



Geo-Locations and System Data of Renewable Energy Installations in Germany

David Manske ^{1,*} , Lukas Grosch ¹, Julius Schmiedt ¹, Nora Mittelstädt ¹ and Daniela Thrän ^{1,2} 

¹ Department of Bioenergy, Helmholtz Centre for Environmental Research GmbH—UFZ, Permoserstraße 15, 04318 Leipzig, Germany

² Bioenergy Systems Department, DBFZ Deutsches Biomasseforschungszentrum gGmbH, Torgauer Str. 116, 04347 Leipzig, Germany

* Correspondence: david.manske@ufz.de; Tel.: +49-3412-434-596

Abstract: Information on geo-locations of renewable energy installations is very useful to investigate spatial, social or environmental questions on their impact at local and national level. However, existing data sets do not provide a sufficiently accurate representation of these installations in Germany over space and time. This work provides a valid approach on how a data set of wind power plants, photovoltaic field systems, bioenergy plants and hydropower plants can be created for Germany based on a data extract from the Core Energy Market Data Register (CEMDR) and publicly available data. Established methods were used (e.g., random forest, image recognition), but new techniques were also developed to fill data gaps or locate misplaced renewable energy installations. In this way, a substantial part of the CEMDR data could be corrected and processed in such a way that it can be freely used in a GIS software by any scientific and non-scientific discipline.

Dataset: <https://doi.org/10.5281/zenodo.6922043>

Dataset License: <http://dcat-ap.de/def/licenses/dl-by-de/2.0>

Keywords: renewable energies; renewable energy plants; renewable infrastructure; wind power plants; photovoltaic field systems; bioenergy plants; hydro power plants; GIS; spatial data



Citation: Manske, D.; Grosch, L.; Schmiedt, J.; Mittelstädt, N.; Thrän, D. Geo-Locations and System Data of Renewable Energy Installations in Germany. *Data* **2022**, *7*, 128. <https://doi.org/10.3390/data7090128>

Academic Editor: Jamal Jokar Arsanjani

Received: 1 August 2022

Accepted: 7 September 2022

Published: 10 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Summary

Renewable energies are an important pillar of power supply in Germany. Thus, the share of renewable power in gross electricity consumption rose from 6.3% in 2000 to 45.2% in 2020 [1]. This increase is due to the nationwide expansion of renewable energies for power generation such as wind power plants or photovoltaic systems. However, the development of renewable energy in Germany has also led to political conflicts with residents, topics of land use, nature conservation or the landscape [2,3]. In order to provide a detailed insight into the spatial expansion of renewable energy plants over time, precise information on plant locations and system data is helpful. This can help to better track environmental issues and impacts of renewables at local and national level [3,4], shed light on the spatial distribution and equity of energy transition, or calculate site-specific generation patterns using numerical simulation models [5,6]. Data sets on the geo-locations of renewable energy installations already exist (e.g., [7,8]). What all these data sets have in common, however, is that the data they contain do not adequately represent the renewable energy plant stock in Germany over space and time. Either data records for plants are missing or existing data records are incomplete, a fact which was already noted by [8,9] before. That is why we have already published a data set and an article on the spatial distribution of wind turbines, photovoltaic field systems, bioenergy plants and river hydropower plants in Germany in 2019 [8]. However, this data collection only covers installations up to the year 2015 with a so far only roughly resolved and partly imprecise level of information

from different sources. In this sense, this contribution is intended as a continuation and further development of the existing work. The aim of this work was to create a data set on the geo-locations and system data of renewable energy installations in Germany that is as error-free as possible. For this purpose, data from the Core Energy Market Data Register (CEMDR) [10] was used and cross-checked with other available sources where necessary. In this work, we focus on onshore and offshore wind power plants, photovoltaic field systems, bioenergy plants and hydropower plants in Germany.

The article is structured as follows: The following Section 2 briefly and precisely describes the format of the compiled data set and how the data can be read and interpreted. In Section 3, the source data used and the procedure for compiling the data set are presented, structured by the types of renewable energy installations mentioned. Finally, Section 4 summarizes and discusses the main outcomes and the significance of the data set.

2. Data Description

The data set described here represents geo-locations and system data of renewable energy installations in Germany up to the cut-off date 7 May 2021 (Version V20210507) and can be downloaded freely available under the CC BY 2.0 DE license at [11]. The data set contains five geodata files in the format GeoJSON with the spatial geometry types points for onshore and offshore wind power plants, bioenergy plants and hydropower plants as well as polygons for photovoltaic field systems, which can be read by any GIS software (e.g., QGIS), and a text file that explains the contents of the data set. The reference coordinate system of the files is WGS 84 (EPSG 4326). Table 1 gives an overall view of the compiled data set by type of renewable energy installation and the number of records included, as well as the total net installed capacity in MW.

Table 1. Composition of the compiled data set according to the type of renewable energy installations.

Renewable Energy Installation	Number of Records	Total Capacity in MW
Offshore Wind Power Plants	1497	7764
Onshore Wind Power Plants	28,156	54,905
Photovoltaic Field Systems	6621	13,807
Bioenergy Plants	19,940	8493
Hydropower Plants	8042	5832

The geodata files of the data set consist of a varying number of variables organised in data frames that represent the corresponding renewable energy installations in a table format with rows and columns. Each row of the data frame represents a renewable energy installation, and each column describes its technical or non-technical characteristics, such as installed system capacity or primary data sources. Table 2 shows the relevant variables that make up the column names of the data frame. Depending on the type of renewable energy installation, more or less variables are available. For example, the variable with the technical data for the hub height (HUB) is only present in the data files that contain wind turbines.

Table 2. Variables used in the compilation of the data set. The short term reflects the variables from the published data set.

