https://doi.org/10.20378/irbo-50523

A Decision Support System for Photovoltaic Potential Estimation

Konstantin Hopf[1] konstantin.hopf@uni-bamberg.de
Michael Kormann[1] michael.kormann@gmx.de
Mariya Sodenkamp[1] mariya.sodenkamp@uni-bamberg.de
Thorsten Staake[1,2] thorsten.staake@uni-bamberg.de

[1)] Management Information Systems / Energy Efficient Systems, University of Bamberg, 96045 Bamberg, Tel: +49 951 863 2236
[2)] Department of Management, Technology and Economics, ETH Zurich

ABSTRACT

With knowledge on the photovoltaic potential of individual residential buildings, solar companies, energy service providers and electric utilities can identify suitable customers for new PV installations and directly address them in renewable energy rollout and maintenance campaigns. However, many currently used solutions for the simulation of energy generation require detailed information about houses (roof tilt, shading, etc.) that is usually not available at scale. On the other hand, the methodologies enabling extraction of such details require costly remote-sensing data from three-dimensional (3D) laser scanners or aerial images. To bridge this gap, we present a decision support system (DSS) that estimates the potential amount of electric energy that could be generated at a given location if a photovoltaic system would be installed. The DSS automatically generates insights about photovoltaic yields of individual roofs by analyzing freely available data sources, including the crowdsourced volunteered geospatial information systems OpenStreetMap and climate databases. The resulting estimates pose a valuable foundation for selecting the most prospective households (e.g., for personal visit and screening by an expert) and targeted solar panel kit offerings, ultimately leading to significant reduction of manual human efforts, and to cost-effective personalized renewables adoption.

CCS CONCEPTS

• ***Information systems~Decision support systems***
• ***Information systems~Data analytics*** • *Information systems~Location based services* • *Information systems~Data mining* • *Hardware~Renewable energy*

KEYWORDS

Crowdsourced Data, Volunteered Geographic Information (VGI), Sensory Data, Data Analysis, Solar Potential, Photovoltaic, Renewable Energy

IML '17, October 17-18, 2017, Liverpool, United Kingdom

ACM ISBN 978-1-4503-5243-7/17/10...$15.00
http://dx.doi.org/10.1145/3109761.3109764

1 INTRODUCTION

Photovoltaic (PV) is one of the most promising energy suppliers in the future energy system and was the second-largest source of newly built renewable energy capacity in 2015 [26]. According to a recent study by Gagnon et al. [17], 39% of U.S. national electric-sector sales could be covered by PV installations on rooftops. By the end of 2015, the cumulative installed solar PV power capacity world-wide was 229 GW, but new investments decline currently due to the drop of subsidies (e.g., attractive incentive programs in Europe ended or will end in the near future) and concerns of investors on how fast renewables can be integrated in the grid infrastructure. These sorrows reduce the long-term investor conviction to invest in PV [25, 28, 53]. Nevertheless, the political will is to achieve large extension of renewables. For example, the EU members committed themselves to the binding goal that at least 27 % of consumed energy shall be produced by renewables by 2030 [13], in 2013 this portion was only 11.8% [14].

Rising energy prices in the future [57] will make PV investments profitable [4], and make them highly attractive for self-consumption or storage settings in residential home owners that can convert their rooftop into a profitable local solar plant already now.

One barrier for private investors to adopt solar installations on their rooftop is their unawareness of the actual potential of their home [55], because they are often unaware of the important determinants for the solar potential of their housing (different rooftop types, tilt, orientation, objects causing shadow, etc.) and how to evaluate the relevant variables for such an investment decision. On the other hand, PV providers go astray manually collecting and updating information about houses in the potentially appropriate areas. Thereby, knowledge on the solar PV potential of single residential building roofs is extremely valuable for solar companies, energy service providers and utilities. By having this location-based information for a large number of residencies energy companies can then select the most suitable households to promote new PV installations or maintenance support for already plugged-in solar panels. Furthermore, having these insights utilities and regional communities (including city planners and energy policy makers) can make more informed decisions about regional renewable energy strategies, better plan their local smart grid infrastructure development and design targeted green incentive campaigns [3, 61]. This will consequently lead to the increase of renewable energy share and support the achievement of ambitious sustainability goals.

The automatic prediction of the roof PV potential has been a subject to considerable amount of research, but existing studies rely on expensive data collection and are therefore regionally limited. ***The work presented in this paper goes beyond the state-of-the-art by presenting a new data-mining-based DSS that utilizes freely available sensory and crowdsourced Volunteered Geographic Information (VGI) data from OpenStreetMap as well as solar irradiation and temperature to automatically assess the PV potential of individual residential building roofs.***

VGI digital data sources have emerged and millions of companies donate their information while users constantly generating new data online [9]. One prominent example is the project OpenStreetMap (OSM), that contains currently 3.8 billion entries on geographic places, streets, buildings, roads etc. [46]. Alongside preloaded digital maps from around the globe, the information within OSM is currently enhanced by a large group of volunteers who upload their sensory information such as satellite images, GPS tracks, but also field surveys to adjust and enrich the data for a region of interest. Some recent works already use OSM data for the PV potential prediction. In [31, 58] the total world solar energy potential on building roofs was estimated. Mainzer et al. [38] tested OSM data with satellite images to calculate the summarized areas of all building roofs in two cities

The methodology presented herein estimates a range of annual electricity energy (in kWh) that can be produced by a PV installation covering individual roofs. We combine several data mining related techniques to develop the decision support system [34]: our approach combines knowledge modeling with predictive analytics. Based on the location data of buildings and related unstructured crowdsourced geographic information [21, 65], we infer knowledge about private houses that is relevant for the PV potential estimation. On the other hand, quantitative domain-specific predictive models are applied for the roof area estimation and the amount of potentially generated PV energy at the given location. Thus, we apply knowledge on PV generation and geographic data to interpret and extract relevant features from OSM database, as well as use Monte-Carlo-Simulation [45] and mathematical models for electricity yield to estimate roof area and PV energy generation. From the engineering viewpoint, energy production yield across the roof was defined by Hookwijk [23] which is the solar energy irradiated at the earth point (known as theoretical or geographical potential) lowered by several influence factors (e.g., conversion efficiency, shading, or losses due to cabling or transformation). The amount of PV power that is actually installed on building roofs can be lower than this estimated technical potential by considering non-technical aspects (e.g., available investment capital, subsidies, laws).

