Energy and AI 17 (2024) 100386

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Review

A systematic review of spatial disaggregation methods for climate action planning

Shruthi Patil [a,b,*], Noah Pflugradt [a], Jann M. Weinand [a], Detlef Stolten [a,c], Jürgen Kropp [d,b]

[a] Institute of Energy and Climate Research, Techno-economic Systems Analysis (IEK-3), Forschungszentrum Jülich, Wilhelm-Johnen-Straße, Jülich, 52425, NRW, Germany
[b] University of Potsdam, Institute for Environmental Science and Geography, Karl-Liebknecht-Str. 24-25, Potsdam-Golm, 14476, Brandenburg, Germany
[c] Chair for Fuel Cells, RWTH Aachen University, c/o IEK-3, Forschungszentrum Jülich, Jülich, 52425, NRW, Germany
[d] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60, 12 03, Potsdam, D-14412, Brandenburg, Germany

HIGHLIGHTS

- Methods relevant for spatial disaggregation of climate action plans are reviewed.
- Key methods: proxy data, machine learning, and geostatistical model-based approaches.
- Appropriate method depends on domain knowledge, avaibility of local-level data, etc.
- Combining different spatial disaggregation methods can enhance accuracy.
- Spatial disaggregation cannot guarantee perfect accuracy in the results.

ARTICLE INFO

Keywords:
Spatial downscaling
Proxy data
Mass-preserving
Climate action plans
Spatial autocorrelation
Machine learning
Geostatistical models

ABSTRACT

National-level climate action plans are often formulated broadly. Spatially disaggregating these plans to individual municipalities can offer substantial benefits, such as enabling regional climate action strategies and for assessing the feasibility of national objectives. Numerous spatial disaggregation approaches can be found in the literature. This study reviews and categorizes these. The review is followed by a discussion of the relevant methods for the disaggregation of climate action plans. It is seen that methods employing proxy data, machine learning models, and geostatistical ones are the most relevant methods for the spatial disaggregation of national energy and climate plans. The analysis offers guidance for selecting appropriate methods based on factors such as data availability at the municipal level and the presence of spatial autocorrelation in the data.

As the urgency of addressing climate change escalates, understanding the spatial aspects of national energy and climate strategies becomes increasingly important. This review will serve as a valuable guide for researchers and practitioners applying spatial disaggregation in this crucial field.

1. Introduction

1.1. Background

As the effects of climate change are being felt across the globe, the need for mitigation and adaptation action is more urgent than ever. The European Union has recognized the gravity of the situation and called on its member states to develop national energy and climate plans [1].

From a regional perspective, since 2008 the Covenant of Mayors for Climate and Energy has been gathering those local governments that have voluntarily pledged to undertake energy and climate action [2]. In a similar manner as for the countries themselves, these cities are required to submit sustainable energy and climate action plans. Although such a bottom-up approach is being taken by many cities, about 33% of the cities in the European Union still lack a plan [3]. Of those that do have a plan, many have developed them autonomously. Therefore, these plans do not necessarily align with national contingencies.

A spatial breakdown of national climate action plans offers a compelling strategy. This allocates sector-specific energy demand and emissions targets and associated mitigation measures to individual municipalities, thereby facilitating regional plans. Such a spatial breakdown of national plans offers the following benefits:

* Corresponding author at: Institute of Energy and Climate Research, Techno-economic Systems Analysis (IEK-3), Forschungszentrum Jülich, Wilhelm-Johnen-Straße, Jülich, 52425, NRW, Germany.
E-mail address: s.patil@fz-juelich.de (S. Patil).

https://doi.org/10.1016/j.egyai.2024.100386
Received 8 January 2024; Received in revised form 15 May 2024; Accepted 6 June 2024
Available online 17 June 2024

List of Abbreviations

LULC	Land Use and Land Cover
GDP	Gross Domestic Product
ATM	Automated Teller Machine
GWR	Geographically Weighted Regression
CAR	Conditional Autoregressive

