Supplementary Information for
Revealing drivers of green technology adoption through explainable AI

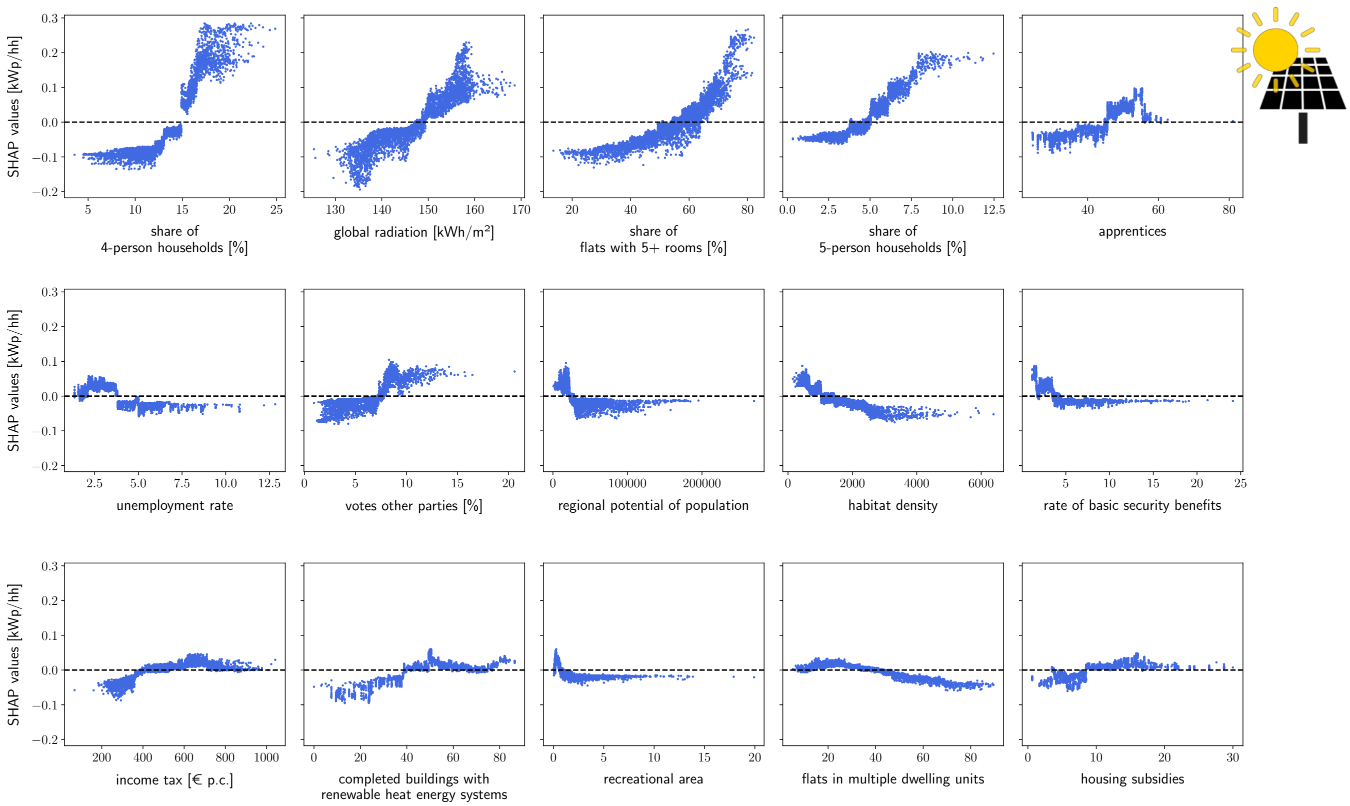
Dorothea Kistinger [1,2,3], Maurizio Titz [4,5], Philipp C. Böttcher [4], Michael T. Schaub [1], Sandra Venghaus [6,7] and Dirk Witthaut [4,5,*]

[1] *Department of Computer Science, RWTH Aachen University, 52074 Aachen, Germany*
[2] *Mercator Research Institute on Global Commons and Climate Change gGmbH, EUREF Campus 19, Torgauer Straße 12-15, 10829 Berlin, Germany*
[3] *Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, 14412 Potsdam, Germany*
[4] *Institute of Climate and Energy Systems: Energy Systems Engineering (ICE-1), Forschungszentrum Jülich, 52428 Jülich, Germany*
[5] *Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany*
[6] *School of Business and Economics, RWTH Aachen University, 52072 Aachen, Germany*
[7] *Institute of Climate and Energy Systems: Jülich Systems Analysis (ICE-2), Forschungszentrum Jülich, 52428 Jülich, Germany*

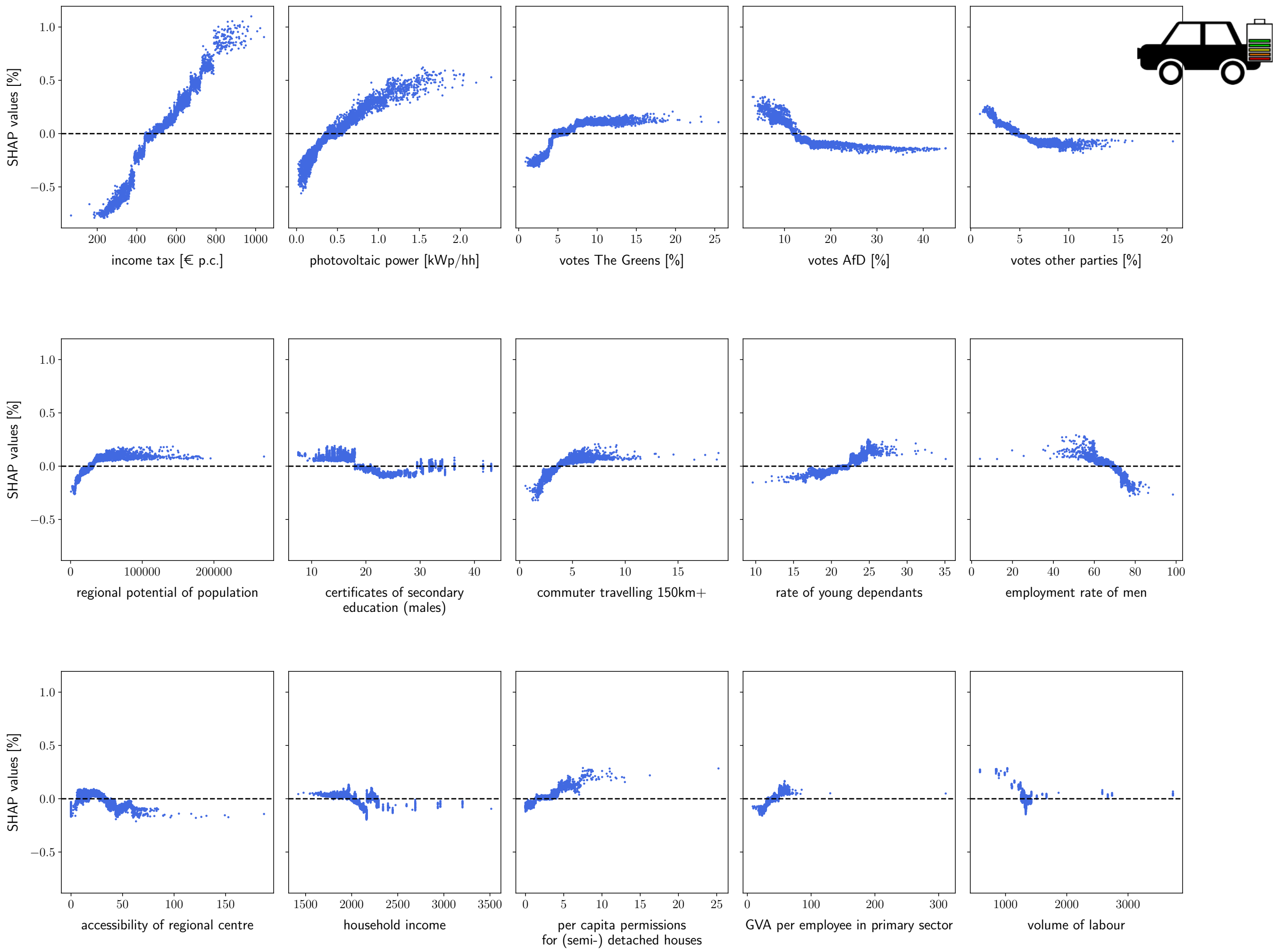
(Dated: October 31, 2024)

This Supplementary Material contains 6 Supplementary Figures and 9 Supplementary Notes.