We have made four assumptions that are justified in the following sections: our approach works for residential buildings with (i) a rectangular basal area, that have (ii) a gabled rooftop with (iii) a tilt-angle of 35°. Besides that, (iv) structural limitations to the rooftop area (such as roof windows, antennas or chimneys) are not assessed individually (due to the lack of data), but included as averages.

This remainder of the paper is structured as follows: An overview to related work is given in the next section. We describe the methodology in Section 3. In Section 4, we show the results of validation with two real-world datasets, and in Section 5, we discuss possible improvements and current limitations.

2 LITERATURE REVIEW

Multiple studies suggest approaches for the prediction of the roof PV potential. The existing works rely upon various data sources with data granularities ranging from detailed three-dimensional data with a raster size of lower than 0.5m, raised in locally limited remote sensing studies [e.g., 33, 48], to globally aggregated statistics [e.g., 31, 58]. The data types include airborne Light Detection and Ranging (LiDAR) data, aerial images from satellite photogrammetry [47, 64], digital earth surface models [18], 3-dimensional (3D) building data [60] or two-dimensional (2D) cadaster data from official land surveying offices. An interested reader is referred to [5, 16, 41, 52] for detailed surveys on the PV potential estimation in urban landscapes. The existing approaches can be divided into four categories that we briefly describe below.

Constant-Value Methods: These works use generalized statistics (e.g., population, gross domestic product, construction statistics in a country) to roughly estimate the overall solar potential of larger regions [e.g., 36]. Due to the simplifications and assumptions, the methods are unsuitable to discriminate between individual buildings.

Manual Selection and Sampling Methods: Samples of rooftops from limited study areas are used to assess the typical solar potential of related buildings. The solar potential estimates for the selected rooftops are then calculated, based on aerial images, three-dimensional LiDAR data, or manual considerations, and the resulting estimate per building is extrapolated to the whole study region [e.g., 6, 47, 64]. Some works seek to draw correlations between population density and available roof area [e.g., 61] to support their estimation. Such manual sampling of houses and their solar potential assessment are hard to automate.

Solar potential estimation based on digital terrain models: Works in this category [e.g., 18, 31, 58] focus on the solar capabilities of complete regions with digital terrain models and environmental data (solar irradiation and weather), using geographic information software (such as ESRI ArcGIS, GRASS GIS, or others). The results are obviously more reliable than those from constant-value works, but assessment of individual roof solar potential is also not possible.

Solar potential estimation of individual roofs: Multiple studies on the solar potential of individual rooftops have been conducted, using 3-dimensional LiDAR data [e.g., 22, 27, 30, 33, 37, 44, 59]. The works rely on advanced cartographic collections from regionally limited remote sensing studies and show the possibility to assess the photovoltaic perspective for individual roofs at a high level of detail. Some authors [7, 12, 48, 49] even estimate the solar potential of building facades for vertical installations. These models achieve performance of up to 9% root-mean-square deviation from the real PV production figures [29]. However, these studies are regionally limited, since expensive data collection is

necessary. Moreover, solar companies and utilities usually do not have access to this kind of detailed data.

3 METHODOLOGY

Figure 1 illustrates the decision support methodology, which encompasses six steps enumerated in the figure and consequently described in the following subsection. As an input, the artifact employs the postal address of a residential building. In addition, the method relies on freely available sensory and crowdsourced data: building geometries from OSM, solar irradiation and temperature. Further the model requires the values of parameters and influence factors (solar panel efficiency, roof shading etc.) to derive the expected range of PV yield on the rooftop per annum.

3.1 Retrieval of the building geometry from OpenStreetMap

At the initial stage, the shape of the building base as a polygon (specified by the coordinates of its corners) is retrieved from the OSM web service. After online request, the given address of the residency is converted to geocoded coordinates.

The data provided by the OSM web service consists of points and polylines on a 2D map, annotated with so called "tags". These tags give semantic meaning to the objects (e.g., they identify lines as streets and polygons as buildings). Tags also contain further information to describe the geometries (e.g., street / building type, name, or opening hours of shops). Technically, tags are key-value pairs that users can add to objects. The OSM community maintains a comprehensive taxonomy of recommended tags, but the existence and quality of the tags associated with objects in the database varies to a large extend [1]. In our implementation, we select the closest building (polygon tagged with the key "building") to the given location. Buildings with a larger Euclidean distance than 50m to the location are excluded to avoid errors that are caused by the lack of data in OSM (this was the case in ca. 15% of our tested addresses).

The building type does not actually influence the roof area estimation, but we extract this information from OSM and use it in our validation. We distinguish thereby between *residential buildings* (that have an estimated rooftop area of lower than 400m^2 and are tagged with the OSM key *building*, together with one of the values: *apartments*, *detached*, *dwelling_house*, *house*, *residential*, *terrace semi_detached*, *semidetached_house*,) and *other buildings* (e.g., commercial, industrial or unspecified buildings). For our purposes, only residential buildings are of interest.

3.2 Estimation of the rooftop area (A_c) and orientation in the space (β)

From the shape of the building base, we extract the rooftop area available for PV installations and the roof orientation in space.

The roof area is mainly determined by the roof type. Since roof type information is rarely existent in OSM [19] we consider the most frequent roof type *gabled roof* in official cadaster data (Table 1 shows the distribution of roof types based on [2] in a random sample of 3,627 buildings in Southern Germany). As the rooftop tilt, we consider α = 35° for all buildings, according to previous studies [36].