Variable	Name	Description
RES	Renewable Energy Source	Renewable energy source with which the plant is operated
CAP	Installed Capacity (kW)	Installed net power of the power generator of the plant in kW
COD	Commissioning Date	Date of commissioning of the plant
DOD	Decommissioning Date	Date of decommissioning of the plant
COY	Commissioning Year	Year of commissioning of the plant
DOY	Decommissioning Year	Year of decommissioning of the plant
HUB	Hub height	Hub height of the wind power plant
ROD	Rotor diameter	Rotor diameter of the wind power plant
ALG	Alignment	Sky alignment of photovoltaic systems
INC	Inclination	Angle of inclination of photovoltaic systems
BPS	Biogas-Production-Site	Identifier which biogas-on-site electricity generation units belong to the same biogas production site
SYS	System	Manufacturer or type of system of the plant, e. g. ENERCON
TYP	Type	Subtypes of the system, e. g. E-115
LAT	Latitude	Latitude of the plant location (WGS 84, EPSG 4326)
LON	Longitude	Longitude of the plant location (WGS 84, EPSG 4326)
SRC	Source	Reference source of the record
SID	Source Identification Number	Unique identifier of the record as given in the reference source
NTE	Note	Note on the record, e.g., whether it has been edited or whether there is a special feature
ACT	Actuality	Actuality of the record as given in the reference source

3. Raw Data and Methods

This section describes the raw data used and the method applied to compile the data set for the renewable energy installation types mentioned. The technical process of data processing is documented in a Git repository and can be reviewed at [12].

The raw data used for this work were taken from the Core Energy Market Data Register (CEMDR) maintained by the Federal Network Agency and reflects the status as of the reporting date 7 May 2021 [10]. The CEMDR data includes all power generation installations in Germany and is provided in XML format. This data source was chosen because it offers the most comprehensive data on renewable energy installations in Germany. However, even though the data provided by the CEMDR is very detailed, the information it contains may be wrong or inaccurate. The main reason for this is incorrect or erroneous information submitted by system operators to the CEMDR. It has been shown, however, that erroneous data of this kind can be corrected by cross-validation with other data sources, by data science techniques or by plant-specific searches [8,9]. Data sources that can be used for cross-validation are, for example, plant data from the four large transmission system operators in Germany (Amprion, 50Hertz, TransnetBW and Tenet TSO). They maintain a public list of information on renewable energy plants that are subsidised under the Renewable Energy Sources Act (EEG), but only with reduced plant information [13]. The federal states also offer data on renewable energy installations in publicly accessible data portals, although the scope and level of detail of the information offered varies (e.g., [14–16]). In addition, there are also commercially accessible databases that can be used for cross-validation if required (e.g., [17,18]).

As target format for the single files of the compiled data set the GeoJSON format was chosen. The GeoJSON format is an open standard format for representing simple geographic features along with their non-spatial attributes [19]. Compared to other ways of making the data available, e.g., via an application programming interface (API), this

format has the advantage that files of this format can be easily shared, read by common GIS software and used by a user group with little IT knowledge.

By using mentioned alternate data sources and existing methods in association with the techniques presented in this work, the CEMDR data can be improved and made more precise and accessible. This makes it possible to obtain a temporally and spatially high-resolution image of the German renewable energy installation stock in the electricity sector. The procedure for compiling the data set is described below for each type of installation mentioned and reflects the work steps of the technical data processing.

3.1. Wind Power Plants

With a share of 46% in 2021, wind energy has the largest share of the total installed electricity generation capacity from renewable energies in Germany [1]. The capacity has been continuously expanded in recent years and has been built both on land and at sea. The data extract from the CEMDR contains a combined total of 30,759 onshore and offshore installations, excluding those that are planned. Table 3 gives an overview of the completeness of the initial data of onshore and offshore wind power plants. What looks like a reasonably complete data set at first sight turns out to be partially wrong on closer inspection, especially with regard to the geographical location of the wind power plants. For example, 500 of the onshore wind power plants had obviously incorrect geographic information (Figure 1a). For this reason, the onshore wind power data were essentially subjected to a review and modified in several successive work steps. The offshore wind data set on the other hand were almost complete and error-free. Only three records had to be deleted because of wrong coordinates, which could not be corrected either.

Table 3. Number, total capacity and number of missing data in the initial data set of onshore and offshore wind power plants of the CEMDR.

Characteristics	Onshore Wind Power Plants	Offshore Wind Power Plants
Number of Records	29,259	1500
Total Capacity in MW	55,633	7775
Missing Capacities	0	0
Missing Commissioning Dates	0	0
Missing Hub Heights	744	1
Missing Rotor Diameters	414	3
Missing Geo-coordinates	709	0

Thus, the initial data set of onshore wind power plants was initially reduced by 691 entries after all records with missing and untraceable geo-coordinates were removed. This was done by a query to delete records without geo coordinates.

Table 4 documents the change in the onshore wind turbine data set during data processing. The records removed were small wind power plants with a generation capacity of less than 30 kW and a total generation capacity of about 5 MW. Although the total number of wind power plants in the data set decreased at this point, the total installed capacity increased by 37 MW, as some obviously incorrect information on installed capacity could be corrected.

The onshore wind power plants with obviously wrong coordinates were shifted to the location of the municipality stored in the initial data extract and then manually placed in their actual location with the help of aerial photographs (Google Maps) or by tracing the geo-information in the other data sources mentioned.

