if a mapping between individual MaStR units and detected systems can be considered correct. This is the case if both attribute sets overlap which is defined as a match. However, only capacity is considered for determining matches since it is included in DNO validation which makes this attribute more reliable than i.e. amount of modules.

For some evaluations based on external data, only detectable systems are considered: For aggregate considerations, a solar unit is defined as detectable if it has been commissioned on or before the latest image capture date on 07/31/2020 and is not being registered as a field system. For unit-level evaluations, this definition is extended by the unit exceeding a net capacity of 30 kWp.

5.2.1 Place of installation

	(1) ≤30kwP		(2) >30kwP		Total	
	U	[%]	U	[%]	U	[%]
(1) BF (Roof, Building, Facade)	15,567	88.01	949	5.37	16,516	93.38
(2) BF (Misc)	118	0.67	13	0.07	131	0.74
(3) Ground-mounted	9	0.05	11	0.06	20	0.11
(4) Plug-In	1,018	5.76	2	0.01	1,020	5.77
(5) No data	1	0.01	nan	0.00	1	0.01
Total	16,713	94.50	975	5.51	17,688	100.01

Table (5.1) Amount of units by place of installation and system size: U=absolute frequency, [%]=relative frequency, BF=building facility.

Internal Validation As discussed in Section 3.1, it is mandatory to register information on the place of installation which has to be verified by the DNO (see Table 3.2). This field is especially relevant in order to determine the fraction of units that is detectable using image and building data. In the Munich dataset, it is missing a value for only a single small unit.

On the coarsest level, units can either be ground-mounted field systems or mounted to a building facility. Extraction of systems which are not mounted to a building requires a different handling compared to roof-mounted systems. Since only 9 small and 11 large units in Munich are field systems (see Table 5.1), this thesis prioritizes roof-mounted systems.

In contrast, the majority of 93.38% of units is installed on a roof or facade as presented in Table 5.1. Another 0.74% of systems are reported to be mounted to a building without their exact mounting location being further specified. According to the field on system tilt angles, 1.23% of all units are facade-mounted systems (see Table 5.7). Consequently, only a small fraction of systems should be expected to be undetectable in the given aerial imagery.

About 5.76% of systems are small units registered as so-called Plug-in systems which can for example be mounted to the balcony. According to the standard for low-voltage networks (VDE-AR-N 4105), those systems with a capacity below 600 watts can be installed privately. Otherwise, installation and acceptance by a specialist electrical company is required [32]. Currently, two units of this type with a capacity exceeding 30 kWp are listed. This indicates erroneous entries either for the capacity or place of installation.

Fusion of MaStR units and OSM buildings Among the 964 large, building-mounted units registered until February 2023, no mapping to an OSM building could be identified for 6.77% of systems. While this could indicate an error in the place of installation, it seems more likely that this lack of correspondence is caused by either an erroneous location or the MaStR point coordinates not being positioned exactly within an OSM building polygon.

In contrast, a correspondence to a building is found for 2 of the 11 large field systems. For the first unit, the mapping is based on the address, for the other it is based on reverse geocoding of coordinates. This finding aligns with the analysis of [36]: It seems to be common in the MaStR that addresses of field systems describe the owner residence instead of the location of installation. Consequently, it is probably rather the locational information that has been registered incorrectly for the two systems than the categorization as a field system.

Fusion of MaStR units and detected systems When considering correspondences to detected systems, only 7.88% of detectable building-mounted systems remain without a correspondence. For the 2 field systems mapped to building, no PV system could be detected on the corresponding roof. This supports the assumption that rather their location is erroneous instead of their general place of installation.

5.2.2 Building Usage

Internal validation Next to the place of installation, the MaStR provides information on the usage area of the building a unit is mounted to (see Table 5.2). However, these details are not further validated by the DNO (see Table 3.2) and are

	(1) ≤30kwP		(2) >30kwP		Total	
	U	[%]	U	[%]	U	[%]
(1) Private Household	13,463	76.11	84	0.47	13,547	76.58
(2) Public Facility	363	2.05	179	1.01	542	3.06
(3) Trade, commerce, services	524	2.96	405	2.29	929	5.25
(4) Industry	30	0.17	50	0.28	80	0.45
(5) Agriculture	311	1.76	127	0.72	438	2.48
(6) Misc	336	1.90	46	0.26	382	2.16
(7) No data	1,686	9.53	84	0.47	1,770	10.00
Total	16,713	94.48	975	5.50	17,688	99.98

Table (5.2) Amount of units by usage area and system size: U=absolute frequency, [%]=relative frequency.

missing for about 10% of installations. Overall, the majority of buildings are used by private households (76.58%), only 3.06% are public facilities. 2.48% of units are related to facilities with an agricultural purpose compared to only 0.45% which are mounted to industrial buildings.

41.54% of large units are related to buildings used for commercial purposes. On the contrary, the most common usage of buildings with small units are private households. Knowledge of this distribution is relevant since private operators may not have the necessary expertise or resources to accurately collect and report data. This could increase the likelihood of errors or inconsistencies. However, to verify this assumption and derive conclusions, i.e. on whether operators might require further assistance during data collection, would require a more detailed analysis on a per-unit level. As discussed, this is not possible in the scope of this research due to limited access to locational information of small systems.

Fusion of MaStR units and OSM buildings As mentioned in Section 3.3, OSM collects information on the type of building as well as its corresponding usage area.

A tag for the type of building is provided for 54.27% of entries. As shown in Figure 5.5, apartments, garages and houses occur most frequently. Since OSM datasets are generated by a hugh variety of contributors, data quality and standardization is still improvable. For example, there exist separate listings for “garage” and “garages” or generic descriptions like “house” are used. Overall, 142 distinct building types are specified for Munich, about half of them having less than 10 occurrences.

Additionally, the OSM tag on building usage purpose specifies 90 different values for Munich, for example “place of worship” or “parking”. However, this tag is only given for 0.8% of entities.

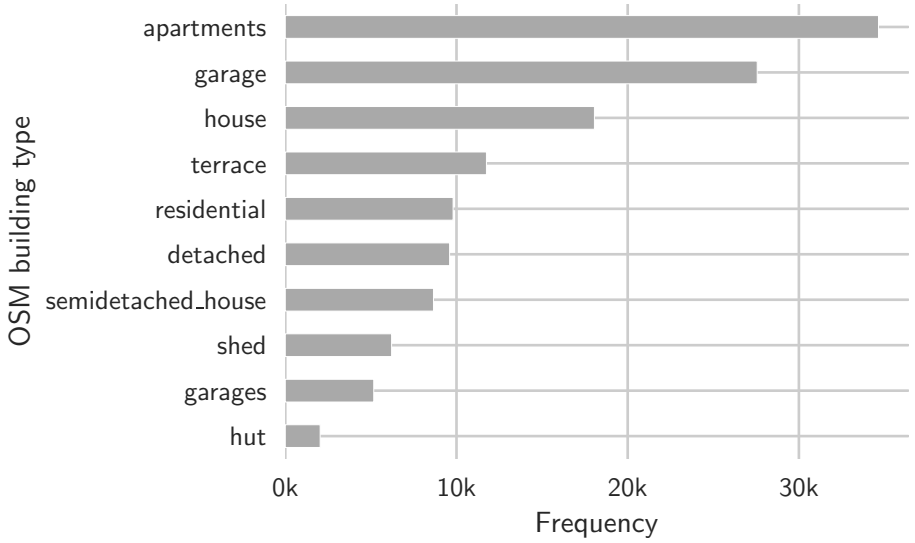


Figure (5.5) The frequency distribution of OSM building types (top 10) in the reference area shows low data quality: Values are not standardized (i.e. “garage” and “garages”) and generic types are used (“house”).

Consequently, both tags on building type and usage purpose do not provide sufficient coverage and data quality to validate or even extend the MaStR in a scalable manner. Additionally, to validate MaStR entries would require a preliminary clustering into corresponding groups like “Public Facility” or “Private Household”.

5.2.3 Location

Internal validation In the MaStR, locational information is provided as address text and geocoordinates. Coordinates as tuples of latitude and longitude are accessible for the 975 units which exceed the capacity threshold of at least 30 kWp (see Section 2.3.3). However, 34 of them lack a value for the street name and 49 for the corresponding house number. A zip code is provided for all systems. Consequently, only large systems with exact location given can be validated on a per-unit level using external sources.

The distribution of large systems across the geographic reference area according to their geocoordinates is visualized in Figure 3.2. Due to the filtering of MaStR entries by the text fields on city and district instead of zip codes or specified coordinates, some entries are kept with coordinates referencing areas way outside of Munich. Some points positioned even further away are not shown in the figure.

Given the reference data described in Section 3.4, 34 of the reported geocoordinates are not positioned within the corresponding zip code area of the unit.

Fusion of MaStR units and OSM buildings Considering the address namespace in OSM, street name, house number and zip code are given for 29.14% of buildings in Munich. This is especially relevant for the join by address text.

99.45% of addresses are unique, indicating that in some cases, the same address is assigned to multiple buildings. This can still be a valid assignment if buildings on the same property and with the same address are listed as separate entities. On the contrary, it can be considered erroneous if the zip code registered in the OSM tag does not match the zip code of the building polygon. However, this is the case for only 424 (0.48%) of these entries.

Additionally, there are 4,736 street and house number combinations which are related to more than one zip code. For example, the street name "Bahnhofstraße" can be found in up to 13 zip code areas. Therefore, information on the correct zip code is highly relevant to identify the correct building. However, for 7.24% of all buildings this information is missing. In total, this leaves lots of buildings without address information, but if data is given, it seems to be matching the given geometry.