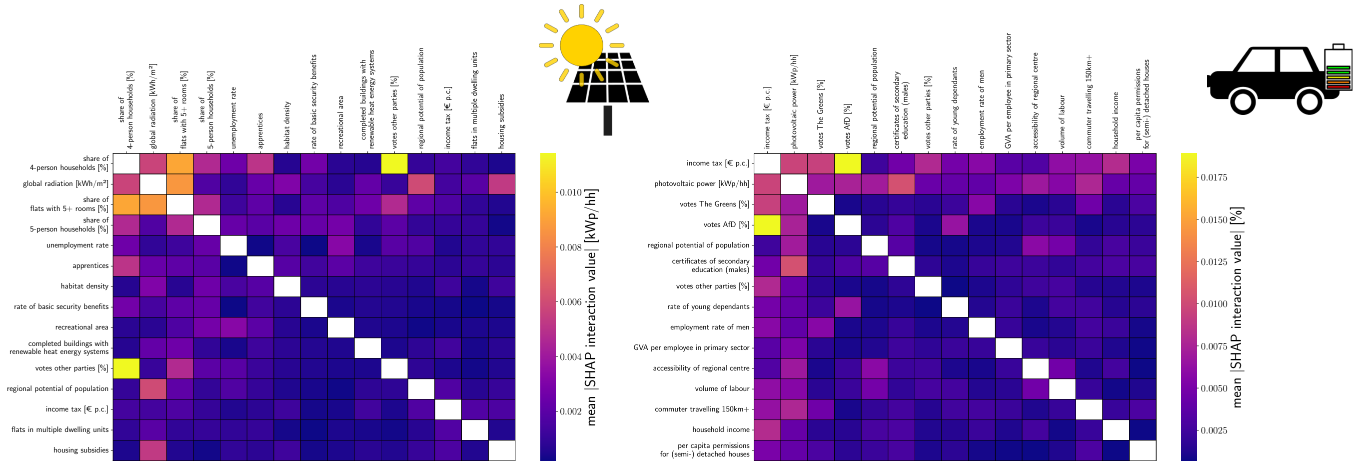
SUPPLEMENTARY FIGURES



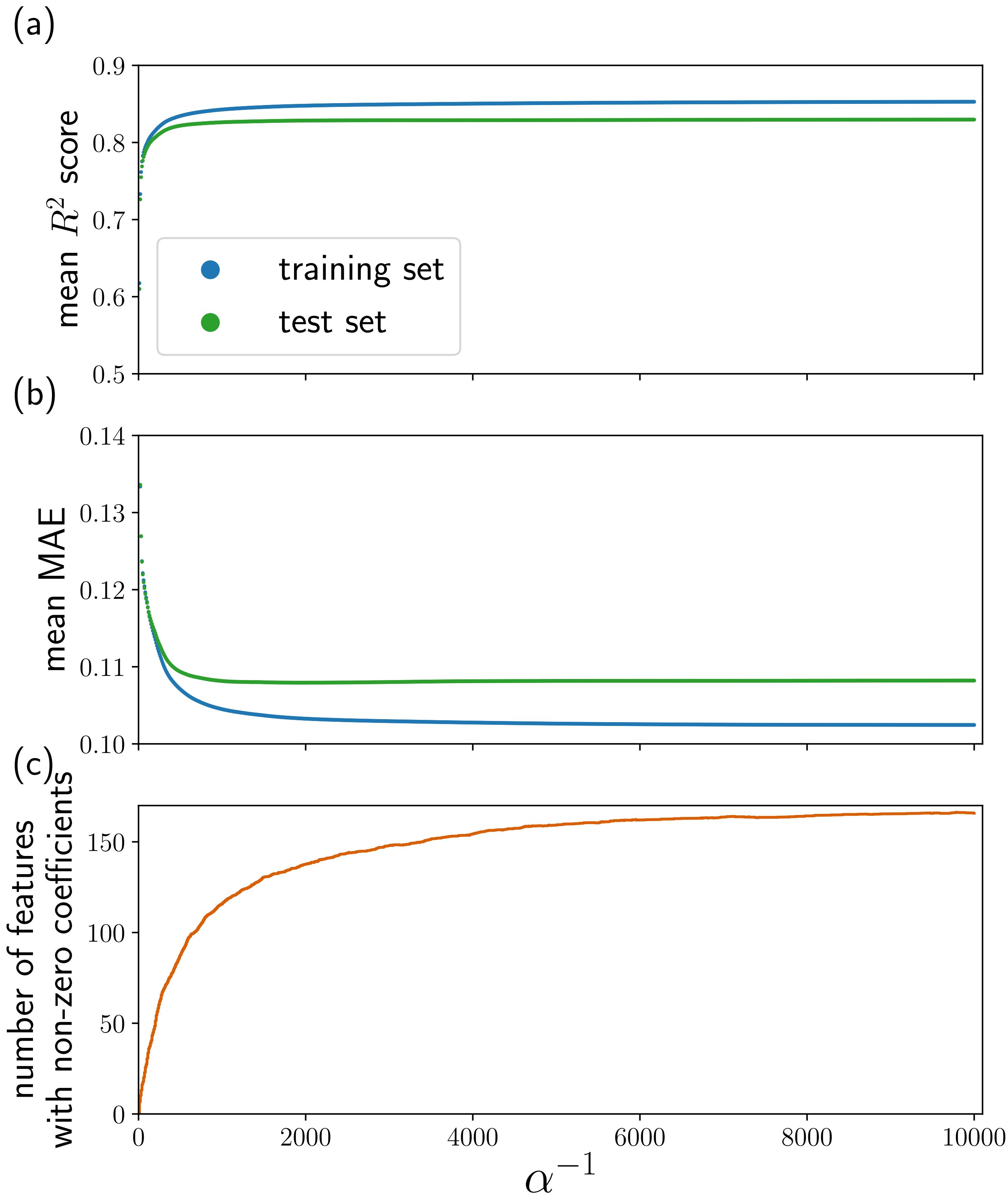
Supplementary Figure 1: **SHAP dependence plots for the adoption of household-scale PV.** The plot shows SHAP values for the GBT model predicting the stock of household-scale PV. For each data point x_n, the SHAP value $\phi_i(x_n; f)$ of feature i is plotted against the respective feature value $x_n^{(i)}$. The plots show the 15 features of the reduced best-performing model.



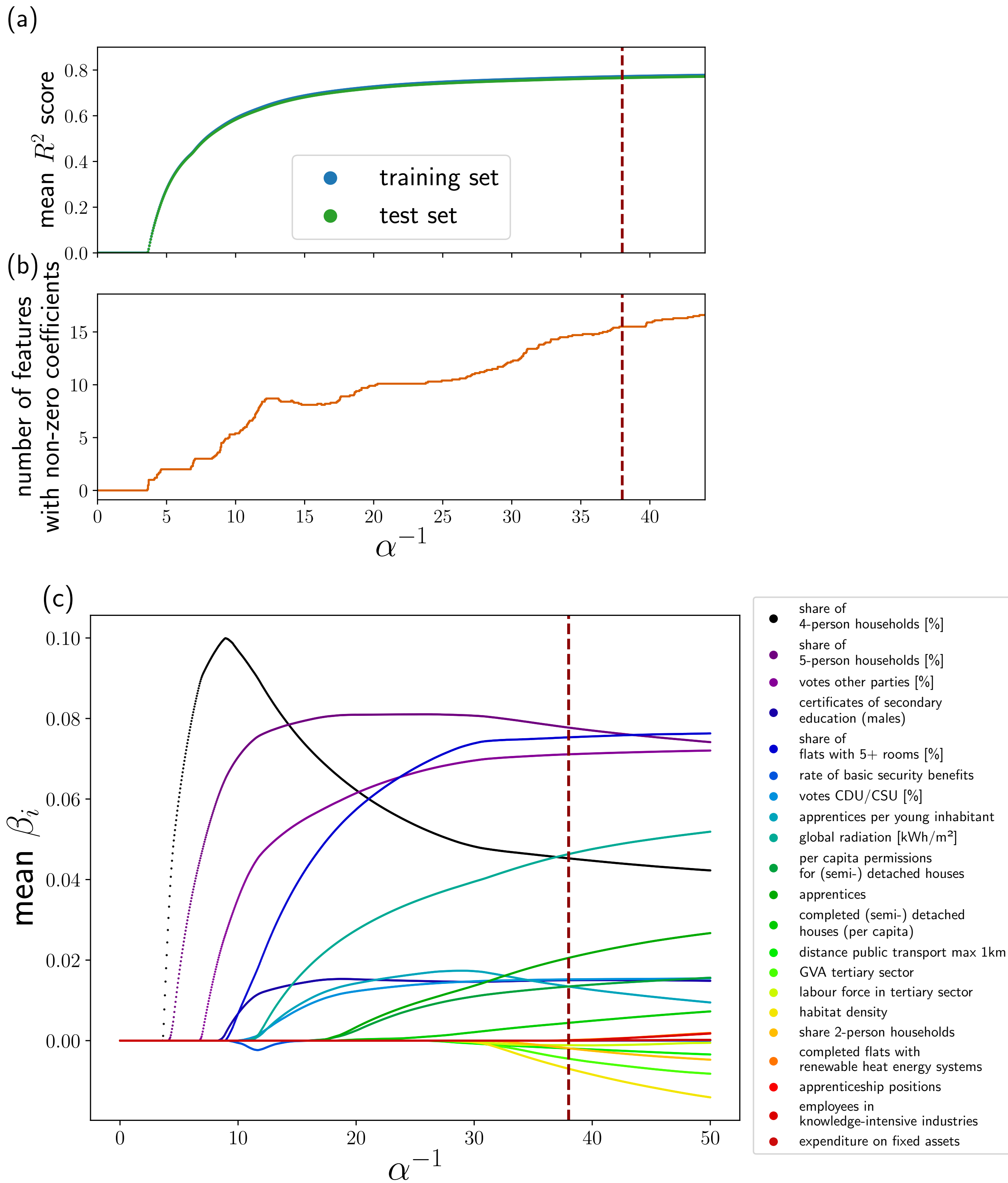
Supplementary Figure 2: **SHAP dependence plots for the adoption of battery-electric vehicles** The plot shows SHAP values for the GBT model predicting the adoption rate of battery-electric vehicles. For each data point x_n, the SHAP value $\phi_i(x_n; f)$ of feature i is plotted against the respective feature value $x_n^{(i)}$. The plots show the 15 features of the reduced best-performing model.



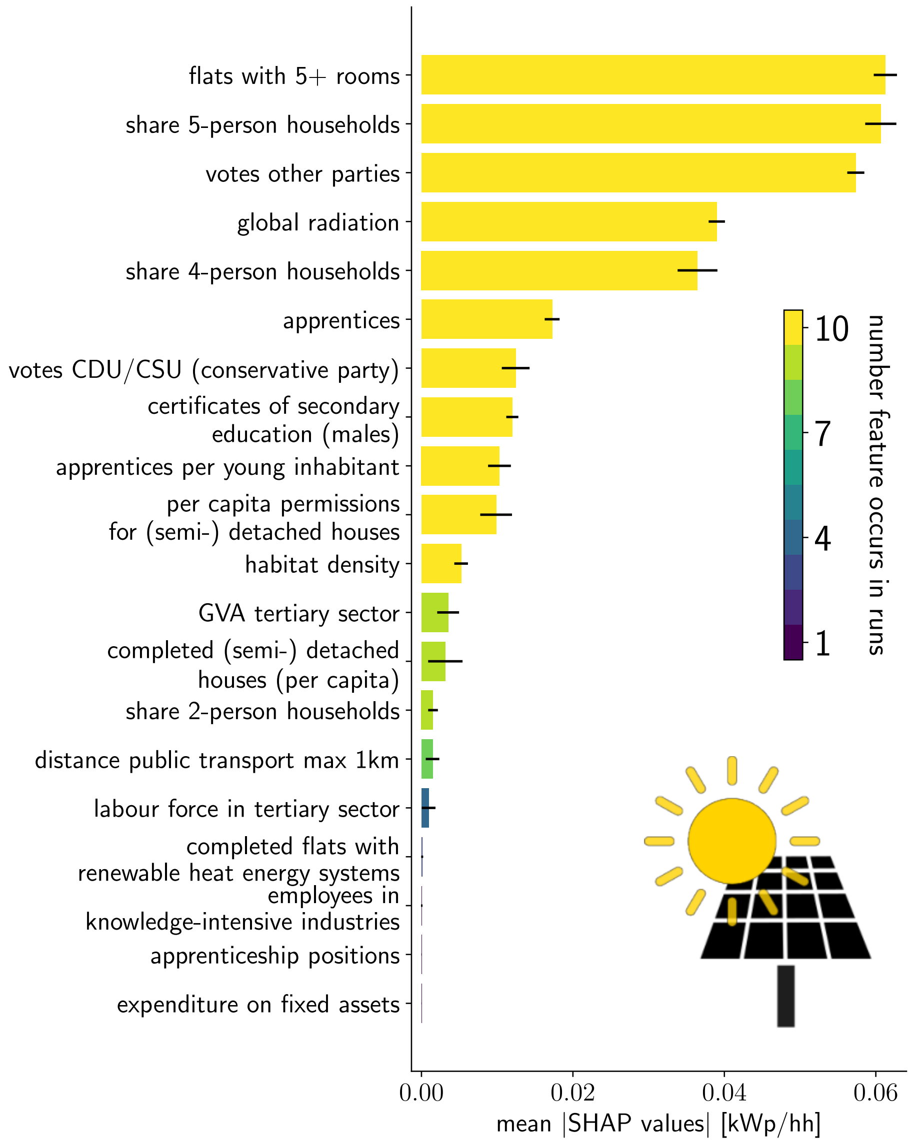
Supplementary Figure 3: **Interactions of input features learned by the GBT models.** The strength of feature interactions is quantified by the magnitude of the SHAP interaction values $|\Phi_{ij}|$ averaged over all data points. In the PV adoption model, the strongest interactions are observed between the share of 4-person households and (i) the share of votes for other parties and (ii) the share of flats with 5+ rooms and the share of flats with 5+ rooms and global radiation. In the BEV adoption model, the strongest interactions are observed between the income tax and the share of votes for the far-right AfD.