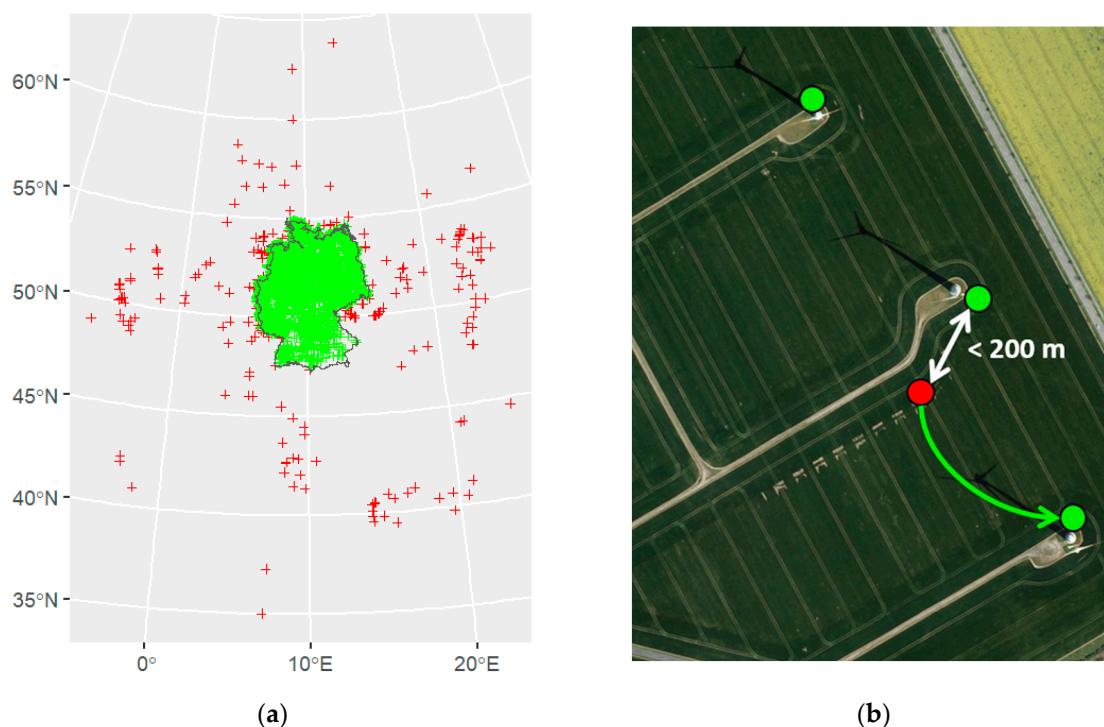


Figure 1. Processing steps in the compilation of the onshore wind turbine data set: (a) Onshore wind power plants with obviously incorrect location information (red crosses) and supposedly correct location information (green crosses). (b) Distance measurement to the nearest wind turbine in the initial data set and manual location correction of an initial misplaced CEMDR data point (red dot).

Table 4. Changes in the number of onshore wind power plants and total capacity during data processing.

Processing Step	Number of Records	Total Capacity in MW
Initial data set	29,259	55,633
After removing all records with missing localisation	28,568	55,670
After checking location accuracy	28,378	55,339
After removing duplicates	28,157	54,927
After correcting invalid data	28,156	54,905

To verify the location accuracy of all other onshore wind turbines, an image recognition model for onshore wind turbines was trained using convolutional neural networks (CNN) and applied to the aerial images (from Google cloud service) of the given geo-coordinates of the wind turbines. For modelling, we used the Keras library and the TensorFlow framework as a backend as they are popular, can be used in R and seem to give the best results in binary image classification compared to other CNN approaches [20,21]. We choose a sequential model for binary classification (wind turbine present or not) and used three fully connected dense layer. The trained model provided sufficiently accurate results with an accuracy of 92% for a setting of 33×33 pixels. New installations from 2018 onwards could not be identified in most cases due to the fact that the aerial images provided by the Google Cloud service were older at the time of classification. Old installations that have already been dismantled before the time of classification could also not be identified as such. This leads to the fact that 7775 records had to be visually checked again manually. Of these, a total of 4786 wind power plants were manually corrected in their position because they contained either incorrect or inaccurate location information. Another 190 records with a total installed capacity of 330 MW could not be assigned to an exact location and were therefore removed from the data set.

In order to check whether there are still misplaced wind turbines, overlaps or duplicates of wind power plants, the distances to the nearest wind power plant were calculated in each case. Since the wind turbines should be at a distance of 2 to 3 times the rotor diameter from each other, depending on the wind direction, and the mean value of the rotor diameter of the data set was about 80 m, wind power plants that were less than 200 m apart were manually checked for correctness of position by aerial photograph comparison (Figure 2b). However, wind turbines could be close to each other if repowering or new construction of the turbine took place at the same or a nearby site and the old turbine had a decommissioning date. The distance check allowed further misplaced wind turbines to be corrected manually and 221 duplicates with a total installed capacity of 412 MW to be identified, which were removed from the data set.

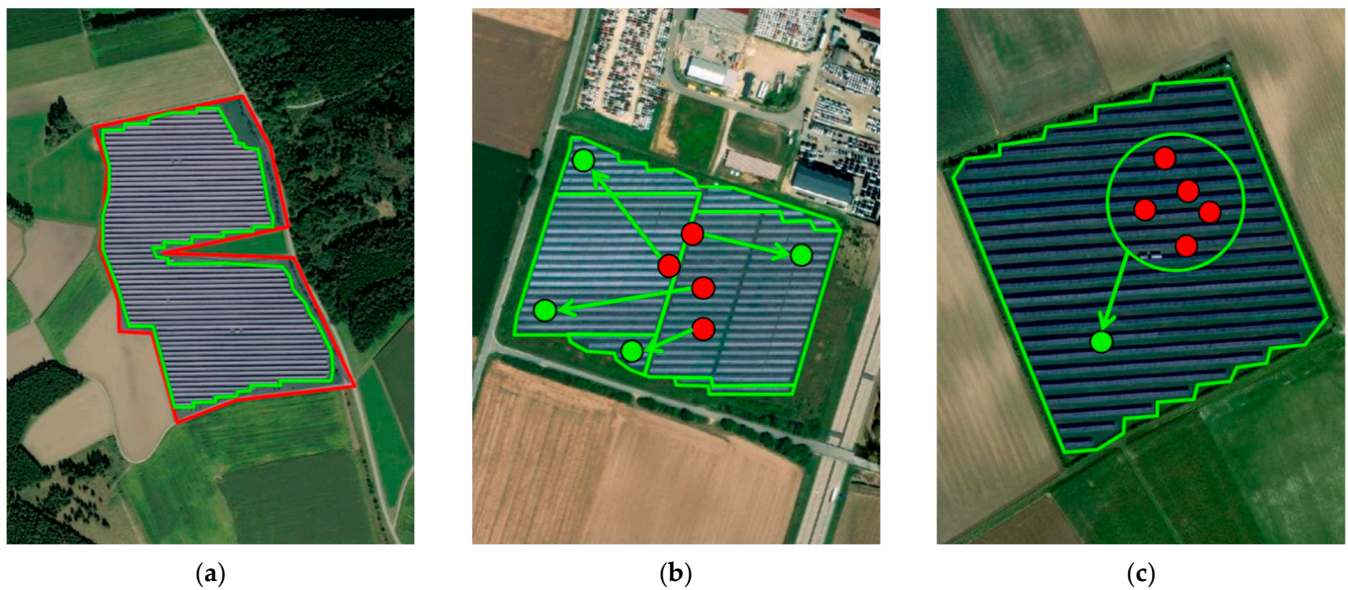


Figure 2. Processing steps in the compilation of the PV field system data set: (a) Adjustments to the PV areas. The areas outlined in red were manually adjusted according to the mapping specifications to the areas outlined in green. (b) Subdivision of PV areas and allocation of CEMDR data points (red dots) to the PV sub-areas (green dots). (c) Consolidation of CEMDR data points (red dots).