We calculate the *available rooftop area for solar installations* A_c with the building footprint area A_F and the rooftop tilt α [36]:

$$A_c = \frac{1}{2} * \frac{A_F}{\cos(\alpha)} \quad (1)$$

In the VGI data, no structural limitations to the available rooftop area (like windows, antennas or chimneys) are included and therefore we leave them out in this study, because even in

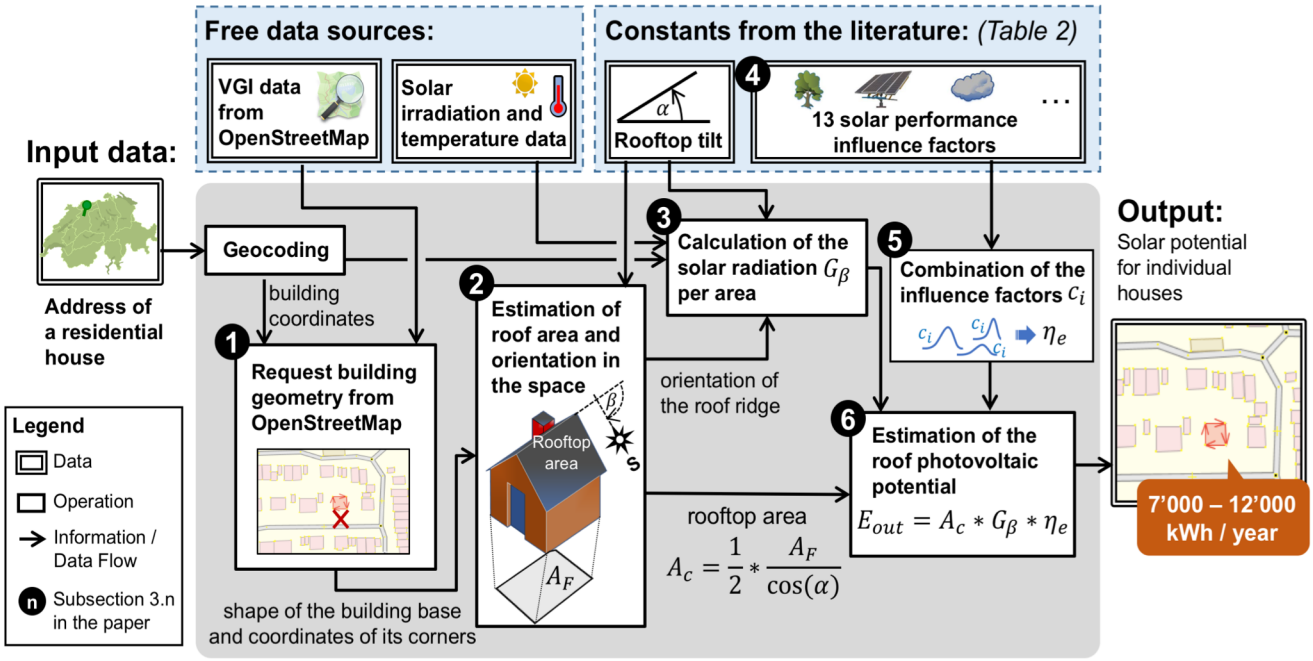

Figure 1: Decision support methodology for estimating solar energy production potential of individual building roofs

studies using highly detailed 3D data, the exact identification of such limitations was not possible [30, 41].

Table 1: Roof types [2] and their frequency in a random sample of 3,627 buildings in Southern Germany.

Roof type	Gabled roofs	Flat roofs	Other (11 types)
			
Frequency	2,328	540	759
Relative frequency	64%	15%	21%

To determine the *roof ridge orientation*, we use the building footprint corner coordinates, identify the longer side of the building and take the angle towards the sun, since a large majority of buildings are predominantly rectangular [54].

3.3 Computation of the amount of solar irradiation per area (G_β)

For the PV potential estimation, the amount of solar energy per area irradiated on the building roof and converted by PV modules to electrical energy is needed. We use Lamigueiro's [32] methodology and implementation that employs monthly solar radiation and temperature data at the specific location of the building (we use data from EUMETSAT [50]), together with the roof ridge orientation β and the roof tilt α to calculate the amount of energy. Since the measurements are faced with inaccuracies and variations (local weather conditions, reflections, etc.), we use the 10-year average readings in order to get a general picture of the solar potential at a specific location.

3.4 Definition of the main features that influence PV yield

The conversion of the irradiated solar energy to electric energy that can be fed into the power grid is subject to losses. We conducted a comprehensive literature review to identify the factors that influence the PV electricity generation, and found 13 factors that we list Table 2 together with published statistics (min., average, max.).

Solar panel efficiency (c_1) refers to the percentage of solar energy that can be technically transformed into electricity by a PV installation. The PV panel efficiencies differ heavily between manufactures [63] and become more efficient with progress in the technical development. With the efficiency of up to 25%, silicon crystalline is today one of the most efficient solar panels. In our estimation, we assume an average efficiency of 16% which was a common standard in the year 2012 when the solar panels of our validation-data were installed [20].

The solar electric energy production is faced with *environmental influences*. First, shading (c_2) reflects what percentage of the roof area is shaded (e.g., by a tree or by neighboring buildings). Dust, snow and other soiling on the surface of a PV module (c_3) prevent solar radiation from reaching the solar cells thus lowering the efficiency. Furthermore, the system shutdowns due to maintenance, grid outages, etc. reduce energy availability and output (c_4).

Besides that, *technical losses* lower the solar energy production. Within the solar cell where direct current (DC) electricity is produced, energy is lost due the wire connection between inverters, transformers and other parts of the installation (c_5). Inverter losses (c_6) happen during the conversion of DC in alternating current (AC) electricity mode. Cable mismatch (c_7) describes the electrical losses caused by slight differences of the manufacturing imperfections between modules in the array and different current-voltage characteristics. The initial light-induced degradation (c_8) describes the deposit of oxygen with silicon caused by a chemical process inside crystalline silicon solar cells during the photovoltaic effect. Further losses arise in the transportation of AC power (c_9), losses due to diodes (c_{10}) and connections of the solar installation and of the transformers (c_{11}) are also considered in the literature.