Supplementary Figure 4: **Performance of the LASSO model for large values of α^{-1} analyzing the adoption of household-scale PV.** We train and evaluate a LASSO model for benchmarking purposes. The analyses displayed in this figure are based on small values of the regularization parameters α. Panel (a): The mean R^2 score. Panel (b): Mean Absolute Error (MAE). Panel (c): Number of features with non-zero coefficients. The plots display the mean values over ten test-train splits.



Supplementary Figure 5: **Performance of the LASSO model for small values of α^{-1} analyzing the adoption of household-scale PV.** We train and evaluate a LASSO model for benchmarking purposes. The analyses displayed in this figure are based on large values of the regularization parameters α. Panel (a): The mean R^2 score. Panel (b): Number of features with non-zero coefficients. Panel (c): The coefficients of the LASSO model. The plots display the mean values over ten test-train splits. The red dashed lines indicate the choice of the reduced lasso model ($\alpha = 0.026324$, i.e., $\alpha^{-1} = \frac{1}{0.026324} \approx 38$).



Supplementary Figure 6: **SHAP feature importances of the LASSO model analyzing household-scale PV.** The figure shows the mean and the standard deviation of the SHAP feature importances over the ten test-train splits.

SUPPLEMENTARY NOTES

Supplementary Note 1: Machine Learning Pipeline and Lasso Results

Implementation of the GBT Models

The GBT models were implemented using LightGBM [1] according to the following procedure for the analysis of the BEV and PV adoption separately. Prior to the analysis, we split the dataset into three parts yielding a training, a validation and a test set. The training set comprised 64%, the validation set 16% and the test set 20% of the data. The actual analysis was two-part: Recursive feature elimination (RFE) was used to iteratively eliminate the least important features and construct a reduced GBT model which embraces the 15 most influential input features. In each iteration hyperparameter optimisation (HPO) yielded a best performing model which served as the foundation for the following iterations of the RFE. Once the best-performing reduced GBT model was identified, we analysed this model using SHAP and SHAP interaction values. The detailed workflow is as follows.

1. Divide data set into training, validation and test set.
2. RFE: For each $i \in \{N, N-l, \ldots, 1\}$ do:
   a HPO: Use random search to find best performing model containing i input features. For this purpose, for each set of hyperparameters (HPs):
      - Perform 5-fold cross-validation (CV) on the training set to determine the model performance for this set of HPs. We perform early stopping on the validation set to determine the required number of boosted trees of the GBT to speed up the training process.
      - Take the mean over the R^2 scores of the folds of the training set used for testing during CV to determine the model performance.
      - Retrain the model on the entire training set applying early stopping on the validation set to determine the ultimate number of trees necessary for this set of HPs.

   b Determine the best performing i-feature model, i.e., the model demonstrating the highest mean R^2 score on the test folds during CV.

   c Compute the SHAP feature importances of the best performing i-feature model on the validation set.

   d Detect the l features with the smallest SHAP feature importances and drop them.
3. Choose the best performing reduced model based on the results of the RFE, i.e., choose the model comprising 15 input features with the highest mean R^2 score on the test folds during CV as the reduced model.
4. Train the reduced model on the joint training and validation set.
5. Determine the performance of the reduced model on the test set.
6. Compute the SHAP and SHAP interaction values of the reduced model on the entire data set.

Following a proposition by Guyon et al. [2], we removed multiple features during early iterations of the RFE to speed up the process. We used the following eliminations scheme for the analysis of the PV adoption: $[10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 5, 5, 5, 5, 5, 5, 5, 5, 5, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$. That is, we first remove 10 features during the first 11 iterations, then 5 and so on. In each iteration of the RFE, we compared models based on 50 random sets of hyperparameters for hyperparameter optimization (HPO). In the case of BEV adoption, we follow the same elimination scheme but for one additional step of removing one feature at the end. This is necessary since here, the distribution of PVs and availability of charging stations are added, while the global radiation is removed from the data set. Thus, this data set contains one additional feature.

We executed the steps 1 to 3 for ten different splits of training, validation and test sets to verify the robustness of the identified reduced model. The reduced model revealing the largest mean R^2 score on the test folds during CV was chosen for the remaining analyses (steps 4 to 6).

During HPO, we used the following settings and intervals:

- loss function: l_2

- number of boosted trees (M): 500
- early stopping rounds (M_{boost}): 30
- maximum number of leaves (J): $2, 3, 4, \ldots, 10$
- minimum sample size in a leaf: $40, 41, 42, 43, \ldots, 200$
- learning rate (λ): $[0.001, 0.1]$
- subsample[3]: $[0.5, 1]$
- subsample frequency: $0, 1, 2, 3, 4, 5$

Implementation of the LASSO

For benchmarking purposes, we trained a LASSO regression model [4] on the data set focusing on household-scale PV. We used the *sklearn.preprocessing.StandardScaler* [5] published in the scikit-learn framework [6] to standardize the data prior to the analysis by the LASSO model.

For the PV data, the training and validation set from the GBT analysis served jointly as the training set LASSO model. We trained LASSO models for values of the regularization parameter α in the interval $[0.0001, 1000]$. Initializing the lasso models, we set the tolerance for optimization to 0.001 and the maximum number of iterations to 20000. Similar to the workflow of the GBT model, we further analyze a LASSO model for a specified value of α. We use the implementation by the *sklearn.model_selection.lasso* package [7] published in the scikit-learn framework [6].

For linear models such as the LASSO, SHAP values are directly given by their transformed regression coefficients.

Discussion of the LASSO analyzing household-scale PV

The regularization parameter α determines the number of input features included in a LASSO model and therefore also the composition of the model. Large values of α implement a large penalty for included input features, resulting in a small number of features with non-zero coefficients. Figures 4 and 5 show the performance of the LASSO models and the number of features as a function of the inverse parameter α^{-1}. The mean R^2 score continuously increases with rising α^{-1} on both the training and the test set, which suggests that overfitting is not an important issue for the LASSO model. A striking result is the large number of features included it takes for the LASSO to achieve high performance.

To compare the performance and feature importances of a LASSO to the results of the reduced GBT model emerging from the Recursive Feature Elimination, we determined a reduced LASSO model. We chose a regularization parameter of $\alpha = 0.026324$ because the resulting reduced models include a similar number of features as the reduced GBT models (15.5 on average for the lasso vs. 15 for the GBT). To ensure robustness, we created a reduced LASSO model for all ten train-test-splits of the data set.

Figure 6 shows the average feature importances of the reduced LASSO models of all test-train splits. A striking results is the similar composition of all ten models: The 11 features with the largest mean values of feature importance are contained in every model. Furthermore, the small values of the standard deviation indicate similar values of SHAP feature importance. We thus conclude that the ensemble of features included in the reduced LASSO model and their impact are robust to the choice of training and test sets.

Supplementary Note 2: Spatial Resolution and Data Aggregation

Spatial Resolution

Our study provides a comprehensive analysis of the adoption of household-scale PV systems on the level of municipal associations (‘Gemeindeverband’). Following a recommendation by the Federal Office for Building and Regional Planning (Bundesamt für Bauwesen und Raumordnung), we decided for the level of municipal associations due to their homogeneity with respect to their legal status and average size compared to municipalities [8]. Table I lists the total number of municipalities and municipal associations for every federal state in Germany.

German administration implemented a taxonomy for the unique identification of municipalities and municipal associations. Every municipal association is identified by the reference key RS (‘Regionalschlüssel’), while every municipality is identified by a reference key RS as well as AGS (‘Allgemeiner Gemeindeschlüssel’). Table II summarizes the definition of the AGS and RS.

For many features, data is not available on the level of municipal associations but only on the level of counties (lower resolution) or municipalities (higher resolution). Identification based on the AGS and RS enabled merging data of these different levels of aggregation. To include data provided on county-level, the RS allowed a simple mapping of the values from counties to the respective municipal associations. The spatial aggregation of data on the municipality level is described below.

Spatial Aggregation and Change of regional keys

For the aggregation of data on the municipality level, we derived a map assigning municipalities to the associated municipal associations based on the data provided by the INKAR database. Municipalities are identified by AGS, municipal associations by RS. Notably, data published by the INKAR database is based on the definition of municipalities and municipal associations as of 31 December 2019 [9, file 1].

The mapping must reflect the fact that the definition of municipalities is not entirely static, but some changes occur over time. The Federal Statistical Office publishes these changes annually [11]. To incorporate data from years prior to 2019, it is necessary for the mapping from municipalities to municipal associations provided by INKAR to account for the modifications in AGS that occurred between the time of the data’s publication and the end of 2019. More specifically, we added pairs of all outdated AGS and the (up-to-date) RS of the corresponding municipal associations. The algorithmic implementation comprised two steps. First, we constructed a mapping of all changes

Table I: **Number of municipalities and municipal associations in federal states.** We follow the definition of municipalities and municipal associations proposed by INKAR [9, file 3]. The table comprises all municipalities and municipal associations contained in INKAR, including the municipal associations that we dropped during the preprocessing.

Federal State	Number of municipalities	Number of municipal associations
Schleswig-Holstein	1108	170
Hamburg	1	1
Lower Saxony	968	431
Bremen	2	2
North Rhine-Westphalia	396	396
Hesse	427	427
Rhineland-Palatinate	2302	178
Baden-Württemberg	1103	462
Bavaria	2233	1562
Saarland	52	52
Berlin	1	1
Brandenburg	417	198
Mecklenburg-Vorpommern	726	116
Saxony	419	309
Saxony-Anhalt	218	122
Thuringia	634	191
Sum	11007	4618

Table II: Composition of AGS (8 digits) [10] and RS (12 digits) [11]. The table gives the numbers of digits encoding the administrative levels.

	federal state	government region	county	municipal association	municipality
AGS	2	1	2		3
RS	2	1	2	4	3

in AGS, i.e., a table listing all changes $\text{AGS}_{\text{old}} \mapsto \text{AGS}_{\text{new}}$. Starting with the year of publication of the data, we iteratively added the changes of AGS from the following years. The implementation considered transitivity in changes of AGS for different years, i.e., if $\text{AGS}_1 \mapsto \text{AGS}_2$ in time period t, it checks whether $\text{AGS}_2 \mapsto \text{AGS}_3$ for all following time periods $t + n$, $n \geq 1$. If there is a change $\text{AGS}_2 \mapsto \text{AGS}_3$, $\text{AGS}_1 \neq \text{AGS}_3$, then we altered the change from $\text{AGS}_1 \mapsto \text{AGS}_2$ to $\text{AGS}_1 \mapsto \text{AGS}_3$. In addition, we applied transitivity to changes of AGS within a time period. That is, provided $\text{AGS}_1 \mapsto \text{AGS}_2$ and $\text{AGS}_2 \mapsto \text{AGS}_3$, where $\text{AGS}_1 \neq \text{AGS}_3$, in the same year, we added $\text{AGS}_1 \mapsto \text{AGS}_3$ (and $\text{AGS}_2 \mapsto \text{AGS}_3$) instead of $\text{AGS}_1 \mapsto \text{AGS}_2$. Two changes in AGS, i.e., $\text{AGS}_1 \mapsto \text{AGS}_2$ and $\text{AGS}_1 \mapsto \text{AGS}_3$, were recorded for a few municipalities. In this case, we considered the change listed first and omitted the following modification. The last changes considered were the changes in AGS from 2019. Overall, this procedure yielded a mapping, i.e., a table, of changes in AGS between the year publication of the data and 2019.