In a further step, missing or obviously incorrect information in the data, such as implausible hub heights or rotor diameters, has been deleted or, if possible, corrected by tracing the installation in the mentioned other available data sources with supposedly correct information. To check the plausibility of the hub height and rotor diameter, the hub heights were set in relation to the total height of the respective wind turbine (which is the sum of the hub height and half the rotor diameter) and checked for a value of less than 0.55 to ensure that the rotor blades are not longer than the hub height.

In a final step, still missing values for hub heights and rotor diameters were added using a method developed by [9]. For this purpose, a random forest was trained and applied to the data set. We used six predictor variables, four of which are technical parameters (installed capacity, hub height, rotor diameter and year of commissioning) and two of which define geographical location (latitude and longitude). Hub height and rotor diameter each represented the response variable. Thus, 598 missing hub heights and 385 rotor diameters were filled in, covering 417 records with one and 283 records with two missing variable values each. A total of four random forest were carried out to fill the gaps and a 4-fold cross-model validation was performed for each trained random forest, with better predictions for rotor diameter following the RSME and R^2 (Table 5).

Table 5. Predictor and response variables according Table 2 and the metrics of the 4-fold correlation validation for each trained random forest.

Predictor Variable	Response Variable	RMSE/R ²
CAP, COY, ROD, LAT, LON	HUB	15.1/0.76
CAP, COY, LAT, LON	HUB	15.8/0.73
CAP, COY, HUB, LAT, LON	ROD	8.4/0.89
CAP, COY, LAT, LON	ROD	8.8/0.88

The final data set of onshore wind turbines consists of 28,159 records with a total installed capacity of 54,905 MW. This is 1103 records less than the initial data set with a reduced total installed capacity of 728 MW, of which 56% of the capacity were identified as duplicates.

3.2. Photovoltaic Field Systems

Besides wind energy, solar energy is one of the most important sources of renewable energy. Thus, a number of Photovoltaic (PV) field systems have been built in recent years. These plants are elevated photovoltaic modules that are usually erected in the open countryside on arable land or grassland.

Since the CEMDR only contains point coordinates for the PV field systems, but the areal extent and size of these installations is important information depending on the context, these areas were mapped in the course of the data collection. For this purpose, existing area data of PV field systems were collected and merged from [8,22,23] before their geometries were adjusted in a GIS software according to the following specifications:

1. The mapping scale was 1:2500.
2. The outer edge of the visibly coherent PV system modules was always mapped (Figure 2a).
3. Areas within the geometries that did not contain PV modules were cut out from a diameter size of more than 25 m.

After the manual adjustment of the geometries, the PV area data set contained polygons with a total area of 20,346 hectares and could be merged with the CEMDR data extract to transfer the system data to the mapped areas. However, before this, the CEMDR data were cleaned of records with incorrect or missing geo-coordinates. Table 6 shows the completeness of the initial data of PV field systems of the CEMDR. Although the proportion of records without geo-coordinates is quite high at 32%, they only account for 0.002% of the total installed capacity and were therefore removed from the data set. These were mainly small PV systems that can be installed in home gardens, for example.

Table 6. Number, total capacity and number of missing data in the initial data set of PV field systems of the CEMDR.

Characteristics	Photovoltaic Field Systems
Number of Records	11,689
Total Capacity in MW	13,758
Missing Capacities	0
Missing Commissioning Dates	0
Missing Geo-coordinates	3826

Thus, 7853 PV records remained for further processing (Table 7). These records were first assigned one of ten categories of sky orientation (north, north-east, east, east-west, south-east, south, south-west, west, north-west or “sun tracked”) and one of five tilt levels from <20, 20–40, 40–60 to >60 or “sun tracked”, based on the information in the CEMDR extract.

Table 7. Changes in the number of records of photovoltaic field systems and total capacity during data processing.

Processing Step	Number of Records	Total Capacity in MW
Initial data set	11,689	13,758
After removing all records with missing localisation	7853	13,706
After allocation and correcting invalid data	6622	13,807

After this initial data preparation, the system data were transferred to the mapped PV areas. However, since the PV records were not always geographically located exactly on the corresponding mapped PV areas in order to perform a simple geographical spatial join, an allocation algorithm with the following logic was first applied:

1. The CEMDR record with the smallest spatial distance to a mapped PV area belongs to this area if all other conditions are met as well.
2. The PV area and the PV record had to be located in the same municipality (we used the local administrative units (LAU)).
3. The ratio between the specified plant capacity and the area had to be within the tolerance range of 0.7 to 1.5 MW/ha.
4. No other record could be assigned to the mapped area under consideration of the rule 1.

If the conditions were not met, the record that was second closest to the PV area was tested and so on. In this way, 2048 PV records could initially be assigned to a PV area, which already corresponded to 26% of the entire PV data set.