Since only 11 large MaStR units are reported to be field systems (see Section 5.2.1) and various join types are implemented, it should be expected that correspondences can be found for most units. Overall, this assumption holds true for 93.23% of large units: Direct correspondences are found for 909 MaStR systems and 1,015 buildings. Due to the implementation of different join types, this results in 2,604 mappings. In the optimal case, a unit should be mapped to a single building with joins on address as well as geocoordinates. While this does not prove that the reported location of the system is correct, it indicates coherence in locational fields.

The co-occurrence of join types is presented in Figure 5.6. Each value in the heatmap represents the fraction of distinct mappings between an OSM building and MaStR unit which was found by a given join type. Generating 66% (792) of distinct mappings, the intersection of MaStR geocoordinates and building polygons yields the most correspondences overall. Of these, 29.29% of correspondences are unique, which is equivalent to a total of 8.91% of mappings which could not be identified by other join types (see Table 5.3). Since reverse geocoding is also based on geocoordinates, 45% of all distinct assignments between a unit and building are created by both these two join types. Additionally, the reverse transformation leads to 4.72% of unique mappings.

A reason for differences in results between both joins is the positioning of

point coordinates: If a point is placed close to but not inside a building polygon, a correspondence cannot be identified. Reverse geocoding identifies address points of the property that the MaStR coordinates are positioned on. Although these are retrieved from OSM, they do not have to correspond to a single building since there might be multiple on a property, i.e. the residence house and a garage. Overall, original coordinates and buildings found by their transformation are only 12 meters apart on average and 321 meters at most. Consequently, reverse geocoding can be applied to create alternative versions of MaStR coordinates and increase the likelihood of a correspondence with a building.

In contrast, the joins on address text and geocoded address are both based on address information instead of given coordinates. As discussed, this information is available for 94.97% of large MaStR units, but only 29.14% of OSM buildings in Munich. Correspondingly, both join types yield less correspondences overall compared to joins based on geocoordinates. Searching for mappings by address text yields only 1.34% of unique results compared to 4.26% for geocoded addresses. The geocoded addresses perform slightly better since they only require address texts to be given for MaStR units but not for buildings. Additionally, this join type shows a high number of co-occurrences with the spatial join on geocoordinates and reverse geocoded addresses. Again, this indicates coherence of the different locational information reported in the MaStR.

Overall, 19.24% of mappings are created by only a single join type. In contrast, address and coordinates yield the same distinct mapping for 533 out of the 975 large MaStR units. Among these correspondences, the assigned buildings are on average only 4 to 5 meters away from the originally registered coordinates (see Table 5.3).

In contrast, for correspondences without overlaps between joins based on address and coordinates this distance is significantly higher. For example, the mean distance between buildings assigned by the address text and reported MaStR coordinates is 10.3 ± 53.7 km.

Fusion of MaStR units and detected systems Images are used to verify that at the location given in the MaStR not only a building but more importantly the correct solar PV system can be found. However, to verify location in this regard requires checking for a match of corresponding attributes, i.e. capacity. Therefore, correctness of mappings is discussed in more detail in Section 5.2.4. This section focuses instead on preliminary analysis of the correspondences created with different join types. As before, mappings between MaStR units and OSM buildings and thus detections follow an n:n relation.

Of the 975 MaStR units with more than 30 kWp, 800 should be detectable

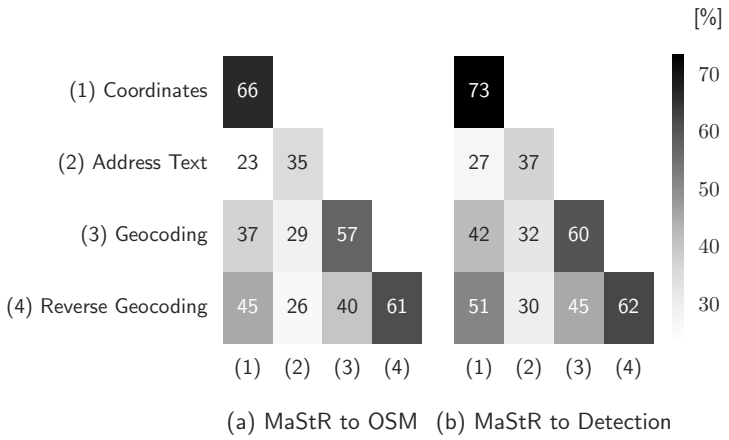


Figure (5.6) Co-occurrence of join types [%]: Labels for join types on the x-axis are shortened to their numeric identifier. Each value represents the share of distinct mappings between a MaStR unit and an OSM building (1) or detection (2) that were found by two join types. The identity diagonal represents the total fraction of distinct mappings created by a single join type. Numbers do not add up 100% since distinct mappings can be found by multiple join types.

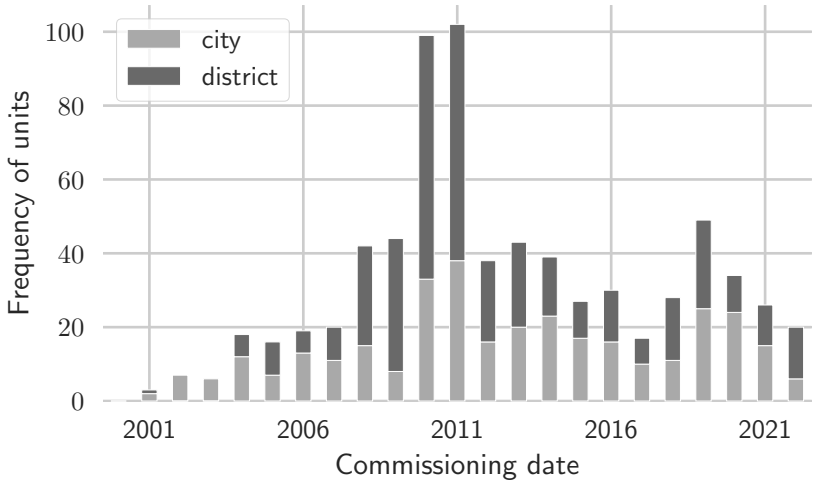


Figure (5.7) Amount of large units with detection correspondences commissioned per year

	Unique joins freq [%]	Dist. w/ loc. overlap [m] mean	std	Dist. w/o loc. overlap [km] mean	std
(1) Coordinates	8.91	-	-	-	-
(2) Address Text	1.34	4.33	17.52	10.33	53.73
(3) Geocoding	4.26	4.93	20.31	6.25	40.71
(4) Reverse Geocoding	4.72	5.18	20.61	0.03	0.05
Total	19.24	4.88	19.79	4.33	34.43

Table (5.3) Statistics for joins between MaStR units and buildings: Relative amount of mappings created by only a single join type and distance (dist.) between the original MaStR coordinates and mapped buildings for cases in which registered address and coordinates do or do not overlap.

in images since they are mounted to a building and have been commissioned until 07/31/2020. In total, 839 mappings between MaStR units and buildings with detection polygons are identified. Those involve 737 distinct MaStR systems of which 91.59% should have been detectable. The remaining systems have been commissioned after the latest image capture on 07/31/2020 and are thus considered as non-detectable. Depending on how much time elapsed between the date the images were taken and the actual start of operation, this does not necessarily indicate an error. For example, the MaStR allows registration of units while they are still under planning (see Table 3.1). However, 46 mapped systems started their operation in 2021 and 2022 which seems like quite a long time between installation and reported start of energy operation (see Figure 5.7). Although the given image data is insufficient to determine the correct start of operation, this finding could indicate that too few units are considered as detectable which could lead to overestimation i.e. of total installed capacity. Overall, no mapping could be identified for 15.62% of detectable large units.

The relative distribution and co-occurrence of join types among mappings between MaStR and detections resembles the one discussed for correspondences between MaStR and OSM (see Figure 5.6). This is expected since for 93.23% of large systems a building correspondence was found and due to their spatial extent, such systems are likely to be detected during classification and segmentation.

As discussed previously, join types may yield overlapping as well as different mappings. Consequently, 87 distinct MaStR units are assigned to up to three detections. However, the reverse relation is also found: 75 buildings have multiple MaStR units assigned to them. This can have two root causes: As defined in Section 3.1, modules installed at different points in time have to be registered as separate units even if they are located on the same roof. For each single unit, this

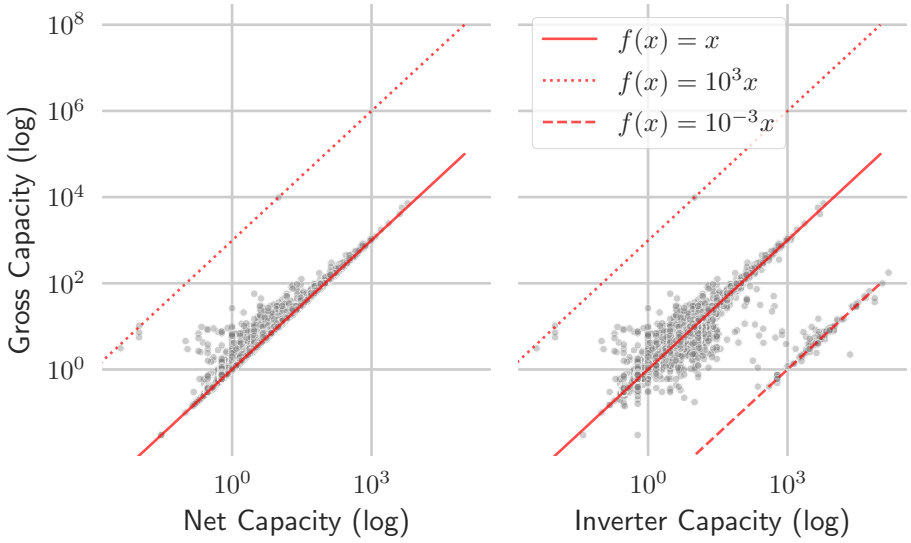


Figure (5.8) Relation between registered net, gross and inverter capacity per unit [kW]: Each grey point represents a single MaStR unit, points scattered around the dashed or dotted red lines indicate conversion errors between kW and MW.