Based on these changes, we extended the mapping of municipalities to municipal associations. To ensure uniqueness, we only considered the AGS changing over time, i.e., contained in the list of changes as a value of the independent variable, which was not covered by the mapping from municipalities to municipal associations provided by INKAR. That is, we considered all AGS_{old} such that $\text{AGS}_{\text{old}} \mapsto \text{AGS}_{\text{new}}$ is contained in the list of changes, but there is no RS^* of a municipal association such that $\text{AGS}_{\text{old}} \mapsto \text{RS}^*$ defined by the mapping derived from INKAR. For all those AGS_{old}, we extended the INKAR-mapping by $\text{AGS}_{\text{old}} \mapsto \text{RS}^*$ if there exists a value RS^* such that $\text{AGS}_{\text{new}} \mapsto \text{RS}^*$.

This procedure yielded an extended mapping of municipalities to municipal associations which enabled integration of data on the municipal level from an earlier point in time into the final data set. We applied this approach three times to derive extended mappings for the integration of the data

- gathered from the 2011 Census (considering changes between 2011 and 2019),
- on the election for the Bundestag in 2017 published by the Regional Database (considering changes between 2017 and 2019) and
- on photovoltaic systems registered at the Marktstammdatenregister platform (considering changes between 2000 and 2019 as most PV systems were installed after 2000).

We took care manually of regional changes on the level of municipalities and municipal associations for the data on electric vehicles and charging stations as data was collected prior and after the publication of the INKAR data set in 2019 (see Supplementary Notes 8 and 9).

Supplementary Note 3: The INKAR Data Set

Overview

The German Federal Institute for Research on Building, Urban Affairs and Spatial Development (Bundesinstitut für Bau-, Stadt- und Raumforschung) provides the INKAR database which embraces roughly 600 indicators from various domains capturing living conditions in Germany and Europe on different levels of spatial aggregation [9, 12]. The lowest level of spatial aggregation INKAR provides most variables on is either the level of municipal associations or counties. The features supplied at the county level are downscaled to the respective municipal associations.

Most indicators were most recently recorded in 2019, which, thus, constitutes the base year for the input data. Whenever available, we included data from 2019. In all other cases, values from 2017, 2018 or 2020 were used. Table III lists all variables from the INKAR database used as well as their year of record and the level of spatial aggregation they are provided on. As almost all data is given either as a share or per capita, we normalize the remaining data that give an absolute number (i.e., 'employees at place of work', 'employees at place of residence', 'unemployed' and 'population of working age') by dividing them by the population of their respective municipal association. In total, 180 variables were used from the INKAR database. Out of these variables, 88 were provided on the level of municipal associations, 92 on county-level.

Data source

The entire INKAR database can be downloaded at [9]. We utilized data from the *inkar_2021*-version [9] which contains four files:

1. *Readme.txt* (file 1) gives metadata on the dataset.
2. *Übersicht der Indikatoren.xlsx* (file 2) contains a detailed description of all features included in the INKAR database. We refer to this file for further details of the definitions of the variables we used from INKAR. We only utilize features listed in the table *Raumbeobachtung DE.*
3. *Referenz Gemeinden, Kreise, NUTS.xlsx* (file 3) provides four tables for identification of municipalities, municipal associations and counties. Two are of major importance for this work:
    - *Gemeinden-Gemeindeverbände* maps the AGSs used to identify municipalities to the municipal associations they belong to. INKAR identifies municipal associations using an identification number that is similar but not equal to the RS.
    - The table *Regionalschlüssel GVB* provides a mapping of the identification numbers of municipal associations used by the INKAR database to the official RS keys.
4. *inkar_2021.csv* (file 4) comprises the actual data of the database.

This version of the INKAR database refers to the administrative definition of municipalities, municipal associations, counties and federal states as of 31.12.2019 [9, file 1]. The file *inkar_2021.csv* provides the data we analyze. Additionally, we deduced a mapping of municipalities identified by their AGS to the corresponding municipal associations encoded by the RS from the two tables mentioned above which are contained in file 3. This mapping constituted the foundation of the mappings used to include data on the municipal level provided by other sources.

Preprocessing Data on the Level of Municipal Associations

As table III reveals, INKAR provides 88 of the relevant variables on the level of municipal associations. Out of all 4618 municipal associations, we dropped 208 due to zero population. In addition, the municipal associations gemfr. Bezirk Lohheide (RS 33519501) and Zandt (RS 93720177) were excluded from the analysis due to missing values for some of the variables related to employment.

The variables 'mean distance public transport' and 'distance public transport max 1km' were missing for seven municipal associations which proved to be islands in the North Sea: Föhr-Amrum (RS 10545488), Helgoland (RS 10560025), Baltrum (RS 34520002), Inselgemeinde Juist (RS 34520013), Nordseebad Wanderooge (RS 34550021), Langeoog (RS 34620007) and Spiekeroog (RS 34620014). Based on the assumption that public transport is not offered on these islands, we set

- 'mean distance public transport' to infinity (for the GBT model and to the maximum non-missing value observed for the LASSO model as the LASSO requires finite values) and
- 'distance public transport max 1km' to zero.

The INKAR dataset comprised 'planning permissions for (semi-) detached houses' ($\text{permissions}_{\text{detached}}$) and 'planning permissions for flats (multiple dwelling units)' ($\text{permissions}_{\text{multiple}}$) indicating the share of permissions for new flats in (semi-) detached houses or multiple dwelling units on all planning permissions for new flats in percent (see file 2 of [9] for further information). Whenever there are no planning permissions for flats granted, i.e., 'planning permissions for flats' per 1000 inhabitants (p.c. permissions) is equal to zero, 'planning permissions for (semi-) detached houses' and 'planning permissions for flats (multiple dwelling units)' are ill-defined and missing. Thus, we substituted these two variables with two per-capita variables:

$$\text{p.c. permissions}_{\text{detached}} = \begin{cases} 0 & \text{if p.c. permissions} = 0 \\ \frac{\text{permissions}_{\text{detached}} \cdot \text{p.c. permissions}}{100} & \text{else} \end{cases}$$

and

$$\text{p.c. permissions}_{\text{multiple}} = \begin{cases} 0 & \text{if p.c. permissions} = 0 \\ \frac{\text{permissions}_{\text{multiple}} \cdot \text{p.c. permissions}}{100} & \text{else.} \end{cases}$$

The input data set comprises 'per capita permissions for (semi-) detached houses' (p.c. $\text{permissions}_{\text{detached}}$) instead of 'planning permissions for (semi-) detached houses' ($\text{permissions}_{\text{detached}}$) and 'per capita permissions for flats (multiple dwelling units)' (p.c. $\text{permissions}_{\text{multiple}}$) instead of 'planning permissions for flats (multiple dwelling units)' $\text{permissions}_{\text{multiple}}$.

Taking these modifications into account, 20 missing values remained in the data set: The 'fiscal capacity' was missing for six and 'studio and two-room apartments' for 14 municipal associations. We used the most recent values available from the past as substitutes.

Preprocessing Data on County-Level

14 of the 92 features provided by INKAR on the county level have missing values.

For the cities of Rostock (RS 13003) and Schwerin (RS 13004), INKAR does not provide values for 'GVA primary sector' and 'GVA per employee in primary sector'. Percentages indicating the contribution of the secondary ('GVA secondary sector') and the tertiary sector ('GVA tertiary sector') allowed for a derivation of the contribution of the primary sector, i.e.,

$$\text{GVA primary sector} = 100 - \text{GVA secondary sector} - \text{GVA tertiary sector}$$

for missing values of 'GVA primary sector'.

The variable 'GVA per employee in primary sector' indicates the GVA generated in the primary sector normalized by the number of employees working in that sector. Similarly, 'GVA per employee' gives the total GVA divided by the number of employees of all sectors. To approximate the missing values of the 'GVA per employee in primary sector' (GVA_{prim}) for the cities of Schwerin and Rostock, we computed the state-wide average ratio of the two indicators and multiplied it by the value of 'GVA per employee' (GVA) of the municipal association, i.e.,

$$\text{GVA}_{\text{prim}}(c) = \text{GVA}(c) \cdot \frac{1}{|S_c|} \sum_{\text{county} \in S_c} \frac{\text{GVA}_{\text{prim}}(\text{county})}{\text{GVA}(\text{county})} \text{ for c} \in \{\text{Rostock,Schwerin}\}$$

where $S_c := \{\text{county}|\text{county in state of c \& county} \neq c \text{ \& } \text{GVA}_{\text{prim}}(\text{county}) \neq \text{NaN}\}$.

For the remaining 12 variables with missing values ('price of building ground', 'employees in industry', 'employees in non-productive industries', 'employees in knowledge-intensive industries', 'apprenticeship positions', 'German university entrance qualification (Abitur)', 'rate support grant', 'expenditure on fixed assets ', 'assignments of investment', 'business volume in mining and manufacturing industry', 'business volume in main construction industry', 'business volume of skilled crafts and trades'), we included the most recent values from previous years into the data set. Table **??** lists the values chosen. INKAR provided substitutes from the past for all missing values except of the variable 'business volume of skilled crafts and trades' for Dessau-Roßlau (RS 15001) and Jerichower Land (RS 15086). For these two counties, we used the state-wide values of the variables.

Variables from INKAR Used

In the following Table III, we provide a complete list of variables from the INKAR database [9] used in this work. The German name specified corresponds to the exact German spelling as provided by INKAR, even if it is grammatically wrong. Further documentation on the data is given by file 2 of the INKAR download. The variables can be identified in file 2 of the INKAR download by the German name and the identification number which is given in the column "ID in INKAR". "Time period" and "Level of aggregation" give the year of record and the level of spatial aggregation the variables are provided on. "Missing values" is true if and only if there are missing values for any municipal associations after removing associations with zero population and the municipal associations of Lohheide (RS 33519501) and Zandt (93720177) from the sample.