Records of PV field systems that could not be allocated by this algorithm were then in a further step manually assigned to a PV area in a GIS software. This brought the challenge of assigning them to the PV areas to which they actually belong, which was a process that was characterized by individual decisions, supported by additional information of the CEMDR (like field system names) and the use of aerial and satellite imagery (Google Maps or Sentinel Data). The manual allocation of PV data points to the mapped PV areas mainly affected PV field systems with initial incorrect coordinates or systems that had been expanded over time and therefore contained several records. If the latter was the case, the mapped areas of such contiguous PV field systems were manually subdivided in a GIS software into independent polygon geometries according to the number and information given in the associated PV records (Figure 2b). Where available and necessary, past aerial imagery was used with Google Earth Pro to support allocation decisions regarding the determination of PV area development over time. Data points that could be consolidated or fell on a PV area that could not be further differentiated, were summarized if the time of commissioning was within one year (Figure 2c). The most recent date was then taken as the commissioning date, as from this date the summarized system information is correct. In the end, 5805 PV records were manually checked and assigned to the obviously associated PV areas, of which ultimately 338 PV records could not be clearly allocated to a PV area and were therefore removed from the data set. This reduced the installed capacity of the PV data set by 269 MW. However, additional CEMDR data were included during this allocation procedure that were not considered PV field systems according to the CEMDR data extract but were classified as such by us. These included, for example, installations built on former landfill sites or in open-cast mining. The inserted data were given the attribute "Structural plants (other)" and in turn increased the total installed capacity of the PV data set by 392 MW.

In the course of the manual allocation, numerous PV areas that were not yet included in the PV area data set already compiled were also mapped and included in it, taking into account the mapping specifications introduced. Sentinel-2 satellite imagery was also used for this, which had the advantage of being more up-to-date compared to Google Maps. On the other hand, there was the disadvantage of lower ground resolution (at best 10×10 m). However, with

the Sentinel-2 imagery, most PV installations could be detected and mapped with reasonable accuracy, although there were limitations with very small installations (<0.3 ha).

In total, about 3000 data points had to be manually shifted to the correct PV area and an additional PV area of 2564 hectares was mapped to which a PV record could be assigned. The final data set thus consisted of PV field systems with a total installed capacity of 13,807 MW and a total area of 22,910 hectares.

3.3. Bioenergy Plants

Bioenergy plays an important role as a renewable energy source to compensate for fluctuations in electricity generation from wind and solar energy. In this context, it encompasses various technologies for electricity generation that are based on the use of biomass. The data extract of the CEMDR showed that about 4% of the entries did not contain information on geo-coordinates (Table 8). In addition, as with the wind power plant data, there were also several obviously incorrectly positioned bioenergy plants. However, unlike wind power plants or PV field systems, which are mostly located in open corridors, bioenergy plants can be located using a street address if no or incorrect geo-coordinates are provided, at least for those that have one stored in the initial data. For this purpose, the addresses of the records with missing geo-coordinates were converted into geo-coordinates using the geocoder of the Federal Agency for Cartography and Geodesy [24].

Table 8. Number, total capacity and number of missing data in the initial data set of bioenergy plants of the CEMDR.

Characteristics	Bioenergy Plants
Number of Records	19,941
Total Capacity in MW	8637
Missing Capacities	0
Missing Commissioning Dates	0
Missing Geo-coordinates	778

In total, there were 838 records with no or obviously incorrect information on geo-coordinates in the initial CEMDR data of the bioenergy plants. For 90% of these records, only the center of the municipality where the plants are located could be geolocated due to insufficient address information, but not their actual site. These were mainly plants with an installed capacity of less than 100 kW. For the remaining 10%, the exact location could be determined.

In order to avoid overlaps of data points in the graphical representation of the plant locations (e.g., because several generation units are located at one site), data points lying on top of each other were offset by a few meters so that each data point can be identified individually in a map viewer. Thus, there were a total of 2313 cases of duplicate or multiple overlapping data points with a total number of 5977 entries. Checking the data set for duplicate entries in relation to the plant-specific information resulted in 6260 cases with the same technical parameters. However, it turns out that these were usually not duplicate entries in the true sense, but mostly twin units, i.e., plants with the same technical characteristics and operated at the same location. Nevertheless, it could not be ruled out that there are duplications.

To determine the main type of biomass used by the respective bioenergy plant, a total of 27 fuel types (Bark, Biodiesel, Biogas (on-site electricity generation), Biogenic liquid waste, Biogenic solid waste, Biomethane, Biomethanol, Burning liquor, Firewood, Landfill gas, Landscape wood, Liquid biogenic substances, Palm oil, Pellets (wood), Reclaimed wood, Sewage gas, Solid biogenic substances, Straw and straw pellets, Sulphite liquor, Turpentine, Vegetable oil, Warm fuels (biogenic commercial waste), Waste wood, Wood, Wood chips, Wood scraps (e.g., joineries), Wood shavings and sawdust) were included in the data set by reading out the initial CEMDR data, which in turn are classified into the three biomass groups gaseous biofuels, solid biofuels and liquid biofuels.

In the case of gaseous biofuels, 91% are biogas (on-site electricity generation), which in turn account for 84% of the total data set. To exclude erroneous localization due to incorrect geo-coordinates of these installations, they were spatially associated with the latest Corine Land Cover data set (CLC 2018) [25]. Records that were located on land cover classes where they were not expected to be, such as forest, wetland or infrastructure, were manually, contextually checked. This led to a manual review of 1069 data sets where 296 records had incorrect geo-locations and were moved to the correct location based on the address in the original CEMDR record (Figure 3a). Only 1 record had to be removed from the data set because its geo-location could not be determined.