1. **Comparative analysis:** Spatially-allocated national plans allow for direct comparisons with local initiatives, revealing discrepancies that may affect the feasibility of national climate action plans. For instance, consider a scenario in which the spatially-allocated national plan sets out a 2030 reduction target for residential building energy demand at 2,900 MWh, whereas a local plan sets it at 1,500 MWh. Further examination of both plans might indicate that the renovation targets in the national plan are not achievable due to a limited number of eligible buildings at the municipal level. Such discrepancies are discussed by Muñoz et al. [4] in a case study focused on the city of Valencia, Spain.
2. **Resource allocation:** For municipalities lacking the capacity to develop their own plans, a spatially-allocated national plan provides a valuable starting point. Programs like the European City Facility [5] require municipalities to develop local climate action plans to qualify for funding. Thus, the spatial allocation not only guides less resourceful municipalities but also aids in securing financial support for implementing climate action.
3. **Strategic grouping:** Spatially-detailed plans facilitate the grouping of municipalities with similar sectoral energy demands and emissions profiles, enhancing the potential for collaboration. This collective approach fosters networking and the joint implementation of measures to meet established targets.

1.2. Motivation for the review

Spatial disaggregation methods have been applied in numerous studies to enhance the spatial resolution of data. These methods vary widely, ranging from simple allocations based on area proportions to advanced methods using machine learning and geostatistical models. The literature not only showcases a variety of spatial disaggregation methods but also reflects diversity within application domains, as well as in the original and target spatial resolutions used, the supporting data employed, and many other nuances. This diversity makes it challenging to gain an overview of the methods and filter the most relevant ones for the spatial disaggregation of climate action plans. In light of this, the objectives of this review are as follows:

1. To systematically review and classify spatial disaggregation methods, with a focus on emerging trends and recent advancements.
2. To evaluate and recommend spatial disaggregation methods specifically suited for national climate action plans.
3. To analyze and provide practical insights into the suitability of different disaggregation methods for various scenarios.

1.3. Scope of the review

Climate action plans typically encompass energy demand and greenhouse gas emission reduction targets across various sectors such as energy, buildings, transportation, industry, and agriculture. As a response to these targets, they include measures such as renewable energy capacity expansion, building renovation, the penetration of new electric vehicles, etc. The focus of the current review is on the spatial disaggregation of data relevant to these areas.

Spatial disaggregation is utilized in various fields, yet certain domains are considered beyond the scope of this review because their data is not directly involved in climate action planning. These excluded domains are:

1. Soil moisture [6].
2. Atmospheric data like daily surface temperature [7], precipitation [8], rainfall [9], wind speed and solar irradiance [10].
3. Increasing image sharpness [11].
4. Hazard risk downscaling such as fire [12] and heat stress [13].
5. Land Use and Land Cover (LULC) mapping [14].
6. Disease mapping [15].

This exclusion is justified because these techniques have been extensively reviewed in their respective fields. For example, Ekström et al. [16] reviewed methods focusing on the downscaling of global climate model simulations to a finer spatial resolution. Meanwhile, various soil mapping approaches are reviewed and evaluated in Vaysse and Lagacherie [17], and methods for disaggregating hydrological extremes can be found in Werner and Cannon [18].

Additionally, small area estimation methods [19] are omitted from this review, as they require the presence of data in target regions, even if it is sparse.

1.4. Methodology of the review

Fig. 1 displays the methodology of the review process employed in this study. As a first step, a heuristic search was performed to become familiar with the topic and to identify the keywords that are used in the literature. Next, a systematic review was conducted using Scopus,[1] based on the search query - *TITLE-ABS-KEY (((spatial AND disaggregation) OR (spatial AND downscaling)) AND (proxy OR ancillary OR covariate OR co-variate OR auxiliary OR surrogate OR synthetic OR simulated))*. The search query was formulated to include common keywords found in the literature. This query looked for these keywords in the title, abstract, and keywords of the corresponding publications. 1,302 papers were obtained using this search query, which were saved in a list on Scopus.

In the first phase, the abstracts of these publications were read. Based on the exclusion criteria previously discussed, the publications were then either discarded or moved to the next phase. In the second phase, the publications were read and various details (Fig. 1) were collected in a spreadsheet.