Table III |

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Arbeitslose	unemployed	*Not included in the analysis but substituted by per capita variable*	1	2019	ma	False
sozialversicherungspflichtige Beschäftigte am Arbeitsort	employees at place of work	*Not included in the analysis but substituted by per capita variable*; employees subject to social insurance contributions at place of work	2	2019	ma	False
sozialversicherungspflichtig Beschäftigte am Wohnort	employees at place of residence	*Not included in the analysis but substituted by per capita variable*; employees subject to social insurance contributions at place of residence	3	2019	ma	False
Bevölkerung gesamt	population		4	2019	ma	False
Erwerbsfähige Bevölkerung (15 bis unter 65 Jahre)	population of working age	*Not included in the analysis but substituted by per capita variable*; population aged between 15 and 65 years	7	2019	ma	False
Bevölkerung (mit Korrektur VZ 1987/Zensus 2011)	population (Census 2011)	population adjusted for the censuses in 1987 (and 1981 in the GDR) and 2011	8	2019	ma	False
Bodenfläche gesamt	floor space	total floor space	9	2019	ma	False
Baugenehmigungen für Wohnungen	planning permissions for flats	planning and building permissions per 1000 inhabitants	46	2019	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Baugenehmigungen für Wohnungen in Ein- und Zweifamilienhäusern	planning permissions for (semi-) detached houses	*Not included in the analysis but substituted by per capita variable*; share of planning and building permissions for flats in detached and semi-detached houses in the total planning and building permissions (percentage)	47	2019	ma	True
Baugenehmigungen für Wohnungen in Mehrfamilienhäusern	planning permissions for flats (multiple dwelling units)	*Not included in analysis but substituted by per capita variable*; share of planning and building permissions for flats in multiple dwelling units in the total planning and building permissions (percentage)	48	2019	ma	True
Fertiggestellte Wohnungen im Bestand	completed flats	completed flats per 1000 flats in stock	49	2019	ma	False
Neubauwohnungen je Einwohner	completed flats (per capita)	completed flats per 1000 inhabitants	52	2019	ma	False
Neubauwohnungen in Ein- und Zweifamilienhäusern je Einwohner	completed (semi-) detached houses (per capita)	completed flats in (semi-) detached houses per 1000 inhabitants	53	2019	ma	False
Neubauwohnungen in Mehrfamilienhäusern	completed flats in multiple dwelling units (per capita)	completed flats in multiple dwelling units per 1000 inhabitants	54	2019	ma	False
Ein- und Zweifamilienhäuser	(semi-) detached houses	share of (semi-) detached houses in total stock of houses (percentage)	57	2019	ma	False
Mehrfamilienhäuser	multiple dwelling units	share of multiple dwelling units with three or more flats in the total stock of houses (percentage)	58	2019	ma	False
Ein- und Zweiraumwohnungen	studio and two-room apartments	share of flats with one or two rooms in total stock of flats (percentage)	59	2019	ma	True
5- und mehr Raum-Wohnungen	flats with 5+ rooms	share of flats with five or more rooms in total stock of flats (percentage)	60	2019	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Wohnungen in Ein- und Zweifamilien-häusern	flats in (semi-) detached houses	share of flats in (semi-) detached houses in all flats (percentage)	62	2019	ma	False
Wohnungen in Mehrfamilien-häusern	flats in multiple dwelling units	share of flats in multiple dwelling units in all flats (percentage)	63	2019	ma	False
Beschäftigtenquote	employment rate		64	2019	ma	False
Beschäftigtenquote Frauen	employment rate of women		65	2019	ma	False
Beschäftigtenquote Männer	employment rate of men		66	2019	ma	False
Verhältnis junge zu alte Erwerbsfähige	proportion young/elderly	proportion of young (15-19) to elderly (60-64) population (percentage)	67	2019	ma	False
Einwohner unter 6 Jahre	population under 6 years		111	2019	ma	False
Einwohner von 6 bis unter 18 Jahren	population (6-17 years)		112	2019	ma	False
Einwohner von 18 bis unter 25 Jahren	population (18-24 years)		113	2019	ma	False
Einwohner von 25 bis unter 30 Jahren	population (25-29 years)		114	2019	ma	False
Einwohner von 30 bis unter 50 Jahren	population (30-49 years)	share of population of given age in total population (percentage)	115	2019	ma	False
Einwohner von 50 bis unter 65 Jahren	population (50-64 years)		116	2019	ma	False
Einwohner 65 Jahre und älter	population (65+ years)		117	2019	ma	False
Einwohner 75 Jahre und älter	population (75+ years)		118	2019	ma	False
Einwohner von 65 bis unter 75 Jahren	population (65-74 years)		123	2019	ma	False
Einwohner unter 3 Jahren	population under 3 years		129	2019	ma	False
Einwohner von 3 bis unter 6 Jahren	population (3-6 years)		130	2019	ma	False
Weibliche Einwohner 75 Jahre und älter	female population (75+ years)	share of women in the age group	119	2019	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Weibliche Einwohner von 18 bis unter 25 Jahren	female population (18-24 years)	share of women in the age group	120	2019	ma	False
Weibliche Einwohner von 25 bis unter 30 Jahren	female population (25-29 years)	share of women in the age group	121	2019	ma	False
Weibliche Einwohner 65 Jahre und älter	female population (65+ years)	share of women in the age group	122	2019	ma	False
Weibliche Einwohner von 65 bis unter 75 Jahren	female population (65-74 years)	share of women in the age group	124	2019	ma	False
Durchschnittsalter der Bevölkerung	mean age		131	2019	ma	False
Abhängigenquote Junge	rate of young dependants	population under 15 years per population of working age (percentage)	138	2019	ma	False
Abhängigenquote Alte	rate of elderly dependants	population with minimum age of 65 years per population of working age (percentage)	139	2019	ma	False
Frauenanteil	percentage of women		140	2019	ma	False
Frauenanteil 20 bis unter 40 Jahre	percentage of women (20-39 years)		141	2019	ma	False
Gesamtwanderungssaldo	net migration	net migration per 1000 inhabitants	146	2019	ma	False
Geborene	births	births per 1000 inhabitants	171	2019	ma	False
Gestorbene	deaths	deaths per 1000 inhabitants	172	2019	ma	False
Natürlicher Saldo	balance of births and deaths	differences between births and deaths per 1000 inhabitants	173	2019	ma	False
Siedlungs- und Verkehrsfläche	settlement and traffic area	share of settlement and traffic area in total area (percentage)	237	2019	ma	False
Siedlungsdichte in km²	habitat density	population per settlement and traffic area	238	2019	ma	False
Erholungsfläche	recreational area	share of recreational area in total area (percentage)	239	2019	ma	False
Erholungsfläche je Einwohner	per capita recreational area		240	2019	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Freifläche	open area	share of open area in total area (percentage)	241	2019	ma	False
Freifläche je Einwohner	per capita open area		242	2019	ma	False
Landwirtschaftsfläche	agricultural land	share of agricultural land in total area (percentage)	243	2019	ma	False
Naturnähere Fläche	semi-natural area	share of semi-natural area in total area (percentage)	244	2019	ma	False
Naturnähere Fläche je Einwohner	per capita semi-natural area		245	2019	ma	False
Waldfläche	forest area	share of forest area in total area (percentage)	246	2019	ma	False
Wasserfläche	water expanse	share of water expanse in total area (percentage)	247	2019	ma	False
Steuerkraft	fiscal capacity	per capita fiscal capacity	274	2019	ma	True
Einkommensteuer	income tax	per capita income tax	275	2019	ma	False
Gewerbesteuer	trade tax	per capita trade tax	276	2019	ma	False
Umsatzsteuer	value-added tax	per capita value-added tax	277	2019	ma	False
Einwohnerdichte	population density		299	2019	ma	False
Einwohner-Arbeitsplatz-Dichte	density of residents and employees		300	2019	ma	False
Regionales Bevölkerungspotenzial	regional potential of population	sum of population within 100km weighted by area of municipality	301	2019	ma	False
Einpendler	in-commuter	share of in-commuters in all employees at place of work (percentage)	351	2019	ma	False
Auspendler	out-commuter	share of out-commuters in all employees at place of residence (percentage)	352	2019	ma	False
Pendlersaldo	net in-commuter	share of difference of in-commuter and out-commuter in all employees at place of work (percentage)	353	2019	ma	False
Pendler mit Arbeitsweg 50 km und mehr	commuter travelling 50km+	share of commuter travelling at least 50km in all employees at place of residence (percentage)	354	2019	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Pendler mit Arbeitsweg 150 km und mehr	commuter travelling 150km+	share of commuter travelling at least 150km in all employees at place of residence (percentage)	355	2019	ma	False
Pendler mit Arbeitsweg 300 km und mehr	commuter travelling 300km+	share of commuter travelling at least 300km in all employees at place of residence (percentage)	356	2019	ma	False
Erreichbarkeit von Autobahnen	accessibility of highways	average journey time to closest highway by car (in minutes)	334	2020	ma	False
Erreichbarkeit von Flughäfen	accessibility of airports	average journey time to closest German international airport by car (in minutes)	335	2020	ma	False
Erreichbarkeit von IC/EC/ICE-Bahnhöfen	accessibility of intercity train station	average journey time to closest intercity train station by car (in minutes)	336	2020	ma	False
Erreichbarkeit von Oberzentren	accessibility of regional centre	average journey time to closest regional centre by car (in minutes)	337	2020	ma	False
Erreichbarkeit von Mittelzentren	accessibility of middle-order centre	average journey time to closest middle-order centre by car (in minutes)	338	2020	ma	False
Nahversorgung Supermärkte Durchschnitts-distanz	mean distance supermarket		339	2017	ma	False
Nahversorgung Supermärkte Anteil der Bev. 1km Radius	distance supermarket max 1km	share of population living within 1km to the closest supermarket (percentage)	340	2017	ma	False
Nahversorgung Apotheken Durchschnitts-distanz	mean distance pharmacy		341	2017	ma	False
Nahversorgung Apotheken Anteil der Bev. 1km Radius	distance pharmacy max 1km	share of population living within 1km to the closest pharmacy (percentage)	342	2017	ma	False
Nahversorgung Grundschulen Durchschnitts-distanz	mean distance primary school		343	2018	ma	False
Nahversorgung Grundschulen Anteil der Bev. 1km Radius	distance primary school max 1km	share of population living within 1km to the closest primary school (percentage)	344	2018	ma	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Nahversorgung Haltestellen des ÖV Durchschnittsdistanz	mean distance public transport	mean distance to the closest stop or station	345	2018	ma	True
Nahversorgung Haltestellen des ÖV Anteil der Bev. 1km Radius	distance public transport max 1km	share of population living within 1km to the closest stop or station (percentage)	346	2018	ma	True
Arbeitslosenquote	unemployment rate	(percentage)	11	2019	c	False
Arbeitslosenquote Frauen	unemployment rate of women	(percentage)	12	2019	c	False
Arbeitslosenquote Männer	unemployment rate of men	(percentage)	14	2019	c	False
Baulandpreise	price of building ground		45	2019	c	True
Fertiggestellte Wohngebäude mit erneuerbarer Heizenergie	completed buildings with renewable heat energy systems	share of buildings with renewable heat energy systems in all completed buildings (percentage)	55	2019	c	False
Fertiggestellte Wohnungen mit erneuerbarer Heizenergie	completed flats with renewable heat energy systems	share of completed flats in buildings with renewable heat energy systems in completed flats in all buildings (percentage)	56	2019	c	False
Wohnfläche	per capita residential floor area		61	2019	c	False
Erwerbsquote	labour force participation rate	share of labour force in total population of working age (percentage)	73	2019	c	False
Selbständigenquote	self-employment rate	(percentage)	76	2019	c	False
Beschäftigte ohne Berufsabschluss	employees without any professional qualification	(percentage)	77	2019	c	False
Beschäftigte mit Berufsabschluss	employees with professional qualification	(percentage)	78	2019	c	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Beschäftigte mit akademischem Berufsabschluss	employees with academic qualification	(percentage)	79	2019	c	False
Beschäftigte mit Anforderungsniveau Experte	expert employees	(percentage)	80	2019	c	False
Beschäftigte mit Anforderungsniveau Spezialist	specialist employees	(percentage)	81	2019	c	False
Beschäftigte mit Anforderungsniveau Fachkraft	skilled employees	(percentage)	82	2019	c	False
Beschäftigte mit Anforderungsniveau Helfer	assistant employees	(percentage)	83	2019	c	False
Industriequote	employees in industry	(percentage)	99	2019	c	True
Dienstleistungsquote	employees in non-productive industries	(percentage)	101	2019	c	True
Beschäftigte in unternehmensbezogenen Dienstleistungen	employees in business services	(percentage)	102	2019	c	False
Erwerbstätige Primärer Sektor	labour force in primary sector	(percentage)	103	2019	c	False
Erwerbstätige Sekundärer Sektor	labour force in secondary sector	(percentage)	104	2019	c	False
Erwerbstätige Tertiärer Sektor	labour force in tertiary sector	(percentage)	105	2019	c	False
Beschäftigte in Kreativbranchen	employees in creative industries	(percentage)	106	2019	c	False
Beschäftigte in wissensintensiven Industrien	employees in knowledge-intensive industries	(percentage)	107	2019	c	True