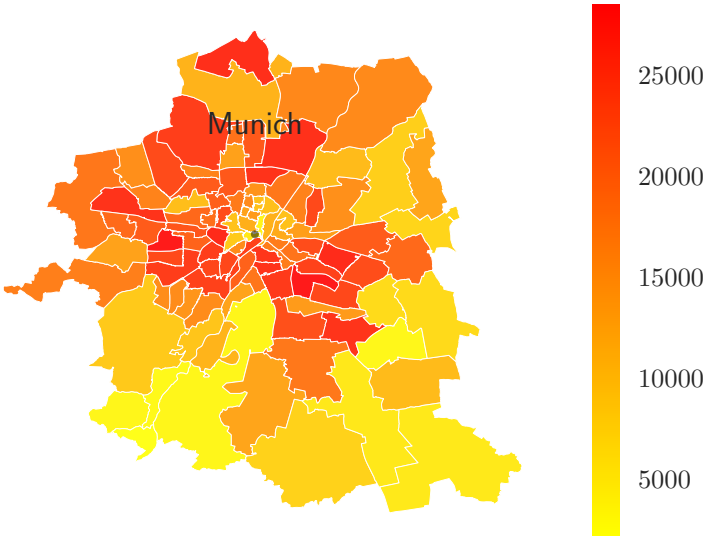
would lead to overestimation of capacities since the entire system per building is considered during estimation. Secondly, it seems to be common especially among solar system operators in the agricultural domain to register several systems on their residential building even if they are not located in its close neighborhood.

While the first case meets the definition of a solar unit according to the MaStR, the latter case would be an erroneous entry in regard to location. Consequently, the discussed join types cannot assign a mapping. This typically leads to underestimation of a reported unit's system area and corresponding capacity.

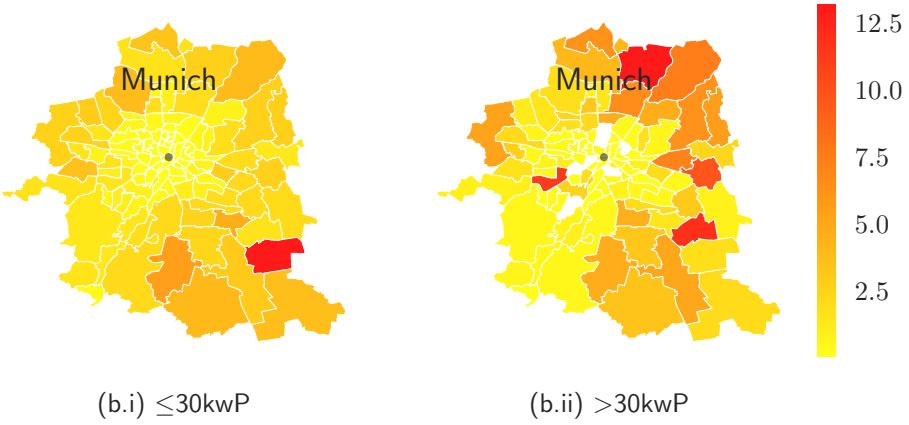
5.2.4 Capacity

Internal Validation As mentioned in Section 3.1, the gross nameplate power registered in Munich until the date of extraction in February 2023 accumulates to 270 MW. Since net capacity is derived automatically as the minimum of gross and inverter capacity, it sums up to only 238 MW. Consequently, net values can only be valid if those two entries are reported correctly and can never be higher than any of them.

Figure 5.8 analyses the relation between the three capacity fields per unit: For the gross and net capacity, scatter points not located on the identity line indicate that



(a) Population per zip code area



(b) Capacity per zip code area [MW] for small (i) and large (ii) systems

Figure (5.9) Comparison of population and capacity distribution for small (b.i) and large (b.ii) units per zip code area: White areas in (b.ii) indicate no allocated solar capacity.

the net capacity is limited by a lower inverter capacity. Moreover, some outliers are scattered around a line identifying a times 10^3 relation between net and gross values. This is most likely caused by a confusion of the units kW and MW. As discussed in Section 2.3.2, for gross and net capacities such errors should have been handled by the department MaStR-QS of the BNetzA.

However, those QA measures might not fully cover inverter capacities since this field is not included in DNO validation for units commissioned before 06/30/2017. Consequently, conversion errors occur more frequently for this field: A non-negligible fraction of entries has a registered gross capacity at only a $1/10^3$ of the inverter value. Since inverter capacity cannot be estimated from images, it is not possible to reproduce calculation of net values as in the MaStR. Consequently, subsequent validation will only reference gross capacities if not stated otherwise. Nonetheless, it should be mentioned that classification into small and large MaStR systems is based on net capacities while detections are grouped by estimated gross capacities. Although this handling introduces inconsistency, it is assumed to not change the results and corresponding interpretations significantly.

Overall, units with $>$ 30 kWp account for only 5.51% of installations, but contribute 47.8% of nameplate power. Among smaller units, the average net capacity ranges around 7.13 $\pm$ 5.67 kWp. In contrast, larger units yield 121.88 kWp on average, but capacity magnitudes vary heavily given the standard deviation of 307.77 kWp in this group.

Considering the distribution of capacity across the reference area (see Figure 5.9b), the number of inhabitants seems to serve as a reverse proxy for installed capacity: As shown in Figures 3.2 and 5.9a, zip code areas in the center cover the city and suburban areas of Munich. In contrast to the southern and northern zip code areas of the district, those regions are smaller in area size but also more densely populated. Only little nameplate power is registered in the city areas for both, small and large systems. The white spaces in Figure 5.9 even indicate that no capacity is allocated to the respective zip codes for large systems. The installed power of large systems focuses on the less densely populated district areas. In contrast, the capacity of small systems is quite evenly distributed across zip codes of the district, with a slight increase at its boundaries.

Fusion of detected systems and OSM buildings Since zip codes are given for all MaStR units, capacities are aggregated on this level and compared to estimated capacities of systems detected in images. This requires to map extracted polygons to buildings to reduce noise and limit detections to building-mounted systems.

The nameplate capacity $\hat{C}$ is approximated according to the formulas given in Section 4.2.3. One of the validation metrics used is the detection ratio ΔC which

		(1) $\leq$ 30kwP		(2) $>$ 30kwP	
		$\hat{C}$ [MW]	ΔC [%]	$\hat{C}$ [MW]	ΔC [%]
city	lower	41.12	113.47	55.65	138.40
	upper	47.41	130.82	80.09	199.18
district	lower	38.14	81.15	53.65	124.68
	upper	44.48	94.64	76.47	177.71

Table (5.4) Aggregate upper and lower PV capacity estimates ($\hat{C}$) and detection ratio (ΔC) by area and system size.

represents the fraction of the detectable capacity $C^{detectable}$ that is identified in the images:

$$\Delta C = \frac{\hat{C}}{C^{detectable}} \tag{5.1}$$

This metric is useful to determine the degree of under- and overestimation in correspondences. As discussed in Section 4.2.3, two capacity limits $\hat{C}_{min}$ and $\hat{C}_{max}$ are calculated. Therefore, the detection ratio is reported as a range defined by these thresholds thereafter.

For aggregate comparison, solar systems corresponding to a nameplate capacity of 166 MW should be detectable in the aerial image dataset. Overall, the panel polygons extracted from the images and mapped to buildings yield a capacity between 188.56 and 248.45 MW. This corresponds to a total system area of detected polygons covering an area of 1.21 km^2. Thereby, the capacity in Munich is overestimated by 13.26 to 49.23%.

As shown in Table 5.4, overestimation is especially high in the city of Munich and for large systems no matter the estimation threshold. This observation is visually observable by comparing detection ratios across zip codes in Figure 5.10. While detected capacities exceed reported values in 35% of district zip codes, this is the case for 81% of city compartments. In contrast, small systems in the district are even underestimated by 5.36 to 18.85%.

Given Equation 5.1, overestimation can have two main root causes: Limitations in the approach for system detection (Section 4.1) causing estimates $\hat{C}$ to be too high or too few units reported as being in operation in the MaStR and thus too little detectable capacity $C^{detectable}$. The research of [39] for Saxony reports a detected amount of units which is 2.5 times the amount reported in the MaStR. On the one hand, this finding supports the latter assumption if a higher amount of units is assumed to indicate higher installed capacity, too. On the other hand, the authors do not compare capacities and also did not align their frequency calculation with the definition of units in the MaStR.

Consequently, due to lack of extensive ground truth data for the reference area

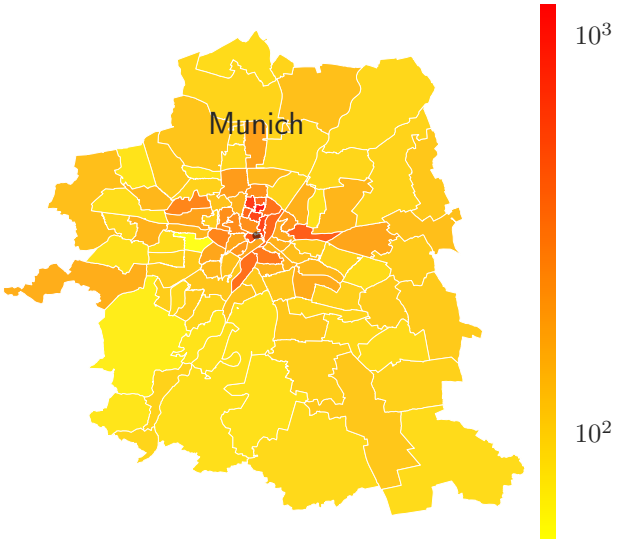


Figure (5.10) Detection Ratio ΔC [%] of all PV systems by zip code area (log scale colorbar): Capacity is heavily overestimated in the city center.

considered in this work, it cannot be determined how many or which of the detected systems are not reported in the MaStR on an aggregate level. A corresponding analysis is performed for unit-level correspondences in the next paragraph.