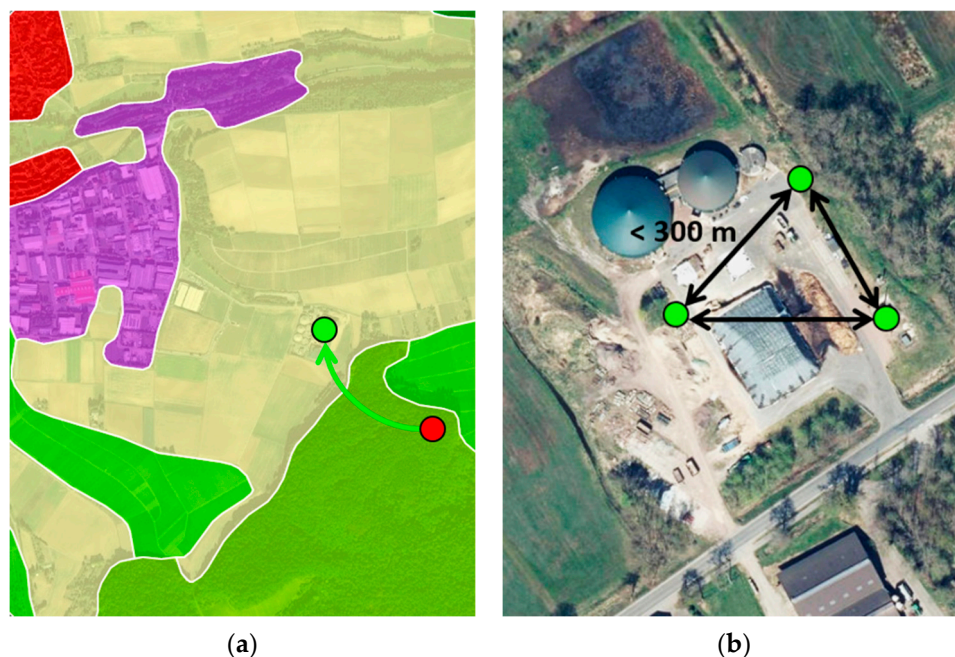


Figure 3. Processing steps in the compilation of the bioenergy data set: (a) Site verification of biogas plants (on-site electricity generation) using Corine Land Cover (colored areas). Data points that were located on land cover classes where they were not expected (red dot), such as forest, wetlands or infrastructure, were manually contextually checked and corrected if necessary (green dot). (b) Measurement of the distance between biogas plants (on-site electricity production) (green dots) in order to virtually group them into biogas production sites with a distance < 300 m from each other.

The biogas plants for on-site electricity generation are generally supplied with gas from a biogas digester in the immediate vicinity. To obtain information about which biogas plants for on-site electricity generation belongs to the same biogas production site, the distance of all biogas on-site electricity generation plants to each other was calculated. Those plants that were less than 300 m away from each other (assuming that this distance covers a typical biogas production site) were grouped by an individual identifier to indicate that these plants are fed by the same biogas production site (Figure 3b). This indexed 9701 virtual biogas production plants and corresponds approximately to the number of biogas production sites actually operated in Germany of 9692 for the year 2021 [26]. The discrepancy between the virtually grouped records and reality is mainly due to the fact that the on-site electricity generation plants are not always located within the selected threshold of 300 m, but in some cases may be somewhat further away.

After correcting all obviously incorrect data, such as unrealistically high information about installed capacities, the final bioenergy plant data set consists of 19,940 records with a total capacity of 8493 MW. Table 9 shows the distribution of the final bioenergy plant data set according to the main fuel groups of biomass used for generator operation, number of plants and installed capacity.

Table 9. Number of records and total installed capacity for the final bioenergy plant data set by fuel group (System).

System	Number of Records	Total Capacity in MW
Gaseous Biofuel	18,466	6602
Solid Biofuel	759	1745
Liquid Biofuel	715	147

3.4. Hydropower Plants

Hydropower is probably one of the longest-used renewable energy sources. There are plants in Germany that have been in operation for over a hundred years. Geographically, there are a particularly large number of hydropower plants in southern Germany, as the conditions for hydropower utilisation are favourable here in the high runoff and precipitation regions of the low mountain ranges and the Alpine region.

In the initial CEMDR data set there were 36 apparently incorrectly located hydropower plants to which the correct geo-coordinates could be assigned. However, the data extract from the CEMDR also showed that 44% of the records do not contain information on geo-coordinates (Table 10). The reason for this is the mainly private use of such plants with low installed generation capacities (less than 30 kW) which are therefore not fully published for data protection reasons [10]. Nevertheless, address data with varying completeness were available for these installations in the CEMDR data extract, from which geo-coordinates could be determined with the help of the geocoder of the Federal Agency for Cartography and Geodesy [24]. Unfortunately, for 3588 of these records, only the geo-coordinates of the centre of the municipality where the installations were located could be assigned. In addition, 15 hydropower generators were identified that are not located on the territory of the Federal Republic of Germany, but in the Alpine region of Austria. They all belong to a network of storage power plants with a total installed generator capacity of 639 MW.

Table 10. Number, total capacity and number of missing data in the initial data set of hydropower plants of the CEMDR.

Characteristics	Hydropower Plants
Number of Records	8046
Total Capacity in MW	6283
Missing Capacities	0
Missing Commissioning Dates	0
Missing Geo-coordinates	3590

The analysis of geographically overlapping records resulted in a total number of 3814 entries, distributed over 1294 cases of duplicate or multiple overlapping data points. They were all moved so that each data point could be identified individually in a map viewer. As with the bioenergy data set, no duplicate records were identified.

A total of five distinguishing features of hydropower plants were included in the data set, which were read from the original CEMDR data set. These were namely hydropower in drinking and service water systems, storage hydropower, wastewater hydropower and run-of-river hydropower. The latter are in turn divided into the three subtypes run-of-river power plants, diversion power plants and residual water power plants.

After correcting manifestly incorrect data and removing four records due to untraceable system data, the final hydropower data set consists of 8042 records with a total capacity of 5832 MW. In terms of the total number of hydropower records collected, run-of-river power plants account for 90%, with the group of run-of-river power plants with 54% followed by diversion power plants with 35% representing the largest subtypes (Table 11). In relation to the total installed hydropower capacity, however, run-of-river hydropower only accounts for 75%. The reason for this is the large storage hydropower plants with a share of 23% of the total installed capacity.

Table 11. Number of records and total installed capacity for the final hydropower plant data set by system characteristics.

System	Number of Records	Total Capacity in MW
Hydropower in Drinking Water System	326	28
Hydropower in Service Water System	147	44
Storage Hydropower	278	1370
Wastewater Hydropower	46	4
Run-of-river Hydropower	7241	4385

4. Discussion and Conclusions

The compilation of the data set of onshore and offshore wind power plants, PV field systems, bioenergy plants as well as hydropower plants for renewable electricity generation has shown that a significant part of the initial CEMDR data extract was incorrect and therefore had to be corrected. For this purpose, existing methods (e.g., [9]) were used, but new techniques (e.g., turbine image recognition, distance checks) were also introduced and further data sources were exploited to fill data gaps or locate misplaced renewable energy installations. In addition, the PV field systems data set was extended to include the corresponding areas taken up by the installations, which enables further analyses, e.g., the calculation of the area size or which land cover classes are affected by PV field plants. The latter aspect contributes to the research value of this particular data set of renewable power plants. In contrast to detailed regional and therefore decentralized data sets from regional planning associations, this data set presents the sites with detail on a national level. The data set is publicly available and can be found at [11]. It is provided as geodata in GeoJSON format, a widely used data format for spatial vector data that can be read by common GIS software.