Table 2: Variables and constants used in our methodology to estimate the PV potential of building roofs

Estimated variables					
A_C	Available rooftop area, see (1)	E_{out}	Estimated PV potential		
A_F	Area of the building footprint	G_β	Mean annual solar irradiation at the roof location		
β	Roof orientation in space	η_e	Power conversion efficiency coefficient		
Constants from literature					
Symbol and name		***Value***			***Ref.***
α	Rooftop tilt	35°			[36, 38]
c_1	Solar panel efficiency	0.16 / 0.25			[20]
		Min, c_i^{min}	***Average***, c_i^*	***Max***, c_i^{max}	
c_2	Shading	0.00	1.00	1.00	[11, 39, 47]
c_3	Soiling	0.75	0.95	0.98	[11, 39, 40]
c_4	Availability	0.00	0.98	0.995	[6, 11, 39]
c_5	Wiring AC	0.98	0.99	0.993	[11, 39]
c_6	Inverter	0.93	0.96	0.98	[4, 11, 39]
c_7	Cable mismatch	0.97	0.98	0.985	[11, 39]
c_8	Initial light-induced degradation	0.90	0.98	0.99	[6, 35, 39]
c_9	Cabling DC	0.97	0.98	0.99	[11, 39]
c_{10}	Diodes and connections	0.99	0.995	0.997	[11, 39]
c_{11}	Transformers	0.96	0.97	0.98	[11, 39]
c_{12}	Manufacturer's nameplate rating	0.85	1.00	1.05	[11, 39, 56]
c_{13}	Error rooftop area (availability for panels)	0.49	0.73	0.95	[43, 51, 61]

Finally, we consider two *other coefficients*. The manufacturer's nameplate rating (c_{12}) is the differences between the solar panel efficiency figures published by the manufacturer and efficiency values that are measured under standard test conditions. The available rooftop has to be reduced due to structural limitations (e.g., windows, chimneys, antennas) by a ratio of the complete rooftop area and the area available for PV (c_{13}).

3.5 Definition of a cumulative performance measure (η_e) based on Monte Carlo simulation

The identified influence factors must be combined to one single *power conversion efficiency coefficient* η_e that is used in the PV potential estimation. A simple multiplication of all factors would ignore the distribution of each factor and leads to a large difference between the minimum and maximum estimated PV potential, due to the large range of some factors (e.g., c_2, c_4, c_{13}). Therefore, we use repeated random sampling for the aggregation of all influence factors $c_2, \dots, c_{13}$. This method is known as Monte Carlo simulation and has its application in math, physics, and business, when probabilistic problems with multiple variables must be solved [15].

In the Monte Carlo simulation, we assume all influence factors $c_2, \dots, c_{13}$ to be independent from each other and they can take a random value between c_i^{min} and c_i^{max}, with the arithmetic mean of c_i^*. The solar panel efficiency (c_1) is considered as a constant, because this factor depends on the technological state of the art and we use it as a parameter to the calculation. We approximate the cumulative density function $C_i(x)$ for each coefficient stepwise, using the cumulative density function $F_{\mu;\sigma}(x)$ of the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where we define the mean as $\mu = c_i^*$, and the standard deviation by $\sigma_{min;i} = \frac{1}{z}(c_i^* - c_i^{min})$ and $\sigma_{max;i} = \frac{1}{z}(c_i^{max} - c_i^*)$ according to c_i^{min} and c_i^{max}; z equals to the number standard deviations between c_i^* and c_i^{min} / c_i^{max}:

$$C_i(x) = \begin{cases} F_{\mu;\sigma_{min;i}}(x) : x < c_i^* \\ c_i^* \quad\quad : x = c_i^* \\ F_{\mu;\sigma_{max;i}}(x) : x > c_i^* \end{cases} \quad (2)$$

For the Monte Carlo simulation, we generated 10,000 independent random values for each influence factor, following the distribution $C_i(x)$ that are within the range of the respective coefficient $[c_i^{min}; c_i^{max}]$. We calculate the aggregated influence factors η_i^* as the product Π of all coefficients, according to [6]:

$$\eta_i^* = \prod_{k=1}^{n=13} c_k = c_1 * c_2 * \dots * c_n \quad (3)$$

The resulting distribution of η_i^* is shown in Figure 2 for $z \in \{2,3,4,5\}$. We choose $z = 3$ for the use in our implementation, because 99.73% of all values in the normal distribution are within the interval of $[\mu - 3\sigma;\ \mu - 3\sigma]$ and the distribution of η_e seems not to be overfitted.

As the result of the aggregated PV performance influence factors, we compute the *power conversion efficiency coefficient* η_e as the expected value of the aggregated η_i^* values:

$$\eta_e = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \eta_i^* \quad (4)$$

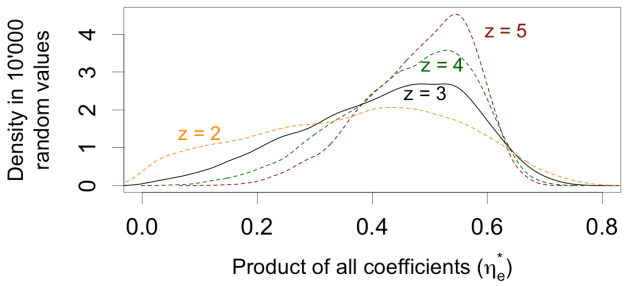

Figure 2: Distribution of the aggregated influence factors η_i^* as a result of Monte Carlo simulation for different numbers of standard deviations (z)

3.6 Roof photovoltaic potential estimation

To finally assess the electric energy E_{out} that can be generated by a PV installation and fed into the grid, we consider Hofierka and Kaňuk's model [22] with three determinants (Equation 5): *Available rooftop* area for solar cell installation A_C (in m^2), *annual solar irradiation* at the roof location G_β (in Wh/m^2), and mean annual *power conversion efficiency coefficient* (power input from the sun / power output from the system) η_e from Equation 4.

$$E_{out} = A_c * G_\beta * \eta_e \quad (5)$$

As an extension to Hofierka and Kaňuk's [22] model, we express the vagueness of this estimation with an interval rather than with a single value and replace η_e with an interval $[\overline{\eta_e}, \underline{\eta_e}]$ that contains 90% of the values for η_i^* (Equation 3). We define $\overline{\eta_e}$ as the 5%-quintile and $\underline{\eta_e}$ as the 95%-quintile of the distribution of η_i^*. Therefore, we provide an estimation of E_{out} and a range of the PV production $[\underline{E_{out}}; \overline{E_{out}}]$.

4 VALIDATION

We make a two-fold validation of our approach. On the one hand, we validate our estimation of the rooftop area A_c using official 3D cadaster data from the Bavarian land surveying office [2], containing detailed information on the roofs of 3,627 buildings. On the other hand, we validate our estimation of the possible PV electricity production E_{out} using real-world production data from 85,806 existing solar collector installations throughout Germany [10]. Our primary focus in both validations lies – in line with the goal of this paper – in assessing the predictive quality of our method for the solar potential estimation of residential buildings with gabled rooftops. Therefore, we distinguish between these buildings and other buildings (e.g. industrial or unspecified building types), and other/flat rooftops, if this information is available to us. The locations of all validation data are illustrated in Figure 3. For both validations, we provide detailed dataset descriptions, followed by the results and an interpretation in the sections below.