Analysing the first root cause instead, two shortcomings can be identified: Due to dense population and lots of buildings in the city area, the models are more likely to come across roof structures that can be mistaken for solar panels. Especially misclassifications of large glass roofs with rectangular structures like the Olympiahalle (see Figure 5.3e) quickly add up. In total, they may well offset the proportion of false negatives which are missed by classifier and/or segmenter.

Additionally, the segmenter detects coherent areas of panels wherefore distances i.e. between rows of modules are considered part of the system area $\hat{A}$, too. For example, "full-black" modules mentioned in Section 5.1.1 lack visible delimiters. Considering a large system with hundreds of panels, this might falsely increase the system area considered for capacity estimation.

Fusion of MaStR units and detected systems The 737 distinct MaStR units with a mapping to a detected system are reported to produce 80.54 MW and thus 96.75% of the detectable capacity in this capacity group. In contrast, the detected systems they are mapped to are estimated to yield a capacity of 61.01 to 80.39 MW.

(a) large unit across multiple factory buildings

(b) large unit across multiple buildings on same property

Figure (5.11) Examples of solar units distributed across neighboring buildings: The orange areas represent OSM buildings, the yellow polygons identify detected PV systems.

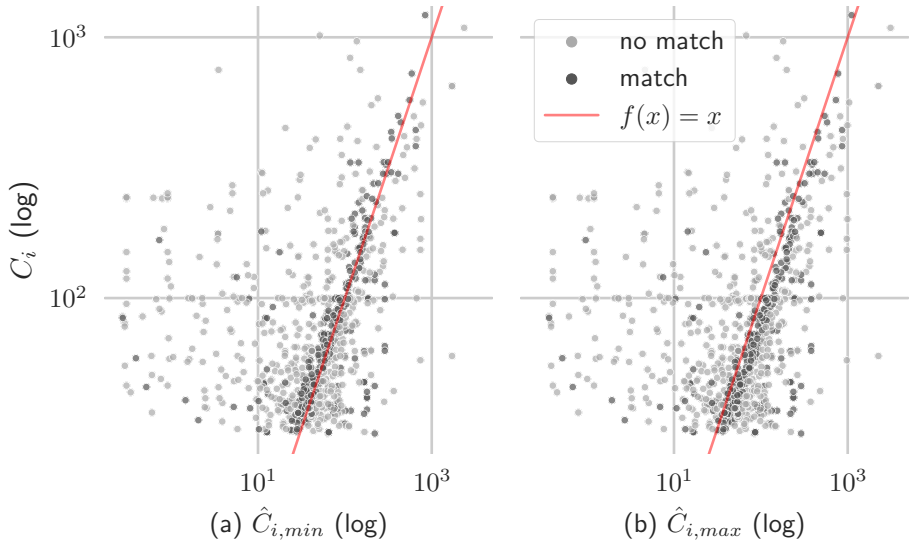


Figure (5.12) Alignment between reported capacity C_i and lower (a) as well as upper (b) capacity estimates $\hat{C}_i$ [kWp]: Each gray point represents a mapping between a large MaStR unit and a detected PV system.

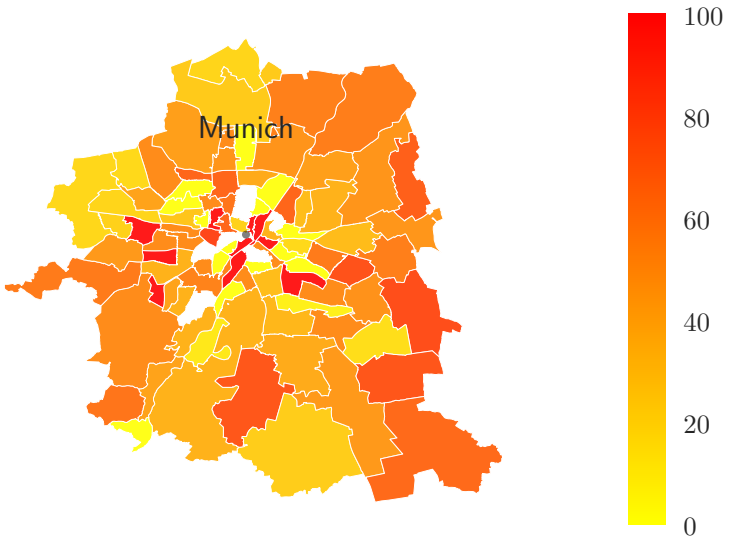


Figure (5.13) Match Ration ΔM [%] of large MaStR units mapped to a detected PV system by zip code area: Less matches are found in the city compared to the district.

Although this might indicate excellent alignment at a first glance, analysing detection ratios per correspondence shows that this relation is way more complex.

The alignment of estimated ($\hat{C}_i$) and registered (C_i) capacities per correspondence is shown in Figure 5.12. Points to the left of the identity represent correspondences with underestimated registered capacities whereas points to the right highlight overestimation. Comparing estimation limits, the lower threshold $\hat{C}_{i,min}$ tends to underestimate 62.41% of mappings. In contrast, assuming high-capacity panels $\hat{C}_{i,max}$ causes 56.69% of capacities to be over-estimated which shifts the distribution visually to the right. However, the lower threshold achieves a mean detection ratio of 101% compared to 133% for higher estimates. Consequently, approximation based on lower panel yields seems to fit the given units better on average.

The distribution scatters to the left of the identity for both limits which shows a tendency towards underestimation of reported unit capacity. Combined visual assessment of images, detections and MaStR locations helps identify two main root causes for such deviation: On the one hand, some mappings include false negatives like partial detections as discussed in Section 5.1.1 and shown in Figure 5.4. On the other hand, observed detections on neighboring and/or adjacent

buildings would have to be grouped to match the extend of the registered unit. For example, Figure 5.11 shows units spread over several buildings. As discussed in Section 5.2.3, adjacent buildings can be considered by the given join types while they fail to map neighboring buildings.

While scattering to the right is less obvious due to usage of log scales, it occurs frequently for low and high thresholds, too. Overestimation can either be caused by dislocated MaStR points or a grouping of units with different commissioning dates on the same roof as explained in Section 5.2.3.

Given the described root causes for over- and underestimation, four types of matches with regard to capacity are evaluated to identify correct mappings: Matches are determined based on single MaStR units as well as aggregate capacities of all unit points located on a building. Additionally, matches for MaStR units located on the same as well as the adjacent building group are considered. The match ratio ΔM represents the fraction of mappings M for which a match or corresponding false duplicate mapping of the same unit to another system can be identified (M^{true}):

$$\Delta M = \frac{M^{true}}{M} \tag{5.2}$$

The colouring of points in Figure 5.12 identifies which mappings are considered matches according to these strategies. In total, 42% of mappings are identified as correct or duplicates of correct mappings, i.e. if the same unit is assigned to multiple buildings. In the district, the match ration is slightly higher with 47% of mappings being found to be correct in contrast to 37% of mappings in the city. This finding is also visualized in Figure 5.13 which shows match ratios across zip codes.

Overall, 36% of the detectable gross capacity can be validated by correspondences in image and building data. While for 3.25% of the detectable capacity no mapping could be created at all, the remainder is included in mappings were estimations do not match reported attributes. Among mappings without matching capacities, about 60% in both, city and district, underestimate the MaStR capacities. Consequently, benchmarking against detected capacities of the adjacent building group only partially helps to identify grouped systems. It would require an extension of the proposed join types to allow mappings across neighboring buildings, too. Given these limitations in the mapping methods, the correspondences not classified as matches cannot automatically be considered to be faulty MaStR entries.

Interestingly, [36, p.8] also report underestimated capacities for their address-level comparisons and higher detection ratios for large compared to small systems. However, they do not analyse the reason for this in more detail. In addition, their mapping is based on a nearest-neighbor search, but they do not further specify how

they handle the challenge of grouped units or whether they consider this a limitation of their methodology. Consequently, an exact comparison of performance between both works is not possible given the described evaluation approach.

5.2.5 Amount of Panels

		(1) ≤30kwP		(2) >30kwP	
		district	city	district	city
Modules per unit	mean	32.17	29.62	624.43	536.47
	std	42.70	46.70	1,383.86	1,249.19
	median	24.00	22.00	303.00	268.00
Gross Capacity per module [kWp]	mean	0.42	0.36	1.24	1.57
	std	4.81	1.11	18.16	18.56
	median	0.31	0.30	0.24	0.25

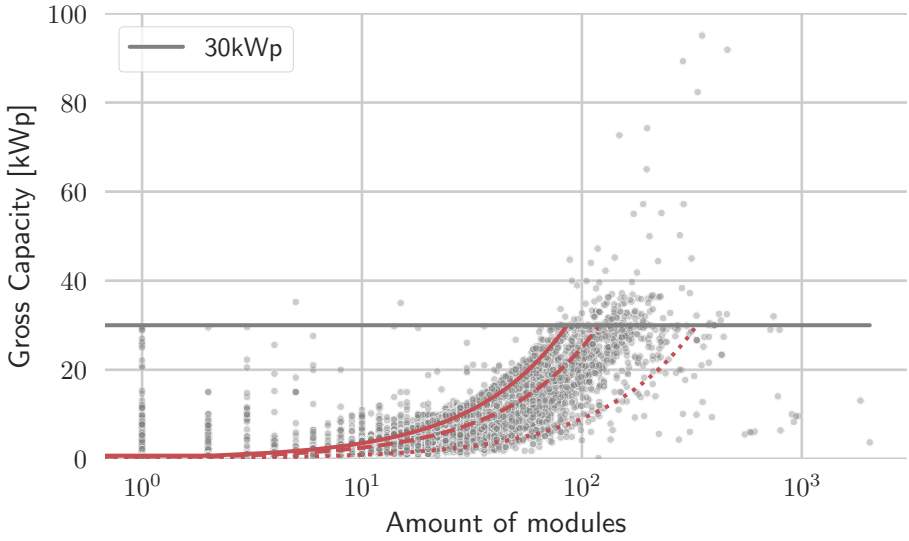
Table (5.5) Measures of central tendency for the amount of modules per unit and corresponding gross capacity per module [kWp] in city and district of Munich.