1.5. Structure of the review

The rest of the paper is structured as follows —Section 2 clarifies the terminologies related to spatial disaggregation, Section 3 reviews the spatial disaggregation methods that moved to phase 2 in the systematic literature review (Fig. 1), Section 4 discusses the reviewed methods in light of the defined objectives and, finally, Section 5 concludes paper.

2. Terminologies

Spatial disaggregation is a method used to improve the detail of data by increasing its spatial resolution. This technique is particularly useful when data is only available for larger areas and must be estimated for smaller, more specific regions. For instance, if we have emissions data available for larger regions like federal states, spatial disaggregation helps to estimate this for smaller regions such as municipalities within those states. In this context, the larger regions (federal states) are known as “source zones” and the smaller regions (municipalities) are called “target zones”. Importantly, each source zone is distinct and does not overlap with others, and each target zone belongs exclusively

[1] https://www.scopus.com/home.uri

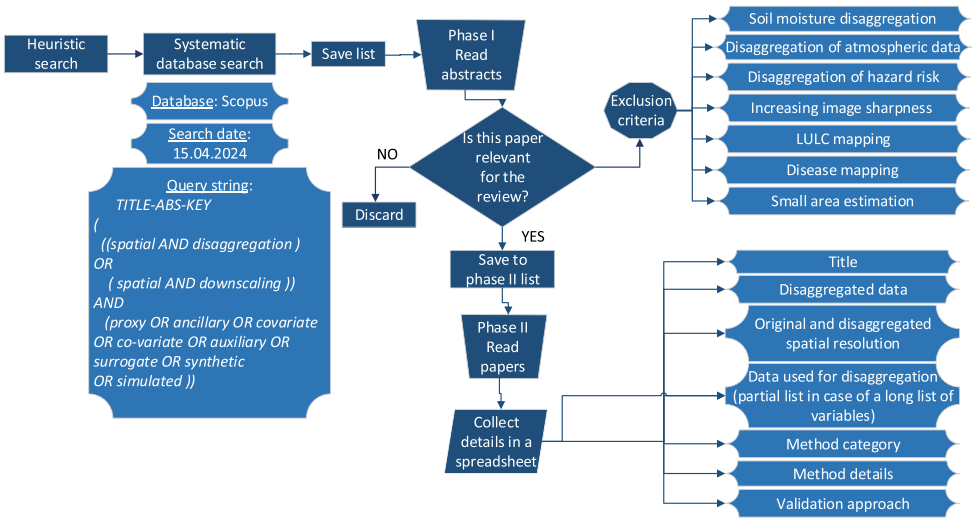

Fig. 1. Flow chart depicting the methodology of the review process.

to one source zone. Source and target zones can be defined as either administrative divisions or geographical grid cells.

The terms "spatial disaggregation" and "spatial downscaling" are commonly used interchangeably in the literature to describe methods for enhancing the spatial resolution of data. However, there is an important distinction between them. As defined by Monteiro et al. [20], "spatial disaggregation" refers to mass-preserving methods, which specifically ensure that the sum of disaggregated values in target zones is equal to the observed value in the parent source zone. This technique is crucial for variables like total emissions or population figures. In contrast, "spatial downscaling" is applied to variables that are not summed, such as temperature, precipitation, and soil moisture, etc. Despite the clarity of these definitions, the two terms are often treated synonymously in the literature.

Several other terms related to spatial disaggregation are also used interchangeably in the literature. This section provides definitions for these, lists the terminologies used in the literature, and selects a terminology to be used in this paper.

Table 1 summarizes this information.

Table 1
Terminologies related to spatial disaggregation that are used in the literature.