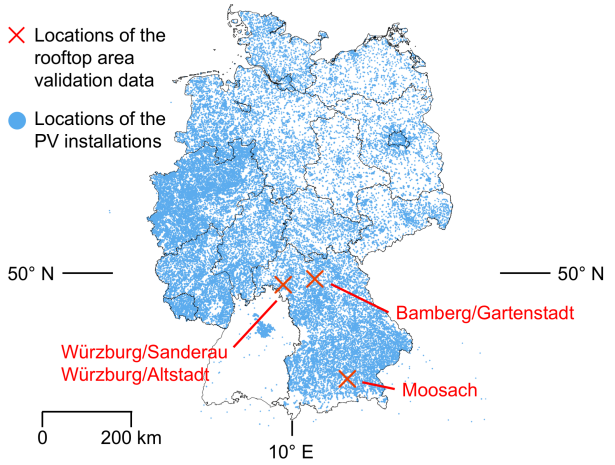

Figure 3: Map of Germany showing the places for rooftop area validation and the locations of the PV installations considered for the validation of PV potential estimates

4.1 Validation of the rooftop area estimate

4.1.1 Validation dataset

The Bavarian land surveying office provided us with laser-scanned 3D cadaster-data for n_A =3,627 houses located in three places in Southern Germany. We selected the places in cooperation with the land surveying office, following the motivation to include residential houses in the countryside, in villages and in suburbs, according to the categorization of Lödl et al. [36]. We chose therefore study areas that show different townscapes: *Würzburg/ Altstadt* is an old district area of a large town, *Würzburg/Sanderau* and *Bamberg/Gartenstadt* are newer districts of towns that are characterized by many residential buildings. In contrast, the village *Moosach* is a rural area with larger buildings and more open space.

The cadaster-data contains information on the buildings and the corresponding rooftops. Each building has one or more roofs associated. For each roof, the actual area $\ddot{A}_C^i$ and the roof type (see Table 1) is known.

4.1.2 Validation results

We validate the methodology for estimation of rooftop area by comparing the calculated area A_C^i (based on OSM data) with the true area $\ddot{A}_C^i$ (based on cadaster data) and calculating the error $\delta_i^A = A_C^i - \ddot{A}_C^i$ for each building i. As measure for the model performance, we use the mean absolute error $MAE = \frac{1}{n_A}\sum_{i=1}^{n_A}\|\delta_i^A\|$ and the model bias error $MBE = \frac{1}{n_A}\sum_{i=1}^{n_A}\delta_i^A$ according to [62] and present the results in Table 3. Negative values of MBE indicate an underestimation. The results are separated by the building type (the categories *residential buildings* and *other buildings* are based on the OSM data, as we describe in Section 3.1) and the roof type (categories *gabled rooftops* and *flat / other rooftops*, based on the information from the cadaster data, as included in Table 1). Existing studies that estimate the PV suitable rooftop area only rarely validate their results with real world data. In the lack of such benchmarks, we consider two random guess estimators (see Table 4) that we use to compare the MAE values with:

1. Random guess: This estimator assumes that all buildings have the same average rooftop size. In the lack of existing statistics on the roof area in Germany, we consider the average floor area of 91.4 m^2 where residents in Germany are living on [8] and take the average rooftop area of 111.58 m^2 (assuming a rooftop tilt of 35°) as a benchmark for the rooftop area estimation.

2. Biased random guess: This estimator assumes that all buildings with the same roof type to have equal roof areas. We take the average rooftop area in our cadaster validation dataset for each of the three rooftop types (gabled, flat, other) and take these values as the estimated rooftop area. We assume that the roof type is known. The MBE for all roofs is therefore 0.

In the main category of interest – residential buildings with a gabled rooftop – our algorithm has an average prediction error of 20.14 m^2. This is 27% of the mean residential building rooftop area (gabled roofs) in our validation data. The estimation for flat/other roof types and other building types is less accurate.

Table 3: Performance of rooftop estimation based on OSM data for different building and rooftop types

	Residential buildings			Other buildings			Residential and other buildings		
Roof type	MAE (m^2)	*MBE (m^2)*	Num. of buildings	MAE (m^2)	*MBE (m^2)*	Num. of buildings	MAE (m^2)	*MBE (m^2)*	Num. of buildings
All	28.30	*-17.2*	1,114	159.97	*27.4*	2,513	119.53	*13.7*	3,627
Gabled	20.14	*-9.6*	954	131.26	*66.7*	1,374	85.73	*35.5*	2,328
Flat / other	76.94	*-62.3*	160	194.50	*-20.0*	1,139	180.11	*-25.2*	1,299

Table 4: Average rooftop area values and random guess estimator results for the rooftop area estimation

		Random guess	Biased random guess		
			gabled roof	flat roof	other roof types
Average rooftop area (m^2)		111.58	110.20	276.55	208.42
All buildings	MAE (m^2)	90.73	95.37		
	MBE (m^2)	*-43.94*	*0*		
Residential buildings with gabled rooftop	MAE (m^2)	49.91	48.94		
	MBE (m^2)	*36.79*	*35.41*		

4.1.3 Interpretation of the results

We interpret the performance of the rooftop area estimation for residential buildings with gabled roofs as good, considering the fact that the prediction is only based on 2D VGI data with a varying data quality [1]. Besides that, in a related study where OSM data was used together with satellite images, an error rate of 12-29% (amount of wrongly identified roof ridge lines) was achieved. In studies using 3D laser-scanning data, errors in the rooftop estimation of 15% are common [41, 61].

The roof area estimation for other buildings high errors and a large positive bias. We find two explanations for that: 1) entries in OSM are sometimes missing (some buildings are not mapped, so that the next building is considered by our implementation, and

many buildings have an unspecified type); 2) multiple houses are often mapped as one building (for example in the case of row houses, or semi-detached houses). One explanation for the underestimated roof area in the category of flat/other roofs lies in the used model (Equation 1) that is adapted to gabled roofs.