Internal Validation On average, the MaStR reports that a small unit consists of 31 ± 45 panels on average compared to 585 ± 1325 panels for large systems. Similar to the capacity, variability is extremely high for large systems. The plausibility of these ranges can be validated by comparing the panel counts with the net and gross capacities they are supposed to produce.

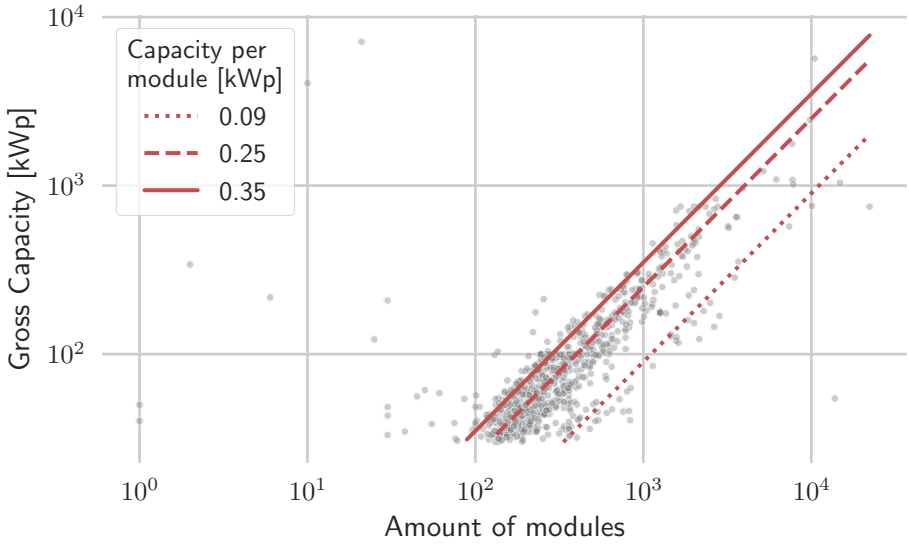
As discussed in Section 4.2.2, the standard yield per panel is between 0.25 and 0.35 kWp. Given the wide spread of panel counts among large systems, the median might give a clearer indication of the corresponding distribution of panel yields: For small systems, 50% of systems yield a gross capacity of up to 0.31 kWp compared to 0.25 kWp for large systems (see Table 5.5). The difference between the city and the district is negligible. Given that most large systems are older systems (see Section 2.3.1), it makes sense that the majority of such installations consists of panels with lower yields. Consequently, the standards assumed for estimation of panel counts seem to fit the given dataset when considering measures of central tendency.

In contrast, Figures 5.14a and 5.14b show the high variability among MaStR entries by visualizing the relation between the amount of modules and gross capacity per unit. Since capacity groups are determined by net capacity, some small units are shown to have higher gross capacities. The majority of systems does not exceed the limit of 0.35 kWp per module for small and large systems, but lots of units seem to consist of panels with lower capacities.

For estimation of panel counts $\hat{P}$, the detected system area $\hat{A}$ is divided by the



(a) Units with a net capacity $\leq$30kWp



(b) Units with a net capacity $>$ 30kwP

Figure (5.14) Alignment between gross capacity and amount of modules per MaStR unit (gray points) for small (a) and large (b) systems: Many units consist of modules with lower nameplate capacity, only few modules produce more than 0.35 kWp (red curve). Some capacities are reported to be generated by unrealistically low or high amounts of modules.

panel area per module A^p (see Equation 4.2). Consequently, larger panel sizes will result in smaller estimated panel counts and vice versa. Given the described distribution of panel yields and corresponding panel sizes in the dataset, the discussed thresholds are likely to underestimate panel counts for many units. Analogously, panel yields are assumed to be higher than they are which might also contribute to overestimation of total capacity for Munich in Section 5.2.4.

Additionally, both Figures make unlikely values with regard to capacity per module observable. For example, the MaStR reports a capacity of about 20 kWp to be achievable by 1 or even more than 100 modules (see Figure 5.14a). Additionally, there are units registered with a single module yielding capacities of up to 30 kWp. Such occurrences might be caused by a confusion between the terminology of units and modules. As explained in Section 3.1, the amount of panels per unit is not validated by the DNO wherefore such erroneous entries will not be validated or corrected after registration.

		(1) $\leq$ 30kwP		(2) $>$ 30kwP	
		$\hat{P}$	ΔP [%]	$\hat{P}$	ΔP [%]
city	lower	164,468	99.10	222,605	105.32
	upper	135,463	81.63	228,841	108.27
district	lower	152,578	75.28	214,585	101.94
	upper	127,086	62.70	218,479	103.79

Table (5.6) Aggregate lower and upper estimates of amount of modules ($\hat{P}$) and detection ratio (ΔP) by area and system size.

Fusion of detected systems and OSM buildings To evaluate the amount of panels reported in the MaStR, aggregates of detected $\hat{P}$ and registered P panel counts over all systems are compared. Analogous to the capacity, the module counts are validated using the corresponding detection ratio:

$$\Delta P = \frac{\hat{P}}{P^{detectable}} = \frac{\hat{A}}{A^p P^{detectable}} \tag{5.3}$$

In the city and district of Munich, 790,501 individual modules should be observable. With the estimation limits discussed in Section 4.2.2, this amount is underestimated by 4.59% to 10.2%. Therefore, the expectation derived by previous internal validation seems to hold true when considering the overall reference area.

However, evaluation by region and system size requires a more nuanced view: For small systems, the number of modules is underestimated in both the city and

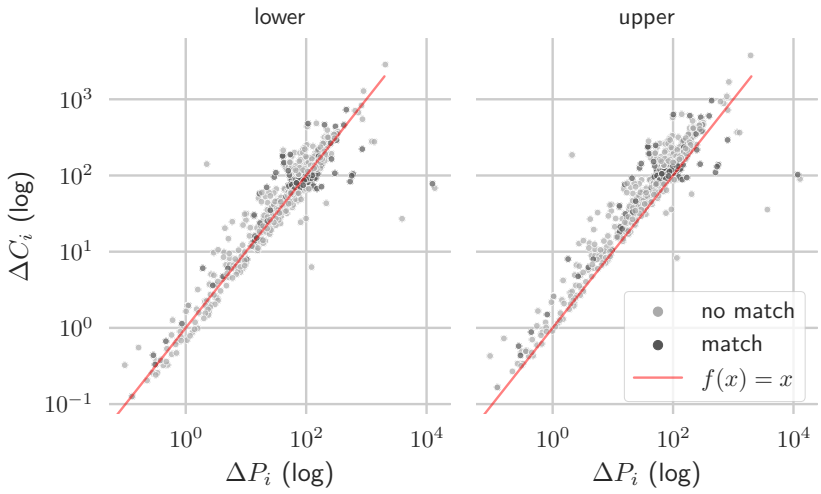


Figure (5.15) Detection Ratio of capacity ΔC_i and amount of modules ΔP_i: Each gray point represents a mapping between a large MaStR unit and a detected system. In contrast to the lower estimation limit (left), applying the higher threshold (right) for capacity and panel count estimation shifts the distribution upwards.

the district. In contrast, panel counts for large systems are consequently overestimated over both regions. As discussed in Section 5.2.4, a stronger tendency towards overestimation for large systems was also found for capacities. As shown in Equation 5.3, higher area sizes $\hat{A}$ cause an increase in estimated panel counts $\hat{P}$ and thus detection ratio ΔP. A more detailed comparison of detection ratios for capacities and panel counts is given in the next paragraph.

Comparing city and district, detection ratios ΔP are lower in the district compared to the city. As shown in Table 5.5, differences in panel yields between both areas are negligibly small. Consequently, this difference might also be caused by estimation of system sizes $\hat{A}$. Since the city has been found to be more affected by false positives, its system area size is more likely to be overestimated in comparison to the district.

Fusion of MaStR units and detected systems In general, capacity and panel counts are estimated based on the same panel size and yield (see Sections 4.2.2 to 4.2.3). As shown in the previous paragraph, this leads to observations with similar interpretations when comparing these estimates to the MaStR. However, corresponding MaStR attributes do not necessarily follow the same linear relationship. Consequently, both detection ratios can differ. Since mappings and matches have

been evaluated in previous sections, this paragraph focuses on differences in detection ratios ΔC and ΔP to validate the amount of modules reported in the MaStR.

As shown in Figure 5.15, detection ratios of capacities typically exceed those of panel counts. While the overestimation for the lower threshold is about 53%, it is already 86% when considering the higher limit. Additionally, the distribution is rather shifted upwards when analyzing the upper threshold: Detection ratios of panel area almost stay the same while the ratio for capacity increases. This can be explained by the 40% increase in panel yield from 0.25 to 0.35 kWp compared to a corresponding increase in area size of only 6% from 1.6 to 1.7 m^2. Consequently, the higher threshold assumes a yield which is too high compared to the assumed size of corresponding modules. Again, the lower estimation threshold seems to be more suitable to approximate distribution of solar system properties in Munich.