Definition	Terminologies	Terminology used in this paper
Region/grid set at the higher spatial level	Source zones [21]	Source zones
Region/grid set at the lower spatial level	Target zones [21]	Target zones
Distribution of data from a set of regions/grids at a higher spatial level to a set of regions/grids at a lower spatial level	Spatial disaggregation [22], spatial downscaling [4], spatialization [23], regionalization [24]	Spatial disaggregation
Data to be disaggregated, for example, emissions, population, energy demand, etc.	Target data [22]	Target data
The data present in the target zones that highly correlate with the target data and, therefore, can be employed in the spatial disaggregation of target data	Proxy data [25] , covariate data [22], co-variate data [26], auxiliary data [27], ancillary data [22], surrogate data [24]	Proxy data
The sum of disaggregated target data values in all the target zones, belonging to a source zone, is equal to the observed target data value in that source zone	Mass-preserving property [22], volume-preserving property [28], pycnophylactic property [29]	Mass-preserving property

3. Literature review

The spatial disaggregation methods introduced in the publications collected in the spreadsheet during phase 2 of the literature review are reviewed in this section. These methods can be categorized into six types (also shown in Fig. 2), based on the core approach or model employed. These are:

1. Areal weighting
2. Dasymetric mapping
3. Proxy data-based
4. Machine learning-based
5. Geostatistical model-based
6. Hybrid techniques

The following subsections address each of these categories.

3.1. Areal weighting

The areal weighting method assumes that the target value within each source zone is evenly distributed in that zone. Based on this assumption, the target data is distributed proportionally to the overlapping area of the target zone and source zone [30]. If a_{sz_I} is the area of the source zone I and a_{tz_i} is the area of a target zone i, then the target value in the target zone, tv_{tz_i} is given by:

$$tv_{tz_i} = tv_{sz_I} * \frac{a_{tz_i}}{a_{sz_I}} \quad (1)$$

As the assumption of even distribution of the target value within each source zone is seldom true, areal weighting has not used as a stand-alone method in recent works. However, in some studies it is combined with other methods as can be seen in the following subsections.

3.2. Dasymetric mapping

Dasymetric maps consist of a set of regions in which the variation of the target value is minimal and features a steep change at their

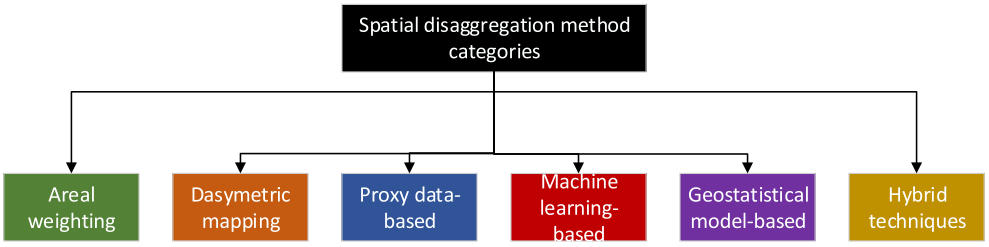

Fig. 2. Categorization of spatial disaggregation methods that were collected during the phase 2 of the systematic literature review.

boundaries [29]. Dasymetric mapping involves the creation of such a map in a top-down approach. The method is most often used to disaggregate population data.

Although it is a type of spatial disaggregation method, dasymetric mapping differs from the other methods in that the target zones are not defined beforehand. Instead, area-class maps such as LULCs are employed [31]. These maps divide a given area into several parts with a LULC type assigned to each. The target zones (also called dasymetric zones herein) are obtained by overlapping the area-class map and the source zones. The result is a set of target zones within each source zone, with each target zone having a single LULC type.

The population (target data) is then distributed to the target zones based on a predefined relationship between the LULC categories and population data. Accurately quantifying this relationship is the key challenge of dasymetric mapping. The following three techniques are described in Mennis [29]:

1. **Binary method**: Each LULC category is deemed either habitable or not. For example, the water and bare land LULC types are uninhabitable, and the rest are habitable. The population is set to 0 in uninhabitable areas, and is evenly distributed to the habitable LULC ones using areal weighting.
2. **Three-class method**: Each LULC category is assigned a percentage such that the percentages sum up to 100%. The total population is then distributed to these areas based on the percentage. It is noteworthy that although the method is named three-class, it can have more than three classes.
3. **Limiting variable method**: A maximum population density is assigned to each LULC category. The process begins with disaggregation using areal weighting. In a second step, if a target zone exceeds the assigned maximum density, the value is redistributed to its neighbors, provided they do not exceed their own assigned density. The second step is repeated until no target zone exceeds its density threshold.