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Anteil Erwerbstätige Verarbeitendes Gewerbe an Industrie	labour force in manufacturing industry	share of labour force working in the manufacturing industry in labour force of entire secondary sector (percentage)	108	2019	c	False
Anteil Erwerbstätige Finanz- und Unternehmensdienstleistungen	labour force in financial and business services	share of labour force in financial and business services in labour force of entire tertiary sector (percentage)	109	2019	c	False
Berufsschüler	vocational school students	vocational school students per 1000 inhabitants	192	2019	c	False
Ausbildungsplätze	apprenticeship positions	apprenticeship positions per 100 requested positions (percentage)	195	2019	c	True
Auszubildende	apprentices	apprentices per 1000 employees	196	2019	c	False
Weibliche Auszubildende	female apprentices	share of women in all apprentices (percentage)	197	2019	c	False
Männliche Auszubildende	male apprentices	share of men in all apprentices (percentage)	198	2019	c	False
Studierende	students at universities	students at universities and universities of applied science per 1000 inhabitants	199	2019	c	False
Studierende an FH	students at universities of applied science	students at universities of applied science per 1000 inhabitants	200	2019	c	False
Auszubildende je 100 Einwohner 15 bis 25 Jahre	apprentices per young inhabitant	apprentices per 100 15- to 24-year old inhabitants	205	2019	c	False
Studierende je 100 Einwohner 18 bis 25 Jahre	students per young inhabitant	university students per 100 18- to 24-year old inhabitants	206	2019	c	False
Schüler	pupils	pupils per 100 inhabitants	207	2019	c	False
Ausländische Schüler	foreign pupils	share of foreign pupils in all pupils (percentage)	208	2019	c	False
Ausländische Schüler je 100 Ausländer 6 bis 18 Jahre	schooling rate of foreign youngsters	foreign pupils per 100 foreign inhabitants aged 6 to 17 years	209	2019	c	False

267

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Schulabgänger mit Hauptschul-abschluss	Certificates of Secondary Education	share of school leavers with a Certificate of Secondary Education in all school leavers (percentage)	210	2019	c	False
Weibliche Schulab-gänger mit Hauptschul-abschluss	Certificates of Secondary Education (females)	(percentage)	211	2019	c	False
Männliche Schulab-gänger mit Hauptschul-abschluss	Certificates of Secondary Education (males)	(percentage)	212	2019	c	False
Schulabgänger mit mittlerem Abschluss	General Certificates of Secondary Education	share of school leavers with a General Certificate of Secondary Education in all school leavers (percentage)	213	2019	c	False
Weibliche Schulab-gänger mit mittlerem Abschluss	General Certificates of Secondary Education (females)	(percentage)	214	2019	c	False
Männliche Schulab-gänger mit mittlerem Abschluss	General Certificates of Secondary Education (males)	(percentage)	215	2019	c	False
Schulabgänger mit allgemeiner Hochschulreife	German university entrance qualification (Abitur)	share of school leavers with a German university entrance qualification (Abitur) in all school leavers (percentage)	216	2019	c	True
Schulabgänger ohne Abschluss	school dropout	share of school dropouts in all school leavers (percentage)	219	2019	c	False
Weibliche Schulab-gänger ohne Abschluss	female school dropout	(percentage)	220	2019	c	False
Männliche Schulab-gänger ohne Abschluss	male school dropout	(percentage)	221	2019	c	False
Bruttoverdienst	mean gross income		222	2019	c	False
Bruttoverdienst im Produzierenden Gewerbe	mean gross income in industry		223	2019	c	False
Medianeinkom-men	median income	median per capita gross income	224	2019	c	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Medianeinkommen 25 bis unter 55-Jährige	median income (25-54 years)	median per capita gross income of 25- to 54-year olds	227	2019	c	False
Medianeinkommen 55 bis unter 65-Jährige	median income (55-64 years)	median per capita gross income of 55- to 64-year olds	228	2019	c	False
Haushaltseinkommen	household income	per capita mean household income	229	2019	c	False
Verbraucherinsolvenzverfahren	insolvency proceedings	insolvency proceedings of consumers per 1000 adults	230	2019	c	False
Schuldnerquote	dept rate		233	2019	c	False
Arbeitsvolumen	volume of labour	yearly working hours per employee	234	2019	c	False
Medianeinkommen anerkannter Berufsabschluss	median income (professional qualification)	median monthly income of employees with any professional qualification	235	2019	c	False
Medianeinkommen akademischer Berufsabschluss	median income (academic qualification)	median monthly income of employees with any academic qualification	236	2019	c	False
Schlüsselzuweisungen	rate support grant	per capita rate support grant	278	2019	c	True
Kommunale Schulden	communal debt	per capita communal debt	279	2019	c	False
Kassenkredite	cash credit	per capita cash credit	280	2019	c	False
Personal der Kommunen	staff of communal administration	members of staff of communal administration per 10000 inhabitants	281	2019	c	False
Ausgaben für Sachinvestitionen	expenditure on fixed assets	per capita expenditure on fixed assets	282	2019	c	True
Zuweisungen für Investitionsfördermaßnahmen	assignments of investment	per capita assignments of investment	283	2019	c	True
Städtebauförderung (langfristig)	long-term urban development promotion program	per capita budgeted national long-term urban development promotion program	284	2019	c	False
Städtebauförderung (kurzfristig)	short-term urban development promotion program	per capita budgeted national short-term urban development promotion program	290	2019	c	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Ländlichkeit	rurality	share of population living in municipalities with less than 150 inhabitants per square km	296	2019	c	False
Bevölkerung in Mittelzentren	population in middle-order centre	share of population living in middle-order centre (percentage)	297	2019	c	False
Bevölkerung in Oberzentren	population in regional centre	share of population living in regional centre (percentage)	298	2019	c	False
SGB II - Quote	rate of basic security benefits	proportion of basic security benefits for job seekers and unemployed (percentage)	302	2019	c	False
Wohngeldhaushalte	housing subsidies	share of household receiving housing subsidies (percentage)	304	2019	c	False
Wohngeldhaushalte (Mietzuschuss)	housing subsidies of tenants	proportion of households receiving housing subsidies that rent a dwelling, i.e., they receive rent subsidies (percentage)	305	2019	c	False
Wohngeldhaushalte (Lastenzuschuss)	housing subsidies of houseowners	proportion of households receiving housing subsidies that own the dwelling (percentage)	306	2019	c	False
Empfänger von Grundsicherung im Alter (Altersarmut)	old-age poverty	proportion of the population aged 65+ years receiving guaranteed minimum pensions (percentage)	307	2019	c	False
Empfänger von Mindest-sicherungen	social welfare	proportion of the population receiving any kind of social welfare (percentage)	311	2019	c	False
Schlafgelegenheiten in Beherber-gungsbetrieben	beds in hospitality businesses	beds in hospitality businesses per 1000 inhabitants	357	2019	c	False
Bruttoinlandsprodukt je Einwohner	per capita GDP	per capita gross domestic product	365	2019	c	False
Bruttoinlandsprodukt je Erwerbstätigen	GDP per employee		366	2019	c	False
Bruttowertschöpfung je Erwerbstätigen	GVA per employee	gross value added per employee	367	2019	c	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Bruttowertschöpfung je Erwerbstätigen Primärer Sektor	GVA per employee in primary sector		368	2019	c	True
Bruttowertschöpfung je Erwerbstätigen Sekundärer Sektor	GVA per employee in secondary sector		369	2019	c	False
Bruttowertschöpfung je Erwerbstätigen Tertiärer Sektor	GVA per employee in tertiary sector		370	2019	c	False
Umsatz im Bergbau u. Verarb. Gewerbe	business volume in mining and manufacturing industry	business volume in mining and manufacturing industry per employee	373	2019	c	True
Umsatz Bauhauptgewerbe	business volume in main construction industry	business volume in main construction industry per employee	374	2019	c	True
Anteil Bruttowertschöpfung Primärer Sektor	GVA primary sector	share of GVA of primary sector in GVA (percentage)	375	2019	c	True
Anteil Bruttowertschöpfung Sekundärer Sektor	GVA secondary sector	share of GVA of secondary sector in GVA (percentage)	376	2019	c	False
Anteil Bruttowertschöpfung Tertiärer Sektor	GVA tertiary sector	share of GVA of tertiary sector in GVA (percentage)	377	2019	c	False
Beschäftigte im Handwerk	craftspeople	share of craftsmen and craftswomen in all employees (percentage)	110	2018	c	False
Kleinstbetriebe	microentities	share of microentities in all enterprises (percentage)	361	2018	c	False
Kleinbetriebe	small enterprises	share of small enterprises in all enterprises (percentage)	362	2018	c	False
Mittlere Unternehmen	medium-sized enterprises	share of medium-sized enterprises in all enterprises (percentage)	363	2018	c	False
Großunternehmen	large-scale enterprises	share of large-scale enterprises in all enterprises (percentage)	364	2018	c	False

Continued on next page

Table III | (Continued)

German name in original INKAR dataset	English name used during analysis	Explanation	ID in INKAR	Time period	Level of aggregation	Missing values
Umsatz im Handwerk	business volume of skilled crafts and trades	business volume of skilled crafts and trades per employee	378	2018	c	True

Supplementary Note 4: The 2011 Census

To obtain data on the size of households and the owner-occupation of dwellings, we consulted the German 2011 Census [13]. To ensure consistency with all areal units considered by INKAR, we utilized data from the Census on the municipal level and mapped them to municipal associations according to the procedure described in Supplementary Note 2.