In contrast, outliers in Figure 5.15 are likely caused by erroneous entries in either MaStR gross capacity or amount of modules. If these points are located closer to an optimal detection ratio of 100% in one dimension, it suggests an error in the other.

5.2.6 Tilt

	(1) $\leq$30kwP		(2) $>$30kwP		Total	
	U	[%]	U	[%]	U	[%]
(1) $< 20°$	2,249	12.71	501	2.83	2,750	15.54
(2) 20-40°	10,885	61.54	412	2.33	11,297	63.87
(3) 40-60°	2,777	15.70	23	0.13	2,800	15.83
(4) $> 60°$	213	1.20	8	0.05	221	1.25
(5) Facade mounted	213	1.20	5	0.03	218	1.23
(6) Tracked	54	0.31	7	0.04	61	0.35
(7) No data	322	1.82	19	0.11	341	1.93
Total	16,713	94.48	975	5.52	17,688	100.00

Table (5.7) Amount of units by tilt angle and system size: U=absolute frequency, [%]=relative frequency.

Internal Validation As discussed in Section 2.1.2, the angle of incidence of solar radiation highly influences electricity yield. Assuming that the panel position is fixed, actual yield will still vary depending on the time of day and season due to positioning of the sun. For example, the sun's arc will be higher and longer in summer compared to winter. Therefore, PV units are installed with tilt angles that are estimated to maximize mean annual solar radiation. At higher latitudes, the tilt angle should be increased since the position of the sun is lower compared to

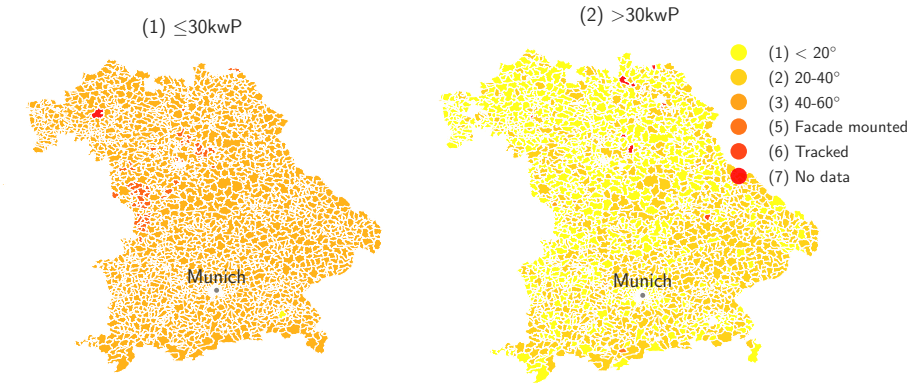


Figure (5.16) Most frequent tilt angle of units in Bavaria per zip code: Larger systems (2) are often installed on flat roofs causing tilt angles to be lower compared to smaller systems (1).

latitudes closer to the equator. According to [27, p.6], the optimal tilt for Munich is 33°.

In the MaStR, the tilt angle is registered in bins of 20°. Most units seem to be configured according to the given recommendation, since 65% of small and 42% of large units are tilted between 20 and 40° (see Table 5.7). For the majority of larger installations however, an angle of less than 20° is recorded. In contrast, only 13% of small units are installed with this tilt. On the one hand, this observation can be found to be true for entire Bavaria as shown in Figure 5.16. On the other hand, a similar finding is reported by [36] who assume that those PV systems are usually installed on flat roofs.

Another option to increase solar coverage are tracked solar PV panels, which adjust their tilt to the positioning of the sun next to other factors. According to [26, p.445], combined horizontal and vertical tracking allow capturing of 17% more solar radiation compared to panels installed at a fixed optimal tilt. In comparison to horizontally mounted modules, the capture rate even increases by 39%. Overall, only 61 units in Munich are tracked solar PV installations, 54 of them small ones.

Additionally, 218 units are reported to be mounted to the facade. Five of these are large ones which seems quite unlikely given the size of such installations. In total, only 1.9% of small and large units are missing any information on tilt configuration.

Fusion of MaStR units and OSM buildings Only 0.27% of OSM buildings in Munich have information on their tilt angle given. Additionally, OSM provides

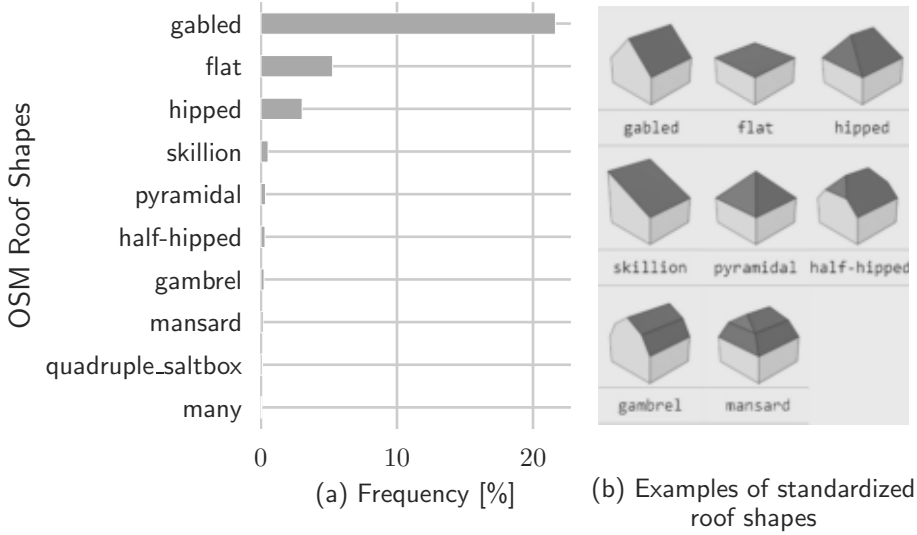


Figure (5.17) Frequency of OSM roof shapes across all buildings in Munich (a) and corresponding standardized visualizations (b) according to [13]: Only few systems have information on the roof shape given. Most roofs are either gabled or flat.

information on standardized roof shapes for 29.91% of buildings. As shown in Figure 5.17, for the majority of OSM entities roofs are either gabled (20.32%) or flat (4.95%). A visual representation of some sample shapes as defined by the OSM documentation are given in Figure 5.17 [13].

Among mappings between MaStR and OSM, no roof shape is specified for 70.78% of assigned buildings. In contrast to the overall distribution of roof shapes shown in Figure 5.17, 13.52% of mappings are reported to be related to a gabled roof, 12.52% to a flat and 1.54% to a skillion roofs. The remaining roof shapes have occurrences below 1%.

While standardized representations allow a distinction between tilted and flat roofs, they do not provide sufficient information on exact tilt angles. Additionally, one would have to assume that a tilt reported as below 20° in the MaStR corresponds to a flat roof. Otherwise, it is unclear whether a non-reported tilt angle indicates no tilt and thus a flat roof or just no information and thus any tilt angle. Consequently, the OSM data does not provide sufficient coverage to validate the MaStR fields on roof architecture in a scalable manner. It is thus not further analysed for unit-to-unit correspondences.

5.2.7 Azimuth orientation

	(1) ≤30kwP		(2) >30kwP		Total	
	U	[%]	U	[%]	U	[%]
(01) North	150	0.85	14	0.08	164	0.93
(02) North-East	142	0.80	7	0.04	149	0.84
(03) East	786	4.44	45	0.25	831	4.69
(04) South-East	1,666	9.42	113	0.64	1,779	10.06
(05) South	8,897	50.30	498	2.82	9,395	53.12
(06) South-West	2,432	13.75	98	0.55	2,530	14.30
(07) West	1,153	6.52	52	0.29	1,205	6.81
(08) North-West	87	0.49	3	0.02	90	0.51
(09) East-West	1,253	7.08	131	0.74	1,384	7.82
(10) Tracked	5	0.03	1	0.01	6	0.04
(11) No data	142	0.80	13	0.07	155	0.87
Total	16,713	94.48	975	5.51	17,688	99.99

Table (5.8) Amount of units by azimuth angle and system size: U=absolute frequency, [%]=relative frequency.

Internal Validation In the MaStR, azimuth angles are reported in categorical bins like North or South as shown in Table 5.8. Overall, only 0.87% of units in Munich lack this information. One large and five small units are tracked and thus do not provide a fixed angle. For achieving optimal yield, it is recommended to orient PV units southwards. A majority of 53.12% of units follows this guidance. Similarly, 24.36% of the units are either southeast or southwest facing, which generally reduces the yield slightly [30]. Another 19.32% of units are oriented either to the east, west or without a clear distinction between both. Especially for large installations, the latter applies to 13.43% of units in this capacity category.

Fusion of MaStR units and OSM buildings In general, if a building's shape is roughly rectangular, it is likely that the building has a north-south orientation, with the longer sides facing east and west. This is because it allows for more sunlight to enter the building's windows and provides more opportunities for natural lighting and passive solar heating. However, this general assumption is insufficient to accurately calculate the orientation of a building based on its outline polygon.

Consequently, it requires an exact architectural specification for reliable validation of MaStR values. However, the orientation of buildings is only known for a minor share of 2.5% of OSM buildings in Munich. As for tilt angles, it is not possible to derive azimuth angles using OSM data only.

5.3 Evaluation of MaStR extension by external data

The following section describes information extracted from aerial images (Section 5.3.1) and OSM building data (Section 5.3.2) and assesses their potential to complement the MaStR. The corresponding findings build on and extend the quantitative analysis performed in previous sections.