Compared to the official figures reported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) in [1], the compiled data set shows good agreement in terms of total installed capacity for 2020 (Table 12). However, it should be noted that the official figures are based, among other sources, on the values recorded in the CEMDR [27].

Table 12. Comparative comparison of the installed capacities of the compiled data set and the figures officially reported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) for 2020 by type of renewable energy plant in MW.

Renewable Energy Installation	Compiled Data Set	Reported by [1]
Offshore Wind Power Plants	7764	7775
Onshore Wind Power Plants	54,116	54,414
Photovoltaic Field Systems	13,669	13,430 ¹
Bioenergy Plants	8412	9295
Hydropower Plants	5829	5436

¹ based on the assumption according to [28] that 25% of the total reported installed photovoltaic capacity of 53,721 MW is accounted for PV field systems.

Even though the compiled data set is almost complete, there are plants that are missing, which may explain the discrepancy in the figures in Table 12. This mainly concerns plants that (1) could not be assigned due to a lack of information and were therefore removed from the data set, (2) contained incorrect or outdated information in the CEMDR extract, and (3) were decommissioned before the official introduction of the CEMDR in 2017 and were therefore not included in the CEMDR data extract. However, according to our own estimates, the latter only affects a few plants, as most of the plants received a 20-year state subsidy with the introduction of the Renewable Energy Sources Act (EEG) in 2000 and should therefore most likely not be decommissioned before 2017. For plants for which no exact location could be determined, it was at least possible to identify the municipality in

which they are located. This applies above all to small hydropower plants, but also to some bioenergy plants.

The compiled data set allows to obtain a very accurate picture of the spatial and temporal distribution of renewable energies in Germany, which can be helpful for monitoring the transformation of the energy system. It can be used for socio-economic and environmental questions in research, infrastructure planning or political discussions, in perspective also at the EU level. For example, it can be made clear, which regions are well advanced in the expansion of renewable energies and which regions still need development assuming at the same time that the natural conditions of the respective region are very diverse. Therefore, the data set might also help social science studies in analyzing questions of justice related to the energy transition, e.g., discussing the urban-rural and the interregional relationship. The data can also be used as basic data for plant-specific modelling of electricity yields with weather data, as developed by [5,6] for wind and PV plants.

To simplify the mapping of PV areas or the detection of renewable energy installations, Artificial Intelligence-based image recognition algorithms could be used in the future to speed up data processing, as already tested in a use case by [29]. In addition, an application programming interface (API) could be established in the future to expose the data set and enable automatic retrieval and integration of the data into external applications [30]. This would improve the applicability of the data set and would have the advantage over a file-based publication of the data set as individual GeoJSON files that users of an API would always be up to date with updates or changes and thus always have a consistent data set.

To update the data set, it can be extended by the desired period. All that is needed is a current data extract from the CEMDR, a comparison with the unique CEMDR number and the deposited time stamp in the existing data set to update it in case of a change and to add new records. The work presented here can help to ensure an efficient and comprehensible process of data preparation during updates.

Author Contributions: Conceptualization, D.M.; methodology, D.M., L.G. and J.S.; validation, D.M., L.G. and J.S.; investigation, D.M., L.G. and J.S.; resources, D.M.; data curation, D.M., L.G. and J.S.; writing—original draft preparation, D.M. and N.M.; writing—review and editing, D.M.; visualization, D.M.; supervision, N.M. and D.T.; project administration, D.M., N.M. and D.T.; funding acquisition, N.M. and D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This data set compilation was funded by the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz, BfN) with grants from the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, BMUV), funding code: 3520860501.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used, the technical documentation and the compilation of the data set are available at <https://git.ufz.de/manske/regeoloc> (accessed on 23 August 2022). The final data set can be accessed at <https://doi.org/10.5281/zenodo.6922043>.

Acknowledgments: The authors would like to thank Jens Ponitka for his valuable contribution of data sources used in the data processing.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Federal Ministry for Economic Affairs and Climate Action (BMWK). Time Series for the Development of Renewable Energy Sources in Germany. 2022. Available online: <https://www.erneuerbare-energien.de> (accessed on 5 May 2022).