4.2 Validation of the solar potential estimate

4.2.1 Data description

We use real production data from existing PV installations in Germany for the second validation step. This data was recorded until 31.12.2013 for accounting the German Renewable Energies Act [66] subsidies and was made available online [10]. Besides the electric energy produced in 2013, the location, the year of construction, and the nominal installed capacity in *kW-peak* is known. We selected all PV installations on building roofs that have been built in 2012 (the 85,806 installations are depicted in Figure 3), which is the year before the data recording ended, since we assume that the newest installations represent the best technical state regarding to solar panel efficiency.

Some PV installations in the dataset have extreme large or low values that may distort our analysis. Therefore, we exclude about 4% of the data points as outliers that match the following criteria: all installations with a real production higher than the 99% quintile ($q_{0.99}$= 64,293 kWh), or lower than the 1% quintile ($q_{0.01}$= 1,690 kWh), such as buildings with a predicted production that is higher than the 99% quintile ($q_{0.99}$= 77,637 kWh), or lower than the 1% quintile ($q_{0.99}$= 623 kWh). Besides that, we exclude 14,092 installations that have no corresponding buildings mapped in the OSM database (as described in Section 3.1). In total, $n_E = 71{,}330$ installations are used for our validation.

4.2.2 Validation results

To validate the prediction of the photovoltaic potential, we compare the actual electrical energy production $\ddot{E}^{i}_{out}$ with the predicted solar potential E^{i}_{out}. For each building i, we compute the error $\delta^{E}_{i} = E^{i}_{out} - E^{\ddot{i}}_{out}$ and use $MAE = \frac{1}{n}\sum_{i=1}^{n}\left\|\delta^{E}_{i}\right\|$ and $MBE = \frac{1}{n}\sum_{i=1}^{n}\delta^{E}_{i}$ as performance metrics.

Table 5: Performance of the PV potential estimation model for different building types and installation sizes

	Residential buildings				Other buildings			
Installation size	MAE (kWh)	MAE (%)	*MBE (kWh)*	Num. of installations	MAE (kWh)	MAE (%)	*MBE (kWh)*	Num. of installations
All	4,805	55.94	*-1,820*	9,434	7,156	66.43	*-1,767*	61,896
Small	3,992	52.80	*-923*	9,110	5,783	67.01	*-253*	57,456
Large	27,643	73.76	*-27,026*	324	24,920	64.72	*-21,354*	4,440

The MAE of our PV potential estimation is 6,845 kWh (65.29% of the average annual PV production in the complete validation dataset). The model underestimates the PV potential on average by 17% (MBE -1774 kWh). The detailed results for *residential buildings* and *other buildings* (the type was obtained from OSM, as described in Section 3.1) are shown in Table 5, separated by the size of the PV installation (based on the nominal installed power in the validation data in *small installations* with ≤30 kWp [66], and *large installations*).

4.2.3 Interpretation of the results

The average estimation error for the PV potential on residential building roofs is 55.15%. The error for small PV installations (≤30 kWp) is even lower at 52.27%.

The comparison of our validation results to the state of the art is difficult, because authors of PV potential studies lack frequently to validate their estimates with real production data [41]. Only Jakubiec and Reinhart [29] assess the performance of their algorithm for laser-scanning and daily solar irradiation data with two real roofs and found that they achieve 9% root-mean-square deviation from the real PV production figures. They also compare two other state of the art algorithms with their estimation and found that these estimations deviate by 32-37% in a test setting with 10 roofs. The results achieved in the test can therefore be interpreted as satisfactory, considering the quality and granularity of the underlying data (2D crowdsourced VGI data and averaged monthly solar irradiation and temperature).

5 LIMITATIONS AND IMPLICATIONS FOR FURTHER RESEARCH

Although the proposed DSS derives a good preliminary estimation of PV potential for individual residential building roofs, the following three limitations must be mentioned: First, the prediction quality depends heavily on the quality of the OSM data, that fortunately increases steadily [1], but in some regions data entries are still sparse. This leads to the problem that the yield prediction be made in particular cases due to the lack of data. In spite of this fact, our approach can be seen as an example for the quality assessment for VGI data based on application needs, as claimed by Mondzech and Sester [42]. Second, the approach is profiled to gabled building roofs with a tilt angle of 35°. Since flat roofs are also common, the parameters could be adjusted to provide also more accurate estimations for other roof types. Third, we included 13 influence factors and made assumptions on them (normal distribution of the factors, fixed value for the solar panel efficiency), that might have different distributions in specific regions (deserts, polar regions, etc.) or might change in the future. Our implementation allows changing these values and adapting the method to other local conditions.

We believe, that our approach can be further extended with a more advanced querying of the building footprints from OSM. For that, a recently proposed algorithm by Hopf et al. [24] can be used, that selects OSM objects not only based on the distance, but also on semantics (objects tagged as *residential buildings* might be more applicable, even if they have a larger distance to the geo-located address, than objects tagged with *greenhouse* or *garage*). Besides that, shadowing objects that are mapped in OSM, like trees or large buildings, or the sparsely recorded information on roof heights and roof types can be incorporated to provide a more accurate prediction of the solar energy assessment. Finally, our implementation could be further extended to provide estimations on further roof types to reduce estimation errors. For that, an

empirical study of existing roof types and the ability to recognize them from 2D data would be necessary.

6 CONCLUSION

In this paper, we presented a novel data-mining–based DSS that uses freely available crowdsourced data in combination with open sensory data (solar irradiation and temperature observations) to automatically assess the PV potential of individual residential building roofs. The estimation result can be used on large scale by solar companies, energy service providers and electric utility companies, supporting their decisions what (potential) residential customers to select for renewable energy rollout campaigns.

In the validation with cadaster data we found, that our approach obtains the rooftop area for residential gabled roofs based on 2D data with an error of 20.14 m^2 (27% of the actual roof area). The validation of the PV potential estimation for residential buildings with real production data showed, that our method has an average error of 52%. For the initial assessment of the residential roof PV potential, without the use of costly and hard to obtain remote-sensing data (e.g., 3D laser-scanning data or aerial images) as used in previous works, the presented results are reasonable.

By going beyond the state of the art, this work makes both practical and theoretical contributions to the field of energy data analytics: Most importantly, we bridge the gap between the needs of solar energy companies to gain information about potential PV kit adopters and free broadly available VGI data. Moreover, this approach provides location-based PV yield estimations for the roofs at the individual level and is useable in manifold ways (for targeted marketing of solar energy providers, for personal decision support by household inhabitants, for policy preparation, etc.) Finally, this approach can be used for the quality assessment for VGI data based on the application needs [42]. All in all, this DSS is an obvious example of how crowdsourced and sensory data mining can contribute to the value generation for energy utilities, household residents and environmental sustainability.