Varying versions of dasymetric mapping are employed in various papers. Mennis and Hultgren [32] disaggregated census data —total population, Hispanic population, number of children and number of households based on an LULC map. Five LULC classes were considered —high-density residential, low-density residential, non-residential developed, vegetated, and water. The limiting variable method was employed. The maximum target variable density was then calculated based on sampling methods such as containment sampling. Here, for each LULC category, the source zones that completely overlapped with the category were sampled. For the set of samples, the maximum target value density for an LULC category was calculated using:

$$\frac{\sum_{i=1}^{n} tv_i}{\sum_{i=1}^{n} a_i} \tag{2}$$

where, tv_i and a_i are the target value and area of the sampled source zone i, respectively, and n is the number of sampled source zones. Different sampling techniques were discussed in the paper. The disaggregated data was then compared to census block-level data

Karunarathne and Lee [21] disaggregated the population in a hilly area using dasymetric mapping. Here, slope, altitude, and LULC maps were employed. The three-class method was employed and disaggregation was performed individually as follows:

1. *Based on slope*: The slope map consisted of six categories of increasing slope. An assumption was made that 80%, 15%, and 5% of the population, respectively, was in the first three categories. The other categories were deemed to be uninhabitable owing to their significant steepness.
2. *Based on altitude*: The altitude map consisted of eight categories of increasing altitude. An assumption was made that 75%, 17%, 5%, and 3% of the population can be found in the first four categories, respectively. The other categories were deemed to be uninhabitable owing to their significant elevations.
3. *Based on LULC*: Weights of 80%, 15%, 4%, and 1% were assigned to the LULC categories of 'home gardens', 'tea', 'rubber' and 'other plantation', respectively. The remaining categories were deemed to be uninhabitable.

Subsequently, the maps were overlapped. Corrections were made based on differences in assigned weights. For example, if an uninhabitable LULC category polygon, such as 'water', lay within the 80% weighted slope polygon, 80% was considered for the entire area of this slope polygon, except the water-covered one. Finally, the population was distributed to the target zones based on the corrected weights.

Apart from the LULC maps, the allocation of population to individual buildings is also seen in the literature. Maroko et al. [33] disaggregated population based on residential building cover. Residential area, building footprint, and height data was collected from various sources. This data was overlapped to calculate the total residential area per building. The population was then allocated to each building based on the share of the residential area. The disaggregated data was then compared to the results obtained using other variations of dasymetric mapping.

A similar approach can be found in Bajat et al. [34]. Population data was allocated to each building based on building height and soil sealing value. The results were compared to the census data available in some residential blocks. Wünsch et al. [35] employed dasymetric mapping to disaggregate residential building assets, i.e., total building costs including fixed assets such as heating and sanitation facilities. Binary dasymetric mapping was used where land use types —"residential areas" and "areas of mixed types" were deemed habitable. The assets were then distributed to these LULC polygons using areal weighting. The results were then benchmarked against available loss estimate data at the municipal level.

3.3. Proxy data-based

In proxy data-based methods, a proxy variable or a set of proxy variables is chosen based on the following two requirements:

1. It should be available in all target zones.
2. It should highly correlate with the target data. In other words, it should mimic the spatial distribution of it.

Table 2
Summary of papers that employed simple proxy data for disaggregation.

Paper	Target data	Proxy data	Source and target resolution
Moran et al. [36]	Source emissions	Several proxies depending on the emission source. For example, vehicle emissions based on number of fueling stations	Country to municipal level
Valencia et al. [37]	Sectoral emissions	Agriculture —agricultural land cover, Buildings —residential land cover weighted by population, Energy and industries - point source locations, Waste —municipal waste disposal sites, Transport —population, vehicle traffic flow and road density	55 km^2 to 500 m^2 grids
Saide et al. [38]	Traffic emissions	Population density, LULCs, road networks, traffic count data, road capacity, etc.	City to 2 km^2 grids
Ramacher et al. [39]	Source emissions	Several proxies depending on the emission source. For example, residential heating based on population density.	6 km^2 to 1 km^2 grids
Lam et al. [40]	Industrial emissions	Blue roof areas obtained from satellite images	A provincial area to 3 km^2 grids
Kuik et al. [41]	Traffic emissions and emissions from industry, residential combustion and product use	Traffic —traffic density, other —population	$\backsim 7km^2$ to $\backsim 1km^2$ grids
Wang et al. [42]	Black carbon emissions	Several proxies based on the source. For example, rural residential burning of coal, firewood, and crop residues —rural population	Country level to 0.1°grids