5.3.1 Potential of MaStR extension by Aerial Image Data

Raw Images Although high resolution, geo-referenced RGB images as analysed in this research are not available on a national level yet, they can provide a useful extension to the MaStR if available.

On the one hand, they simplify manual detection of solar systems and their distinction in regard to i.e. solar thermal detectors or roof structures like windows and dormers. For example, they can be used to verify whether a solar unit is installed at the reported point coordinates and address or to verify the size of a system. They also proved helpful to verify the correctness of detections and corresponding mappings to the MaStR.

On the other hand, they are likely to yield better results during automated detection of solar PV systems than lower resolution images (see Section 3.2). Relevant features of solar units like the rectangular structures of single panels are not as blurred and thus clearer to detect.

Additionally, the link of MaStR units to a corresponding image area can support further use cases. For example, [29] applied semantic segmentation to roof segments and superstructures for analysis of solar potentials. Given the localizability of a system, such methods could support an analysis of each unit's extensibility.

PV System Polygons The evaluation of locational information (Section 5.2.3) and mappings between MaStR units and detected systems (Section 5.2.4) indicated the limitations of using point coordinates for fusion of MaStR and external data. To allow extensive correspondence search, it is required to know the exact spatial extend of each solar unit.

On the one hand, this would simplify mapping for units which are spread over several adjacent or neighboring buildings. Instead of referencing a single point, a multipolygon which groups all parts of a unit could be referenced. This would improve transparency about the exact location and ownership of solar installations.

On the other hand, it would help clarify which part of a detected system was commissioned at which point in time. To not have to extend given geocoordinates manually, the detection approach discussed in this research can support in automated generation of unit polygons (see Section 4.1).

System Area To estimate PV attributes like capacity and amount of modules, the ground area covered by a detected system is used. Additionally, knowledge of the system area can be used for analysis of land cover classes according to [34]. It can also support deep learning approaches mentioned previously which aim to determine the extensibility of solar units, i.e. by pre-filtering of units on their roof to area ratio.

Although this field is given in the MaStR, it is only mandatory for field systems and found missing for 99.9% of entries in Munich. Additionally, it is not validated by a DNO (see Table 3.2).

However, the methods explained in Section 4.2.1 allow automated calculation of system size from detected polygons. Consequently, this research can help fill the current gap for this parameter in the MaStR.

5.3.2 Potential of MaStR extension by Building Data

Roof or Building polygons and area As shown in this research, OSM building polygons allow to verify the reported location of roof-mounted systems. While knowing the shape of a building does not provide accurate information on its orientation (see Section 5.2.7), it allows calculation of its roof size. Next to the polygon area, this is another relevant parameter for unit extensibility analysis and could provide a useful extension to the MaStR.

Building properties Analysis in previous sections showed that the OSM tags on orientation, tilt, building type and usage purpose do not provide sufficient coverage and data quality to extend the MaStR in a scalable manner. While the address namespace is useful for validating the MaStR, it does not contain additional information beyond what is already registered. The same holds true for the information returned by the Nominatim API when geo- and reverse geocoding addresses.

The standardized roof shapes given in OSM for 29.91% of buildings in Munich may not suffice to estimate tilt and orientation (see Section 5.2.6). Nonetheless, they could be used to extend the MaStR entries and specify whether a system is installed on a gabled versus a flat roof. Overall, energy companies and utilities may be interested in this information due to the effect of tilt on energy output.

For other OSM tags, only spot checks on potentially relevant attributes for solar systems have been performed. For example, building height and levels could be helpful when analysing shading on a solar system by neighboring buildings. While exact height is only given for two buildings, the amount of levels is specified for 36.25% of mappings between MaStR and OSM. Consequently, although building properties specified in OSM are not provided for all buildings they might be a starting point to complement roof-mounted systems in the MaStR.

Chapter 6

Conclusion

This chapter presents the central findings of this work as well as their critical discussion (Section 6.1). Finally, it highlights limitations and corresponding opportunities for further research (Section 6.2).

6.1 Discussion

In summary, this research aims to explore methods for automated validation of data on solar units in the MaStR (RQ1). These are based on creating correspondences between MaStR records and external sources like aerial images and buildings. Additionally, this work presents an overview of information contained in and extracted from these alternative datasets to complement solar MaStR units (RQ2).

The methods are evaluated by conducting a study on MaStR and OSM extracts as well as images of the city and district of Munich. The CV techniques of image classification and semantic segmentation are applied to detect solar systems in high-resolution images. After extracting their boundary polygons, system attributes like capacity and amount of modules are estimated and compared to MaStR data. Building polygons, addresses and geocoding services provided by OSM are used to validate locational information of MaStR units. Additionally, OSM data supports the generation of correspondences with regard to roof-mounted solar units between the different data sources. Finally, validation performance is reported for a selection of MaStR fields and the potential of image and OSM building data to complement the MaStR is discussed.

In conclusion, the discussed methods allow different extends of validation depending on the accessibility and coverage of i.e. high resolution images, locational information in the MaStR and building properties. However, even without external data, an internal validation based on plausibility checks of individual MaStR fields

and their relation lead to identification of several errors: Unrealistic capacities for installation types like plug-in systems or facade mounted units have been found. Additionally, there exist conversion errors with regard to kWp and MW for gross and net but especially for inverter capacities. For lots of units, reported capacities are related to panel counts that would require unrealistic panel yields which could indicate a confusion of units and modules.

Using openly accessible OSM data allows to confirm the place of installation and i.e. identify field systems that are registered with a location that corresponds to a building. Additionally, it supports in verifying whether geocoordinates of roof mounted systems are actually placed on a building and whether their coordinates and addresses overlap. The implementation of several join types proved helpful to reduce dependency on exact positioning of MaStR coordinates for identifying correspondences between solar units and buildings. However, since the MaStR only reports point coordinates, these methods are limited when it comes to the generation of mappings for units spread over several buildings. Additionally, OSM data coverage and quality proved to be insufficient to validate data on building type, usage purpose, tilt and azimuth angles. However, building properties like roof shape or levels per building are available and could complement the MaStR for many solar units.

Access to high resolution images allows to validate solar PV system attributes like capacity and the amount of modules forming a unit. The total detectable capacity is overestimated by at least 13% which indicates an incorrectly registered amount of operating units at the time of image capture. Additionally, 36% of gross capacity for detectable large systems can be verified based on the discussed mapping and estimation approaches. To address the limitation of MaStR point coordinates, system polygons generated in this research can be used to complement the registry. Moreover, high-resolution images support several use cases and should be added as reference if available.

Overall, the presented results demonstrate the basic ability of the proposed methods to validate several MaStR fields and detect a non-negligible fraction of errors in the registry. Additionally, existing records can be complemented by information extracted from external sources. Although several limitations still need to be addressed (see Section 6.2), this demonstrates the potential of the proposed methodology to not only validate MaStR records but more importantly automatically generate them.

6.2 Limitations and Future Work

The discussion of performance on the Munich dataset shows limitations of the suggested approach and leads to identification of several opportunities for further research:

Alternative External Data Sources Although OSM data has the advantage of being openly accessible, it has been shown to have limited coverage for relevant attributes like tilt or orientation. Next to an image dataset, the LDBV also published a dataset with building properties at LoD2 for Bavaria in the beginning of 2023 [31]. It includes polygons for buildings and roofs and attributes on roof tilt and azimuth angles. Since it is published by an official institution in contrast to OSM, this dataset might provide a more suitable base for validation and extension of the MaStR.

PV System Detection A main problem with regard to solar PV system detection has been found to be the sensitivity of the models to certain roof structures (see Section 5.1.1). To address this issue, the fraction of negative samples showing such special patterns like glass roofs could be increased during model training and testing. It remains unclear to what extend solar PV modules are confused with solar thermal collector panels. Consequently, it might be recommendable to review the training samples or even extend the approach to detect both types according to [39]. Additionally, it seems to be recommendable to include the threshold for binarization of segmentation masks in hyperparameter tuning as suggested by [36]. This might support in finding a balance between elimination of false positives versus false negatives. In extension to the fully-supervised models, [16] discussed a semi-supervised approach for automated detection of solar panels in the Munich image dataset. It could provide further insights to compare the performance of both model types. To improve scalability of the discussed methods, both models should be re-trained on images of lower resolution, i.e. the images for Bavaria published on the OpenData platform (see Section 3.2). To also reduce sensitivity of models to the location of solar panels in the images, the training images should not be centered on the panels as done by [16] but rather include a variety of placements.

Estimation of PV System Attributes Validation of system attributes like the amount of modules and capacity is based on approximations assuming standard module properties. While there is little alternative to such an approach with regard to capacity, alternatives for detecting the amount of modules could be considered. For example, image processing techniques such as edge detection could support

identification of more accurate panel counts. In addition, [15] suggested to train deep learning models for estimating implementation details like azimuth angles from images.

Data Fusion Since mapping approaches discussed in this work are based on OSM buildings, no valid correspondences for field systems can be generated. Additionally, no mappings can be generated for neighboring buildings which results in a low match rate. The discussed methods could be extended i.e. by a nearest neighbor matching as applied by [36].

Bibliography

[1] Kyle Bradbury, Raghav Saboo, Timothy L Johnson, Jordan M Malof, Arjun Devarajan, Wuming Zhang, Leslie M Collins, and Richard G Newell. Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data*, 3(1), 2016.

[2] Bundesnetzagentur. Das Marktstammdatenregister - Gesamtkonzept. 2018.

[3] Bundesnetzagentur. Struktur der Daten zu Marktakteuren, Einheiten und Gruppierungsobjekten im Marktstammdatenregister. 2018.