Acknowledgements

The research presented in this paper was financially supported by Swiss Federal Office of Energy (Grant number SI/501202-01), and Eureka member countries and European Union (EUROSTARS Grant number E!9859 - BENgine II). We thank Denis Stühler for his contribution to implement an earlier version of the decision support system.

7 REFERENCES

[1] Ballatore, A. and Zipf, A. 2015. A Conceptual Quality Framework for Volunteered Geographic Information. *Spatial Information Theory*. S.I. Fabrikant et al., eds. Springer International Publishing. 89–107.

[2] Bavarian Land Suveying Office 2015. *Kundeninformation LoD2 (in German)*. Bayerisches Landesamt für Digitalisierung, Breitband und Vermessung.

[3] Branker, K. and Pearce, J.M. 2010. Financial return for government support of large-scale thin-film solar photovoltaic manufacturing in Canada. *Energy Policy*. 38, 8 (Aug. 2010), 4291–4303.

[4] Burger, B. et al. 2016. *Photovoltaics Report*. Frauenhofer ISE.

[5] Byrne, J. et al. 2015. A review of the solar city concept and methods to assess rooftop solar electric potential, with an illustrative application to the city of Seoul. *Renewable and Sustainable Energy Reviews*. 41, (Jan. 2015), 830–844.

[6] de Castro, C. et al. 2013. Global solar electric potential: A review of their technical and sustainable limits. *Renewable and Sustainable Energy Reviews*. 28, (Dec. 2013), 824–835.

[7] Catita, C. et al. 2014. Extending solar potential analysis in buildings to vertical facades. *Computers & Geosciences*. 66, (May 2014), 1–12.

[8] DESTATIS: Wohnungsbestand in Deutschland (in German): 2016. *https://www.destatis.de/*. Accessed: 2016-08-26.

[9] Elwood, S. et al. 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*. 102, 3 (Mai 2012), 571–590.

[10] EnergyMap - Auf dem Weg zu 100% EE - Der Datenbestand: 2015. *http://www.energymap.info/download.html*. Accessed: 2016-02-20.

[11] Enphase Energy Inc. 2014. *Guide to PVWatts Derate Factors for Enphase Systems When Using PV System Design Tools*. National Renewable Energy Laboratory (NREL), Golden, CO.

[12] Esclapés, J. et al. 2014. A method to evaluate the adaptability of photovoltaic energy on urban façades. *Solar Energy*. 105, (2014), 414–427.

[13] European Commission 2014. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions a Policy Framework for Climate and Energy in the Period from 2020 to 2030*.

[14] Eurostat 2015. *Consumption of energy - Statistics Explained*. Eurostat.

[15] Fishman, G. 2013. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer Science & Business Media.

[16] Freitas, S. et al. 2015. Modelling solar potential in the urban environment: State-of-the-art review. *Renewable and Sustainable Energy Reviews*. 41, (Jan. 2015), 915–931.

[17] Gagnon, P. et al. 2016. *Rooftop Solar Photovoltaic Technical Potential in the United States: A Detailed Assessment*. Technical Report #NREL/TP-6A20-65298. National Renewable Energy Laboratory.

[18] Gherboudj, I. and Ghedira, H. 2016. Assessment of solar energy potential over the United Arab Emirates using remote sensing and weather forecast data. *Renewable and Sustainable Energy Reviews*. 55, (Mar. 2016), 1210–1224.

[19] Goetz, M. and Zipf, A. 2012. Towards defining a framework for the automatic derivation of 3D CityGML models from volunteered geographic information. *International Journal of 3-D Information Modeling (IJ3DIM)*. 1, 2 (2012), 1–16.

[20] Green, M.A. et al. 2014. Solar cell efficiency tables (version 44): Solar cell efficiency tables. *Progress in Photovoltaics: Research and Applications*. 22, 7 (Jul. 2014), 701–710.

[21] Guo, D. and Mennis, J. 2009. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*. 33, 6 (Nov. 2009), 403–408.

[22] Hofierka, J. and Kaňuk, J. 2009. Assessment of photovoltaic potential in urban areas using open-source solar radiation tools. *Renewable Energy*. 34, 10 (Oct. 2009), 2206–2214.

[23] Hoogwijk, M.M. 2004. *On the Global and Regional Potential of Renewable Energy Sources*. University of Utrecht.

[24] Hopf, K. et al. 2015. Identifying the Geographical Scope of Prohibition Signs. *COSIT 2015* (Santa Fe, NM: USA, 2015), 247–267.

[25] IEA 2014. *Medium-Term Renewable Energy Market Report 2014: Market Analysis and Forecasts to 2020*.

[26] IEA 2015. *Medium-Term Renewable Energy Market Report 2015 - Market Analysis and Forecasts to 2020*.

[27] Jacques, D.A. et al. 2014. Methodology for the assessment of PV capacity over a city region using low-resolution LiDAR data and application to the City of Leeds (UK). *Applied Energy*. 124, (Jul. 2014), 28–34.

[28] Jäger-Waldau, A. 2014. *PV Status Report 2014*. Technical Report #Report EUR 26990 EN. European Commission.

[29] Jakubiec, J.A. and Reinhart, C.F. 2013. A method for predicting city-wide electricity gains from photovoltaic panels based on LiDAR and GIS data combined with hourly Daysim simulations. *Solar Energy*. 93, (Jul. 2013), 127–143.

[30] Jochem, A. et al. 2009. Automatic Roof Plane Detection and Analysis in Airborne Lidar Point Clouds for Solar Potential Assessment. *Sensors*. 9, 7 (Jul. 2009), 5241–5262.

[31] Korfiati, A. et al. 2016. Estimation of the Global Solar Energy Potential and Photovoltaic Cost with the use of Open Data. *International Journal of Sustainable Energy Planning and Management*. 9, (2016), 17–30.

[32] Lamigueiro, O.P. 2012. solaR: Solar Radiation and Photovoltaic Systems with R. *Journal of Statistical Software*. 50, 9 (2012).

[33] Levinson, R. et al. 2009. Solar access of residential rooftops in four California cities. *Solar Energy*. 83, 12 (2009), 2120–2135.