The target data is then distributed to the target zones based on the share of proxy data in each target zone.

The proxy data is chosen based on domain knowledge and the availability of data in the study area, at the target zone level. For example, Muñoz et al. [4] compared national energy and climate plans with local ones. For this, the authors chose the city of Valencia in Spain as a case study. The study identified the measures that were common in both national and local plans, such as energy efficiency improvements and the renovation of appliances in residential buildings. Simple proxies were used to disaggregate the national values to the city level. For example, if the national plans included that n number of buildings needed renovation, the number of old residential buildings in each city was used as a proxy to disaggregate this value. The result was then compared with local plans to identify potential misalignments. A similar approach of disaggregation based on simple proxies was found in other publications. These are summarized in Table 2.

The advantages of proxy data-based methods are:

1. They are simple and straightforward.
2. They are easily explainable based on the domain knowledge.
3. They can be employed even if the number of source zones are as few as one.

However, the main challenge associated with the proxy data-based methods is that the relevant proxy data to disaggregate a particular target dataset is not always readily available. The following three issues can arise:

1. A single proxy data is not sufficient to mimic the spatial distribution of the target data. Different datasets must be combined in order to synthesize appropriate proxy data.
2. The proxy data might be missing in some target regions. These values must be filled before they can be employed in the disaggregation.
3. The proxy data might not be available at the target zone level and might be available at an intermediate spatial resolution instead. In this case, the proxy data must first be disaggregated to the target zones.

Several publications have addressed these issues. Kuenen et al. [25] disaggregated sectoral greenhouse gas emissions from the country level to the grid one at a resolution of 0.05°* 0.1°. The proxy data was manually defined for each sector based on domain knowledge. Where feasible, the proxy data was synthesized by combining data from different sources and filling the gaps in the resulting data. For example, the road network locations found in Open Street Map [43] and the traffic volumes per vehicle type and roads categories found in Open Transport Map [44] were combined, resulting in a traffic intensity map. This map provided the traffic volume per vehicle type and road category in each target zone. Where traffic volume data was found to be missing, it was estimated by determining the relationship between the traffic volume per vehicle type, road category, and population density in the target zones where this data was available. This relationship was then used to impute the missing traffic volumes. The traffic intensity map was then used to disaggregate the emissions from the road transport sector. The authors note the impossibility of further delineation of the proxy data. For example, the dependency of road transport emissions on factors

Table 3
Summary of papers that employed synthesized proxy data for disaggregation.

Paper	Target data	Proxy data	Source and target resolution
Hernández et al. [46]	Traffic emissions	Road network overlapped with vehicle counts.	Metropolitan area to 1 km^2 grids
Chakraborty et al. [47]	Biomass potential	Gross primary production, calculated at the grid level using satellite data.	Districts to 1 km^2 grids
Mu et al. [48]	Traffic emissions	Road network with assigned weights for road segments (e.g., 0.05 for residential, 1 for freeway, etc.). The weighted length of road segments is used as a proxy.	11 km^2 to 100 m^2 grids
Gately et al. [49]	Traffic emissions	Vehicle fuel consumption, calculated based on vehicle miles traveled and emission factors across different vehicle classes and road categories.	Road level to 1 km^2 grids

such as vehicle speed, traffic flow and traffic jams was not considered, due to the unavailability of this information.