2. Kühne, O. Neue Landschaftskonflikte—Überlegungen zu den physischen Manifestationen der Energiewende auf der Grundlage der Konflikttheorie Ralf Dahrendorfs. In *Bausteine der Energiewende*; Kühne, O., Weber, F., Eds.; Springer Fachmedien: Wiesbaden, Germany, 2018; pp. 163–186. ISBN 978-3-658-19509-0.
3. Thrän, D.; Bunzel, K.; Klenke, R.; Koblenz, B.; Lorenz, C.; Majer, S.; Manske, D.; Massmann, E.; Oehmichen, G.; Peters, W.; et al. *Naturschutzfachliches Monitoring des Ausbaus der Erneuerbaren Energien im Strombereich und Entwicklung von Instrumenten zur Verminderung der Beeinträchtigung von Natur und Landschaft*; Bundesamt für Naturschutz: Bad Godesberg, Germany, 2020. [[CrossRef](#)]
4. Ponitka, J.; Boettner, S. Challenges of Future Energy Landscapes in Germany—A Nature Conservation Perspective. *Energy Sustain. Soc.* **2020**, *10*, 17. [[CrossRef](#)]
5. Lehneis, R.; Manske, D.; Thrän, D. Generation of Spatiotemporally Resolved Power Production Data of PV Systems in Germany. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 621. [[CrossRef](#)]
6. Lehneis, R.; Manske, D.; Thrän, D. Modeling of the German Wind Power Production with High Spatiotemporal Resolution. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 104. [[CrossRef](#)]
7. Dunnett, S.; Sorichetta, A.; Taylor, G.; Eigenbrod, F. Harmonised Global Datasets of Wind and Solar Farm Locations and Power. *Sci. Data* **2020**, *7*, 1–12. [[CrossRef](#)] [[PubMed](#)]
8. Eichhorn, M.; Scheffelowitz, M.; Reichmuth, M.; Lorenz, C.; Louca, K.; Schiffler, A.; Keuneke, R.; Bauschmann, M.; Ponitka, J.; Manske, D.; et al. Spatial Distribution of Wind Turbines, Photovoltaic Field Systems, Bioenergy, and River Hydro Power Plants in Germany. *Data* **2019**, *4*, 29. [[CrossRef](#)]
9. Becker, R.; Thrän, D. Completion of Wind Turbine Data Sets for Wind Integration Studies Applying Random Forests and k-Nearest Neighbors. *Appl. Energy* **2017**, *208*, 252–262. [[CrossRef](#)]
10. Federal Network Agency (BNetzA). Core Energy Market Data Register (MaStR). Available online: <https://www.marktstammdatenregister.de/MaStR/> (accessed on 7 May 2021).
11. Manske, D.; Grosch, L.; Schmiedt, J. *Geo-Locations and System Data of Renewable Energy Installations in Germany [Data Set]*; Version V20210507; Zenodo: Genève, Switzerland, 2022. [[CrossRef](#)]
12. Manske, D. Regeoloc. Gitlab Repository. 2022. Available online: <https://git.ufz.de/manske/regeoloc> (accessed on 23 August 2022).
13. Amprion GmbH; TransnetBW GmbH; TenneT TSO GmbH; 50 Hertz Transmission GmbH. EEG-Anlagenstammdaten. Available online: <https://www.netztransparenz.de/EEG/Anlagenstammdaten> (accessed on 14 May 2021).
14. Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie (StMWi). Energieatlas Bayern. Available online: <http://www.energieatlas.bayern.de> (accessed on 1 June 2021).
15. Hessisches Landesamt für Naturschutz, Umwelt und Geologie (HLNUG). Windenergie in Hessen. Available online: <https://www.hlnug.de/themen/luft/windenergie-in-hessen> (accessed on 7 June 2021).
16. Landesanstalt für Umwelt Baden-Württemberg (LUBW). Daten- und Kartendienst der LUBW. Available online: <https://udo.lubw.baden-wuerttemberg.de/public/index.xhtml> (accessed on 7 June 2021).
17. Deutsche Energie-Agentur (Dena). Einspeiseatlas. Available online: <https://www.biogaspartner.de/einspeiseatlas/> (accessed on 30 March 2022).
18. Pierrot, M. The Wind Power. Available online: <https://www.thewindpower.net/> (accessed on 13 July 2021).
19. Butler, H.; Daly, M.; Doyle, A.; Gillies, S.; Schaub, T.; Hagen, S. *The GeoJSON Format*; Internet Engineering Task Force: Fremont, CA, USA, 2016. [[CrossRef](#)]
20. Kim, S.; Wimmer, H.; Kim, J. Analysis of Deep Learning Libraries: Keras, PyTorch, and MXnet. In Proceedings of the 2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA), Las Vegas, NV, USA, 25–27 May 2022; pp. 54–62. [[CrossRef](#)]
21. Lee, H.; Song, J. Introduction to Convolutional Neural Network Using Keras; an Understanding from a Statistician. *Commun. Stat. Appl. Methods* **2019**, *26*, 591–610. [[CrossRef](#)]
22. Federal Agency for Cartography and Geodesy (BKG). Digitales Basis-Landschaftsmodell. Available online: <https://gdz.bkg.bund.de/index.php/default/digitales-basis-landschaftsmodell-ebenen-basis-dlm-ebenen.html> (accessed on 17 June 2021).
23. Geofabrik. OpenStreetMap Data Extracts. Available online: <http://download.geofabrik.de/> (accessed on 23 June 2021).
24. Federal Agency for Cartography and Geodesy (BKG). BKG Geocoder. Available online: <https://gdz.bkg.bund.de/index.php/default/webanwendungen/bkg-geocoder.html> (accessed on 13 October 2021).
25. Copernicus Programme. CORINE Land Cover. Available online: <https://land.copernicus.eu/pan-european/corine-land-cover> (accessed on 11 May 2022).
26. Fachverband Biogas. Branchenzahlen 2020 und Prognose der Branchenentwicklung. 2021. Available online: [https://www.biogas.org/edcom/webfvb.nsf/id/DE_Branchenzahlen/\\$file/21-10-14_Biogas_Branchenzahlen-2020_Prognose-2021.pdf](https://www.biogas.org/edcom/webfvb.nsf/id/DE_Branchenzahlen/$file/21-10-14_Biogas_Branchenzahlen-2020_Prognose-2021.pdf) (accessed on 28 July 2022).
27. Walker, M.; Bickel, D.P.; Musiol, D.F.; Nieder, T.; Schneider, S.; Schrempf, L.; Memmler, M. Datenquellen und Methodik der AGEE-Stat-Zeitserien zur Entwicklung der erneuerbaren Energien in Deutschland. Stromerzeugung und installierte Leistung. 2016. Available online: https://www.umweltbundesamt.de/sites/default/files/medien/377/publikationen/2016-1-15_dokumentation_agee-stat-zr_stromerzeugung_leistung_final.pdf (accessed on 24 August 2022).
28. Lewicki, P. Photovoltaik. 2021. Available online: <https://www.umweltbundesamt.de/themen/klima-energie/erneuerbare-energien/photovoltaik> (accessed on 24 August 2022).

-
29. Schulz, M.; Boughattas, B.; Wendel, F. DetEEktor: Mask R-CNN Based Neural Network for Energy Plant Identification on Aerial Photographs. *Energy AI* **2021**, *5*, 100069. [[CrossRef](#)]
 30. Górski, T.; Wojtach, E. Use Case API-Design Pattern for Shared Data. In Proceedings of the 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 18–20 December 2018; pp. 1–8. [[CrossRef](#)]