[34] Liao, S.-H. et al. 2012. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*. 39, 12 (Sep. 2012), 11303–11311.

[35] Lindroos, J. and Savin, H. 2016. Review of light-induced degradation in crystalline silicon solar cells. *Solar Energy Materials and Solar Cells*. 147, (Apr. 2016), 115–126.

[36] Lödl, M. et al. 2010. Abschätzung des Photovoltaik-Potentials auf Dachflächen in Deutschland. *11. Symposium Energieinnovation* (Graz, Austria, 12.02 2010).

[37] Ludwig, D. et al. 2009. Location analysis for solar panels by LiDAR-data with geoprocessing—SUN-AREA. *Proceedings of the 23rd International Conference on Informatics for Environmental Protection* (Berlin, Germany, 2009), 9–11.

[38] Mainzer, K. et al. 2016. Rooftop Pv Potential Estimations: Automated Orthographic Satellite Image Recognition Based on Publicly Available Data. *Proceedings of European PV Solar Energy Conference and Exhibition (EU PVSEC)* (Munich, Germany, 24.06 2016).

[39] Marion, B. et al. 2005. Performance parameters for grid-connected PV systems. *Conference Record of the Thirty-first IEEE Photovoltaic Specialists Conference, 2005.* (2005), 1601–1606.

[40] Mejia, F.A. and Kleissl, J. 2013. Soiling losses for solar photovoltaic systems in California. *Solar Energy*. 95, (Sep. 2013), 357–363.

[41] Melius, J. et al. 2013. *Estimating Rooftop Suitability for Pv: A Review of Methods, Patents, and Validation Techniques*. Technical Report #NREL/TP-6A20-60593. National Renewable Energy Laboratory (NREL), Golden, CO.

[42] Mondzech, J. and Sester, M. 2011. Quality Analysis of OpenStreetMap Data Based on Application Needs. *Cartographica: The International Journal for Geographic Information and Geovisualization*. 46, 2 (2011), 115–125.

[43] Montavon, M. et al. 2004. Solar energy utilisation potential of three different swiss urban sites. *Energie und Umweltforschung im Bauwesen, Zurich*. (2004), 503–510.

[44] Nguyen, H.T. et al. 2012. The Application of LiDAR to Assessment of Rooftop Solar Photovoltaic Deployment Potential in a Municipal District Unit. *Sensors*. 12, 12 (Apr. 2012), 4534–4558.

[45] Olson, D.L. and Wu, D. 2005. Decision Making with Uncertainty and Data Mining. *Advanced Data Mining and Applications* (Jul. 2005), 1–9.

[46] OpenStreetMap Statistics: 2016. *http://www.openstreetmap.org/stats/data_stats.html*. Accessed: 2016-07-22.

[47] Ordóñez, J. et al. 2010. Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain). *Renewable and Sustainable Energy Reviews*. 14, 7 (Sep. 2010), 2122–2130.

[48] Redweik, P. et al. 2013. Solar energy potential on roofs and facades in an urban landscape. *Solar Energy*. 97, (Nov. 2013), 332–341.

[49] Regvat, R. et al. 2014. 3D-punktbasierte Solarpotenzialanalyse für Gebäudefassaden mit freien Geodaten. *AGIT 2014 – Symposium und Fachmesse Angewandte Geoinformatik* (Salzburg, Austria, 2014).

[50] Satellite Application Facility on Climate Monitoring: 2016. *http://www.eumetsat.int/website/home/Satellites/GroundSegment/Safs/ClimateMonitoring/index.html*. Accessed: 2016-06-02.

[51] Scartezzini, J.-L. et al. 2002. Computer evaluation of the solar energy potential in an urban environment. *EuroSun, Bologna*. (2002).

[52] Schallenberg-Rodríguez, J. 2013. Photovoltaic techno-economical potential on roofs in regions and islands: The case of the Canary Islands. Methodological review and methodology proposal. *Renewable and Sustainable Energy Reviews*. 20, (Apr. 2013), 219–239.

[53] Schmela, M. et al. 2016. *Global Market Outlook for Solar Power 2016 - 2020*. Technical Report #ISBN 9789090298146. SolarPower Europe.

[54] Steadman, P. 2006. Why are most buildings rectangular? *Architectural Research Quarterly*. 10, 2 (2006), 119–130.

[55] Stigka, E.K. et al. 2014. Social acceptance of renewable energy sources: A review of contingent valuation applications. *Renewable and Sustainable Energy Reviews*. 32, (Apr. 2014), 100–106.

[56] Topi, M. et al. 2007. Effective efficiency of PV modules under field conditions. *Progress in Photovoltaics: Research and Applications*. 15, 1 (Jan. 2007), 19–26.

[57] U.S. EIA 2015. *Annual Energy Outlook 2015 with projections to 2040*. U.S. Energy Information Administration.

[58] Veronesi, F. et al. 2015. Evaluating the Use of Open Data to Estimate the Global Solar Energy Potential. *Proceedings of the 2nd International Conference on Energy and Environment* (Portugal, 19.06 2015).

[59] Voegtle, T. et al. 2005. Airborne laserscanning data for determination of suitable areas for photovoltaics. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 36, 3/W19 (2005), 215–220.

[60] Wieland, M. et al. 2015. Computing solar radiation on CityGML building data. *18th AGILE international conference on geographic informaton science* (2015).

[61] Wiginton, L.K. et al. 2010. Quantifying rooftop solar photovoltaic potential for regional renewable energy policy. *Computers, Environment and Urban Systems*. 34, 4 (Jul. 2010), 345–357.

[62] Willmott, C.J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. 30, 1 (Dec. 2005), 79–82.

[63] Wirth, H. and Schneider, K. 2012. Aktuelle Fakten zur Photovoltaik in Deutschland. *Fraunhofer ISE*. (2012).

[64] Wittmann, H. et al. 1997. Identification of roof areas suited for solar energy conversion systems. *Renewable Energy*. 11, 1 (1997), 25–36.

[65] Xintong, G. et al. 2014. Brief survey of crowdsourcing for data mining. *Expert Systems with Applications*. 41, 17 (Dezember 2014), 7987–7994.

[66] 2014. *German Renewable Energy Sources Act (Erneuerbare-Energien-Gesetz, EEG)*.