Due to the high level of granularity, the data sets contained imprecise values for reasons of protection of data privacy. We neglected the imprecision and included the values without further modification.

Sufficient data was available for all municipalities except the municipality of Dierfeld (AGS 7231021), which constitutes a very small municipality of eleven inhabitants. For Dierfeld, all values needed for the computations of the ownership occupation ratio and the variables related to the size of households were missing. As the municipality belongs to the municipal association of Wittlich-Land (RS 72315008) along with five other municipalities [9, file 3], we excluded Dierfeld from the data set and based the ownership occupation and household sizes for Wittlich-Land on the five municipalities remaining.

The data set on dwellings and buildings (German: 'Gebäude und Wohnungen') of the Census [14] contains the number of owner-occupied flats ($\#\ \text{flats}_{\text{owner}}(M)$, German name in the Census: "Wohnungen in Gebäuden mit Wohnraum nach Art der Nutzung - Art der Wohnungsnutzung - Von Eigentümer/-in bewohnt") and the number of all flats ($\#\ \text{flats}_{\text{total}}(M)$, German name in Census: "Wohnungen in Gebäuden mit Wohnraum nach Art der Nutzung - Insgesamt") existing in all municipalities. To obtain the 'ownership occupation ratio' (share of $\text{flats}_{\text{owner}}(\text{MA})$), we computed

$$\text{share of flats}_{\text{owner}}(\text{MA})\ [\%] = 100 \cdot \frac{\sum\limits_{M\in MA} \#\ \text{flats}_{\text{owner}}(M)}{\sum\limits_{M\in MA} \#\ \text{flats}_{\text{total}}(M)}. \tag{1}$$

The data set on households and families (German: 'Haushalte und Familien') of the Census [15] served as the foundation of the variables on the size of households we included in the input data set. Specifically, we utilized the variables

- $\#\ \text{households}_{\text{total}}$: the number of all households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Insgesamt"),
- $\#\ \text{households}_{\text{1-p.}}$: the number of 1-person households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 1 Person"),
- $\#\ \text{households}_{\text{2-p.}}$: the number of 2-person households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 2 Personen"),
- $\#\ \text{households}_{\text{3-p.}}$: the number of 3-person households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 3 Personen"),
- $\#\ \text{households}_{\text{4-p.}}$: the number of 4-person households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 4 Personen"),
- $\#\ \text{households}_{\text{5-p.}}$: the number of 5-person households (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 5 Personen") and
- $\#\ \text{households}_{\text{6+-p.}}$: the number households with at least 6 inhabitants (German name in the Census: "Private Haushalte nach Größe des privaten Haushalts - Größe des privaten Haushalts - 6 und mehr Personen").

Having excluded the municipality of Dierfeld, $\#\ \text{households}_{\text{total}}$ was available for all municipalities. However, $\#\ \text{households}_{\text{1-p.}}$, $\#\ \text{households}_{\text{2-p.}}$, $\ldots$, $\#\ \text{households}_{\text{6+-p.}}$ were missing for some municipalities. Analyses revealed that for all municipalities with missing values, the non-missing numbers of households of different sizes added up to the total number of households. Thus, we substituted the missing values by zero and obtained a complete data set on the level of municipalities.

Based on this data set, the variables

$$\text{'share of X-person households' (MA) [\%]} = 100 \cdot \frac{\sum\limits_{M\in MA} \#\ \text{households}_{\text{X-p.}}(M)}{\sum\limits_{M\in MA} \#\ \text{households}_{\text{total}}(M)} \tag{2}$$

for $X \in \{1, 2, 3, 4, 5, 6+\}$ defined the share of households of the respective sizes on the level of municipal associations. We referred to these variables as 'share of 1-person households', 'share of 2-person households', ... in the input data set. In addition to the shares of sizes of households, we added the total number of households in a municipal association ('count households') as an input feature. It is defined as the sum over all numbers of households of municipalities (M) belonging to a municipal association MA,

$$\text{'count households' (MA)} = \sum_{M \in MA} \# \text{ households}_{\text{total}}(M). \tag{3}$$

Supplementary Note 5: Regional Database of Germany

The Regional Database of Germany (Regionaldatenbank Deutschland) [16] run by the Federal Statistical Office and the regional bureaus of statistics publishes election results on the level of municipalities and counties and states.

To obtain election results for municipal associations, we started from the following variables provided by [16]

- the total number of valid votes ($\#\text{votes}_{\text{total}}$, German name in original dataset: "Gültige Zweitstimmen - Anzahl"),
- the number of votes for the CDU/CSU ($\#\text{votes}_{\text{CDU}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - CDU/CSU - Anzahl"),
- the number of votes for the SPD ($\#\text{votes}_{\text{SPD}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - SPD - Anzahl"),
- the number of votes for the Greens ($\#\text{votes}_{\text{Green}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - Grüne - Anzahl"),
- the number of votes for the FDP ($\#\text{votes}_{\text{FDP}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - FDP - Anzahl"),
- the number of votes for The Left ($\#\text{votes}_{\text{Left}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - DIE LINKE - Anzahl"),
- the number of votes for the AfD ($\#\text{votes}_{\text{AfD}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - AfD - Anzahl") and
- the number of votes for other parties ($\#\text{votes}_{\text{other}}$, German name in original dataset: "Gültige Zweitstimmen - Parteien - Sonstige Parteien - Anzahl").

The preprocessing required three steps:

1. First, as the Regional Database does not consider city-states as municipalities but only as counties, we distinguished between city and area states to include the cities of Bremen, Bremerhaven, Hamburg and Berlin manually as municipal associations.
2. Second, we applied the extended mapping of municipalities to municipal associations to upscale the election data in area states (see Supplementary Note 2). The following paragraph describes this step in more detail. The resulting data set missed data for 102 municipal associations, which all were cities. The Regional Database does not consider these cities as municipalities but only as counties.
3. We added the election results of these cities.

On step 2: The set of all municipalities of the areal states comprised 13,384 municipalities. Election data was entirely missing for 2444 of these municipalities, which we thus excluded from further analysis. For another 260 municipalities, the number of votes was missing for at least one party. As the sum of all non-missing numbers of votes for parties was equal to the total number of votes in each municipality, we substituted zero for the missing values. Applying the extended mapping of municipalities to municipal associations described in Supplementary Note 2 and summing up election data over all municipalities belonging to a municipal association, i.e.,

$$\#\text{ votes}_{\text{party}}(MA) = \sum_{M \in MA} \#\text{ votes}_{\text{party}}(M) \text{ for party} \in \{\text{AfD, CDU, FDP, Green, Left, SPD, other}\} \quad (4)$$

$$\text{and } \#\text{ votes}_{\text{total}}(MA) = \sum_{M \in MA} \#\text{ votes}_{\text{total}}(M), \quad (5)$$

yielded election data on the level of municipal associations.

For the resulting complete set of election results for areal and city states, we divided the numbers of votes for specific parties by the total number of votes, which led to the relative share of votes for each party, i.e,

$$\text{share of votes}_{\text{party}}(MA) = 100 \cdot \frac{\#\text{ votes}_{\text{party}}(MA)}{\#\text{ votes}_{\text{total}}(MA)}. \quad (6)$$

We included these percentages of votes for all parties in the input data set and refer to the variables in the data set by the respective party names, i.e., AfD, CDU/CSU, FDP, The Greens, The Left, SPD and other parties.

Supplementary Note 6: Solar Radiation Data

Overview

Data on the average global radiation is computed from two data sets. The first data set published by the Federal Statistical Office consists of a registry of all municipalities as of 31.12.2019, which includes the longitudinal and latitudinal position of the geographical middle points of all municipalities [17]. The second data set provides the average global radiation (in $\frac{W}{m^2}$) on an optimally inclined surface as a geographical grid [18] and is described in detail below. Based on these two data sets, we computed the approximate global radiation present in all municipal associations by applying the following steps:

Input: registry of municipalities, set of all municipal associations considered by preprocessed INKAR dataset (*INKAR set*), radiation grid
for each municipality M in registry **do**
    (Step 1) Get the radiation of the grid cell that contains the coordinates of M.
end for
for each municipal association MA in the INKAR set **do**
    (Step 2) Take the average over the solar radiation of all municipalities in MA.
end for
Output: Average global radiation for all municipal associations contained in the INKAR set.

The SARAH database

The Joint Research Centre of the European Commission supplies the *SARAH Solar Radiation Data* [18], which is data on the average global radiation in Europe, Asia and Africa between 2005 and 2016. We chose to include the "Yearly average global irradiance on an optimally inclined surface ($\frac{W}{m^2}$)" [19] assuming that most resident PV systems are installed on inclined roofs. We referred to the website of the SARAH Solar Radiation Data initiative [18] and two continuative studies [20, 21] for further information on the radiation data.

The radiation data is provided as a grid covering latitudes from $40°S$ to $62°30'N$ and longitudes from $65°W$ to $128°E$ [19]. Encoding latitude and longitude as one variable each, this is equivalent to latitudes from $-40°$ ($\text{lati}_{\text{south end}}$) to $62°30'$ and longitudes from $-65°$ ($\text{long}_{\text{west end}}$) to $128°$. The grid cells are $3'$ wide which corresponds to $0.05°$ ($\text{width}_{\text{grid cell}}$). In step 1, we identify the column index of the grid cell covering the latitudinal and longitudinal coordinates of M by

$$\text{col}_M = \left\lceil \frac{\text{long}_M - \text{long}_{\text{westend}}}{\text{width}_{\text{grid cell}}} \right\rceil - 1 = \left\lceil \frac{\text{long}_M + 65°}{0.05°} \right\rceil - 1 \tag{7}$$

and the row index by

$$\text{row}_M = \#\text{ rows of grid} - 1 - \left\lceil \frac{\text{lati}_M - \text{lati}_{\text{southend}}}{\text{width}_{\text{grid cell}}} \right\rceil \tag{8}$$

$$= \#\text{ rows of grid} - 1 - \left\lceil \frac{\text{lati}_M + 40°}{0.05°} \right\rceil, \tag{9}$$

where '# rows of grid' corresponds to the total number of rows of the grid. Global radiation of the grid cell then yielded the radiation for M. Taking the mean over the global radiation recorded for all municipalities belonging to a municipal association according to the registry (*Step* 2 of the algorithm stated above lead to an approximation of the radiation present in the association.