[4] Bundesnetzagentur. Bundesnetzagentur - Datennutzungsgesetz. `https://www.bundesnetzagentur.de/DE/Fachthemen/Digitalisierung/Daten/DNG/start.html`, 2023. Accessed: 2023-02-15.

[5] Bundesnetzagentur. Datendownload | MaStR. `https://www.marktstammdatenregister.de/MaStR/Datendownload`, 2023. Accessed: 2023-01-20.

[6] Bundesnetzagentur. Historie des MaStR. `https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Monitoringberichte/Marktstammdatenregister/Historie/start.html`, 2023. Accessed: 2023-01-19.

[7] Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen, Bundesnetzagentur, and Bundeskartellamt. Monitoringbericht 2022. 2022.

[8] GeoPy Contributors. Welcome to GeoPy's documentation! — GeoPy 2.3.0 documentation. `https://geopy.readthedocs.io/en/stable/index.html?highlight=nominatim#nominatim`, 2018. Accessed: 2023-02-16.

[9] OpenStreetMap Contributors. Copyright and License. https://www.openstreetmap.org/copyright. Accessed: 2023-02-15.

[10] OpenStreetMap Contributors. Nominatim Documentation. https://nominatim.org/release-docs/develop/. Accessed: 2023-02-16.

[11] OpenStreetMap Contributors. Key:building. https://wiki.openstreetmap.org/wiki/Key:building, 2022. Accessed: 2023-02-18.

[12] OpenStreetMap Contributors. Tags. https://wiki.openstreetmap.org/wiki/Tags, 2022. Accessed: 2023-02-16.

[13] OpenStreetMap Contributors. Simple 3D Buildings. https://wiki.openstreetmap.org/wiki/Simple_3D_Buildings#Roof_shape, 2023. Accessed: 2023-02-18.

[14] Esri Deutschland. Postleitzahlengebiete - OSM. https://opendata-esri-de.opendata.arcgis.com/datasets/5b203df4357844c8a6715d7d411a8341_0, 2020. Accessed: 2023-02-19.

[15] Ayobami S. Edun, Kirsten Perry, Joel B. Harley, and Chris Deline. Unsupervised azimuth estimation of solar arrays in low-resolution satellite imagery through semantic segmentation and Hough transform. *Applied Energy*, 298, 2021.

[16] Yasmin Elsharnoby. Exploring Different Supervision Techniques for Deep Learning-based Image Segmentation of Photovoltaic Systems in Munich. Master's thesis, Technical University of Munich, 2022.

[17] Energieexperten. PV Modul-Größen im Überblick. https://www.energie-experten.org/erneuerbare-energien/photovoltaik/solarmodule/groesse, 2022. Accessed: 2023-02-07.

[18] Energieexperten. Schwarze Solarmodule: Technik, Leistung & Kosten. https://www.energie-experten.org/erneuerbare-energien/photovoltaik/solarmodule/schwarze, 2022. Accessed: 2023-03-19.

[19] Bundesamt für Kartographie und Geodäsie. Postleitzahlgebiete Deutschland. https://gdz.bkg.bund.de/index.php/default/postleitzahlgebiete-deutschland-plz.html, 2022. Accessed: 2023-02-19.

[20] Bundesministerium für Wirtschaft und Klimaschutz. Neue Möglichkeiten und Einsatzfelder für Concentrated Solar Power. https://www.german-energy-solutions.de/GES/Redaktion/DE/Meldungen/Aktuelle-Meldungen/2021/20210618-interview-branche-des-monats-csp.html, 2021. Accessed: 2023-02-04.

[21] Geofabrik GmbH. Geofabrik Download Server. http://download.geofabrik.de/, 2018. Accessed: 2023-02-16.

[22] Klokan Technologies GmbH. ETRS89 / UTM zone 32N - EPSG:25832. https://epsg.io, 2022. Accessed: 2023-02-04.

[23] GovData. Datenlizenz Deutschland-Namensnennung-Version 2.0. https://www.govdata.de/dl-de/by-2-0. Accessed: 2023-01-20.

[24] Abdishakur Hassan and Jayakrishnan Vijayaraghavan. *Geospatial Data Science Quick Start Guide: Effective techniques for performing smarter geospatial analysis using location intelligence*. Packt Publishing Ltd, 2019.

[25] Ludwig Hülk, Guido Pleßmann, Christoph Muschner, Florian Kotthoff, and Deniz Tepe. open-MaStR. https://github.com/OpenEnergyPlatform/open-MaStR/, 2022. Accessed: 2023-01-19.

[26] Mark Z. Jacobson. *100% Clean, Renewable Energy and Storage for Everything*. Cambridge University Press, 2020.

[27] Mark Z. Jacobson and Vijaysinh Jadhav. World estimates of PV optimal tilt angles and ratios of sunlight incident upon tilted and tracked PV panels relative to horizontal panels. *Solar Energy*, 169, 2018.

[28] Gabriel Kasmi, Laurent Dubus, Philippe Blanc, and Yves-Marie Saint-Drenan. Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed pv mapping. In *Workshop on Machine Learning for Earth Observation (MACLEAN), in Conjunction with the ECML/PKDD 2022*, 2022.

[29] Sebastian Krapf, Lukas Bogenrieder, Fabian Netzler, Georg Balke, and Markus Lienkamp. RID—Roof Information Dataset for Computer Vision-Based Photovoltaic Potential Assessment. *Remote Sensing*, 14(10), 2022.

[30] Nadine Kümpel. So bestimmen Sie die Größe Ihrer Photovoltaikanlage. `https://www.wegatech.de/ratgeber/photovoltaik/planung-und-installation/dimensionierung-pv-anlage/`, 2022. Accessed: 2023-01-29.

[31] Landesamt für Digitalisierung, Breitband und Vermessung. BayernAtlas. `https://www.bayernatlas.de`, 2023. Accessed: 2023-03-22.

[32] Jahn Mahn, Pina Merkert, and Andrijan Möcker. Balkonkraftwerke: Fragen und Antworten zu Installation und Angeboten. `https://www.heise.de/ratgeber`, 2023. Accessed: 2023-02-11.

[33] Jordan M. Malof, Leslie M. Collins, and Kyle Bradbury. A deep convolutional neural network, with pre-training, for solar photovoltaic array detection in aerial imagery. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[34] David Manske, Lukas Grosch, Julius Schmiedt, Nora Mittelstädt, and Daniela Thrän. Geo-Locations and System Data of Renewable Energy Installations in Germany. *Data*, 7(9), 2022.

[35] Mapbox. rasterio.features module. `https://rasterio.readthedocs.io/en/latest/api/rasterio.features.html#rasterio.features.shapes`, 2018. Accessed: 2023-02-21.

[36] Kevin Mayer, Benjamin Rausch, Marie-Louise Arlt, Gunther Gust, Zhecheng Wang, Dirk Neumann, and Ram Rajagopal. 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310, 2022.

[37] QGIS project. A Gentle Introduction to GIS — QGIS Documentation documentation. `https://docs.qgis.org/3.22/en/docs/gentle_gis_introduction/index.html`, 2023. Accessed: 2023-02-04.

[38] Wolfram Schneider. BBBike Extract Service. `https://extract.bbbike.org/`, 2023. Accessed: 2023-02-16.

[39] Maximilian Schulz, Bilel Boughattas, and Frank Wendel. DetEEktor: Mask R-CNN based neural network for energy plant identification on aerial photographs. *Energy and AI*, 5, 2021.

[40] Henrikki Tenkanen. Pyrosm - OpenStreetMap PBF data parser for Python. `https://pyrosm.readthedocs.io/en/latest/index.html`, 2020. Accessed: 2023-02-16.

Appendix A

Program Code and Resources

The source code and a documentation are available at the GitHub Repository: `https://github.com/esthervogt/MTMASTRS`. Access to this private repository was provided to Florian Kotthoff (fortiss) and Sascha Marton (InES). In case of access or permission issues, please reach out to: `esther.vogt@students.uni-mannheim.de`.

Appendix B

Acronyms

AC Alternating Current . . . 7
AGS Official Municipality Code . . . 4
BAFA Federal Office of Economics and Export Control . . . 12
BKartA Federal Cartel Office . . . 12
BMEL Federal Ministry of Food and Agriculture . . . 12
BMWK Federal Ministry for Economic Affairs and Climate Action . . . 1
BNetzA Federal Network Agency . . . 1
CRS Coordinate Reference System . . . 7
CSP Concentrated Solar Power . . . 6
CV Computer Vision . . . 22
DBMS Database Management System . . . 14
DC Direct Current . . . 7
DIBT German Institute for Building Technology . . . 27
DNO Distribution Network Operator . . . 1
DUA Act governing the use of public sector data . . . 19
EEG Renewable Energy Act . . . 10
EnWG Electricity and Gas Supply Act . . . 1
EPSG European Petroleum Survey Group . . . 7
kWh Hourly DC Watts . . . 7
kWp Peak DC Watts . . . 7
LDBV Bavarian Agency for Digitisation, High-Speed Internet and Surveying . . . 19
LoD2 Level of Detail 2 . . . 3
MaStR Marktstammdatenregister . . . ii
MaStRV Regulation on the registration of energy industry data . . . 8
MW Mega Watts . . . 10

NRW North-Rhine Westphalia 3
OSM Open Street Map ii
PBF Protocolbuffer Binary Format 20
PV Photovoltaics ii
QA Quality Assurance 10
RES Renewable Energy Sources 1
SEE Electricity Producing Unit 14
SEL Technical Electricity Producing Location 15
STC Standard Test Conditions 7
TIFF Tagged Image File Format 8
TSO Transmission System Operator 1
UBA German Environment Agency 12
V GeoBund Contract on the continuous transmission of official digital geodata of the federal states for use in the federal area 21

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Masterarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Mannheim, den 23.03.2023

Esther Vogt