Supplementary Note 7: Data of Solar PV Systems

Overview

Owners of photovoltaic systems are obliged by law to register installations directly or indirectly connected to the power grid with the *Marktstammdatenregister* run by the Federal Network Agency (Bundesnetzagentur) [22]. Given the completeness and accessibility of the database, the *Markstammdatenregister* served as a source for data on PV systems.

PV systems are typically characterized by their peak power. Most PV systems with a peak power of up to $10kW_p$ are installed on the rooftops of residential buildings, while larger installations of up to $30kW_p$ are often placed on top of industrial buildings or stables [23]. To focus on small-scale PV systems installed on residential buildings, we restricted our analysis to PV systems with a net nominal power of at most 10 kW in accordance with the literature [24, 25].

Our ML model is designed to predict the accumulated (peak) power of all PV systems installed within a municipal association. The number of PV installations evidently depends on the size of a municipal association; hence, we normalize our target. As our study aims to analyze the decision of households for or against the installation of a PV system, we normalize the accumulated peak power by the number of households in a municipal association,

$$\frac{\text{accumulated (peak) power of PV systems } (\leq 10\ kW_p) \text{ installed in the MA before } 01/01/2022}{\text{number of households living in the MA}} \quad \left[\frac{kW_p}{hh}\right]. \tag{10}$$

Data processing

The *Marktstammdatenregister* provides a download of the entire database [26]. Table IV gives the filters we applied to extract relevant data from the download.

For all PV installations, the *Marktstammdatenregister* includes the AGS key of the municipalities in which it is installed. We summed up the values of (approximated peak) power ('Nettonennleistung') of all PV systems installed in a municipality to derive the accumulated power. An aggregation according to the mapping introduced in Supplementary Note 2 yielded the accumulated power of all PV systems installed within a municipal association. Normalizing by the number of households living in a municipal association (given by 'count households' derived from the 2011 Census) led to the target feature used for our analysis given in equation 10. We refer to the target variable as 'accumulated power' in the data set.

The net nominal power only varies slightly from the peak power for most small-scale PV systems and thus approximates the peak power in this work. Therefore, talking of the peak power of PV installations, we refer to this approximation, i.e., the net nominal power ('Nettonennleistung').

The *Marktstammdatenregister* indicates an unrealistic commissioning date, e.g., 1950 or 1906, for a small number of PV systems, which may go back to mistakes during registration. Nonetheless, we included the installations in the data set strictly sticking to the filtering presented in table IV.

Table IV: Filtering applied to the PV data taken from the Marktstammdatenregister.

Name of Filter	Values
registration date (Registrierungsdatum der Einheit)	before 01/11/2022 (vor 01/11/2022)
commissioning date (Inbetriebnahmedatum der Einheit)	before 01/01/2022 (vor 01/01/2022)
operating status (Betriebsstatus)	unequal to planned (entspricht nicht 'In Planung')
energy source (Energieträger)	solar radiation energy (entspricht 'Solare Strahlungsenergie')
net nominal power (Nettonennleistung)	$\leq 10kW_p$

Supplementary Note 8: Data of BEV Adoption

The German Federal Motor Transport Authority (Kraftfahrt-Bundesamt, KBA) reports the number of registered vehicles for every municipality in Germany as of the 1st of January, the 1st of April, the 1st of July and the 1st of October for every year [27]. Municipalities are identified by their name and by the first five digits of the RS. Hence, every entry can be easily mapped to a municipal association and we can aggregate the number of vehicles per municipal association.

The number of electric vehicles is reported on the level of municipalities since 1st of October 2022. The KBA distinguishes between (i) battery electric vehicles and (ii) plug-in hybrid electric vehicles. Furthermore, it differentiates between (i) private owners and commercial owners. In this article, we consider only privately owned battery electric vehicles. The target of our ML model is obtained by dividing this number by the total number of privately owned vehicles per municipal association.

Subsequently, the changing of RS over time had to be considered to merge the data sets with the charging station data and the data set used in the analysis of the adoption of PV, containing PV installations and socio-economic features. This was done by examining the changes (e.g., by fusion or splitting up of a municipal association) over time and aggregating the respective municipal associations to arrive at a consistent RS. A list of the considered changed is provided in the data folder.

Supplementary Note 9: Data of charging stations

The Federal Network Agency (Bundesnetzagentur) registers all public charging stations with the individual address and the geospatial coordinates [28]. These geospatial coordinates are mapped to municipal associations using shape files provided by the German Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodäsie) using geopandas [29]. Shapefiles are provided for all municipal associations in Germany at different reference dates [30].

In some cases, the geospatial data is obviously wrong as it maps to points outside of Germany. In these cases, we have manually determined the respective municipal association using the given street address.

As in 8, the changing of RS over time was considered by mapping them to common values. Thus, we could arrive at data set containing both the data on charging stations per capita and the share of BEVs, as well as, the data previouly used in the analysis of PV adoption.

SUPPLEMENTARY REFERENCES

* d.witthaut@fz-juelich.de

[1] Microsoft Corporation. LightGBM - Parameters. https://lightgbm.readthedocs.io/en/latest/Parameters.html (2023). Revision ffb29862, accessed: 17/04/2023.

[2] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389–422 (2002).

[3] The interval includes the range of values identified in the literature ([0.5, 0.8]) but does not limit the HPO to it.

[4] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).

[5] scikit-learn developers. Scikit-learn - sklearn.preprocessing.StandardScaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler (2023).

[6] Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[7] scikit-learn developers. Scikit-learn - sklearn.linear_model.Lasso. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (2023).

[8] Bundesamt für Bauwesen und Raumordnung. Laufende Raumbeobachtung - Raumabgrenzungen. https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Raumabgrenzungen/deutschland/gemeinden/gemeinden-gemeindeverbaende/gemeinden.html;jsessionid=D96D39FABA443452F9492FD8BA246FF6.live11312. 2020-2023, accessed: 29/03/2023.

[9] Bundesinstitut für Bau-, Stadt- und Raumforschung. INKAR 2021 (2021). Available at https://www.bbr-server.de/imagemap/inkar/download/inkar_2021.zip, accessed: 06/09/2022.

[10] Statistisches Bundesamt. Regionalschlüssel (RS). https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Glossar/regionalschluessel.html (2023). Accessed: 29/03/2023.

[11] Statistisches Bundesamt. Namens- und Gebietsänderungen der Gemeinden: Gebietsänderungen (Namens-, Grenz- und Schlüsseländerungen) (2023). Available at https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Namens-Grenz-Aenderung/namens-grenz-aenderung.html, accessed: 04/01/2023.

[12] Bundesinstitut für Bau-, Stadt und Raumforschung. INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung. https://www.inkar.de/ (2022).

[13] Statistische Ämter des Bundes und der Länder. Zensus 2011. https://www.zensus2011.de/DE/Home/home_node.htmlhttps://www.zensus2011.de/DE/Zensus2011/zensus2011_node.html (2020). Accessed: 16/05/2023.

[14] Statistische Ämter des Bundes und der Länder. Zensus 2011 - Gebäude und Wohnungen. Available at https://www.zensus2011.de/SharedDocs/Downloads/DE/Pressemitteilung/DemografischeGrunddaten/csv_GebaudeWohnungen.zip?__blob=publicationFile&v=2, accessed: 14/12/2022.

[15] Statistische Ämter des Bundes und der Länder. Zensus 2011 - Haushalte und Familien. Available at https://www.zensus2011.de/SharedDocs/Downloads/DE/Pressemitteilung/DemografischeGrunddaten/csv_HaushalteFamilien.zip?__blob=publicationFile&v=2, accessed: 14/12/2022.

[16] Statistische Ämter des Bundes und der Länder. Regionaldatenbank Deutschland. https://www.regionalstatistik.de/genesis/online/ (2023). Accessed: 09/11/2022.

[17] Statistisches Bundesamt. Alle politisch selbständigen gemeinden mit ausgewählten merkmalen am 31.12.2019. Available at https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GVAuszugJ/31122019_Auszug_GV.html, accessed: 14/12/2022.

[18] European Commission's Joint Research Centre. SARAH Solar Radiation Data. Available at https://joint-research-centre.ec.europa.eu/pvgis-online-tool/pvgis-data-download/sarah-solar-radiation-data_en, accessed: 16/12/2022.

[19] European Commission's Joint Research Centre. Yearly average global irradiance on an optimally inclined surface (W/m2). Available at http://re.jrc.ec.europa.eu/pvg_download/sarahdata/gh_opt_year_sarah.zip, accessed: 16/12/2022.

[20] Urraca, R. *et al.* Extensive validation of CM SAF surface radiation products over Europe. *Remote Sensing of Environment* **199**, 171–186 (2017).

[21] Urraca, R. *et al.* Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. *Solar Energy* **164**, 339–354 (2018).

[22] Bundesministerium der Justiz & Bundesamt für Justiz. Verordnung über das zentrale elektronische Verzeichnis energiewirtschaftlicher Daten (Marktstammdatenregisterverordnung - MaStRV) (2017). Date of entry into force: 10/04/2017, available at https://www.gesetze-im-internet.de/mastrv/BJNR084210017.html, accessed: 17/05/2023.

[23] Linder, S. *Räumliche Diffusion von Photovoltaik-Anlagen in Baden-Württemberg*. Ph.D. thesis, Universität Würzburg, Würzburg (2013).

[24] Baginski, J. P. & Weber, C. Coherent Estimations for Residential Photovoltaic Uptake in Germany Including Spatial Spillover Effects. *HEMF Working Paper* (2019). Available at https://ssrn.com/abstract=3357359.

[25] Dharshing, S. Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany. *Energy Research & Social Science* **23**, 113–124 (2017).

[26] Markstammdatenregister. Datendownload - Gesamtdatenauszug vom Vortag. https://www.marktstammdatenregister.de/MaStR/Einheit/Einheiten/OeffentlicheEinheitenuebersicht (2023). Accessed: 16/11/2022.

[27] Kraftfahrt-Bundesamt. Bestand an Kraftfahrzeugen und Kraftfahrzeuganhängern nach ausgewaehlten Merkmalen (Bun-

deslaendern und Fahrzeugklassen), vierteljaehrlich (FZ 27). URL https://www.kba.de/DE/Statistik/Produktkatalog/produkte/Fahrzeuge/fz27_b_uebersicht.html?nn=3514348.

[28] Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen. Liste der Ladesäulen (Stand 1. Mai 2023). https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/E-Mobilitaet/Ladesaeulenkarte/start.html. (visited on 05/01/2023).

[29] Jordahl, K. *et al.* geopandas/geopandas: v0.8.1. https://doi.org/10.5281/zenodo.3946761 (2020).

[30] Bundesamt für Kartographie und Geodäsie (BKG). Verwaltungsgebiete 1:250 000, Stand 31.12. (VG250 31.12.). https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete/verwaltungsgebiete-1-250-000-stand-31-12-vg250-31-12.html. (visited on 06/01/2023).