Another example of proxy data synthesis can be found in Zasina and Zawadzki [45] who aimed to disaggregate emissions due to domestic combustion. They argued that the population density alone could not account for this target data. A population that is supplied with heat via district heating infrastructure produces less emissions than one that relies on domestic combustion. Therefore, the population density in grids, in whose neighborhoods, a district heating infrastructure point is found, was reduced by 50%. This adjusted population data was then used as a proxy. A synthesis of the proxy data can be found in several other publications. These are summarized in Table 3.

Maes et al. [50] addressed the issue of the unavailability of proxies at target spatial resolutions. They aimed to disaggregate industrial emissions from country-level to 250 m^2 grids. They identified employment numbers as being relevant proxy data. However, employment data was not directly available for 250 m^2 grids. Therefore, the employment numbers were first disaggregated based on the LULC category, "industrial and commercial units" using dasymetric mapping. Subsequently, the disaggregated employment data was used as a proxy for industrial emissions disaggregation.

A similar two-step disaggregation can be seen in Alam et al. [51], who disaggregated traffic emissions from the national level to 0.5 km^2 grids. Vehicle fleets weighted by their corresponding mileages were used as proxies to first disaggregate national-level emissions to the county level. The county-level emissions were then disaggregated to the grid level using traffic volume as a proxy.

In some cases of emissions disaggregation, the proxy data was merely used to derive the emission factor values. For example, Righi et al. [52] aimed to disaggregate non-industrial NOx emissions from the provincial level to 100 m^2 grids. Based on domain knowledge, they identified population, building volume, and domestic gas consumption as the relevant proxies. They calculated the emission factor of NOx relative to each proxy variable at the provincial level. The emission factor was defined as the ratio between the total NOx emissions and the proxy value at the provincial scale. Then, the total emissions of each grid cell were calculated by multiplying the emissions factors by the proxy variable value in each cell.

Guevara et al. [53] developed an innovative approach to tackling the issues discussed earlier. For each emission source, they identified three proxies. For instance, for industrial combustion emissions, the proxies were industrial land use, urban land use, and urban population. The disaggregation began with the first proxy. If none of the grid cells intersecting the corresponding municipality contained information on this proxy, they moved to the next one. If all three proxies were inadequate, areal weighting was applied. This method aimed to address potential data gaps for specific proxies by allowing the selection of alternative ones.

In the literature, spatial disaggregation is applied not only to current data but also to future projections. The disaggregation of projections is especially relevant for national climate plans, which include future targets such as reducing emissions by a specific amount by 2030. Some studies emphasize the disaggregation of projections. For instance, Ran et al. [54] disaggregated population-related emission projections from the national level to 12 km^2 grids. In this work, population and land-use change projections, such as housing units, were used as proxies where available. Some land uses, such as road networks, were kept constant across future years due to the lack of available projections. The disaggregation was performed for each future year separately, based on the proxy values for that specific year.

3.4. Machine learning-based

Proxy data-based disaggregation methods often depend on domain expertise and the manual weighting of proxies, as exemplified by Mu et al. [48], in which different road categories were manually weighted. Machine learning models, on the other hand, offer a more nuanced approach by accurately combining various proxies with appropriate weights through learning the intricate relationships between target and proxy data.

Some studies have utilized machine learning techniques like random forests [55] or gradient boosting [56] to establish the relationship between target and proxy data. The method typically involves aggregating proxy data at the source zone level, training a machine learning model, and then using this model to predict the target data at the target zone level.

While this approach eliminates the manual effort needed to select and combine relevant proxies, it has a notable drawback in that it does not inherently preserve mass. In most cases, the predicted data at the target zone level are simply used as weights, and the target data is disaggregated based on these.

A machine learning-based approach, specifically using random forest, was seen in Wan et al. [57], who disaggregated population data from the tract level to the block group one. First, a set of proxies were synthesized based on land-impervious class cover. Examples of synthetic proxies include the total area of a particular impervious class, edge density, i.e., the ratio between the total length of a particular impervious class grid's edges and the total area, etc. A random forest model is then trained using synthesized proxies as predictors and population as a target. The model predictions at the block group level are used as weights to perform population disaggregation.

Patel et al. [58] followed a similar approach to disaggregate population data from the administrative level to 1 km^2 grids. They employed geotweet densities, land cover, night-time lights, temperature, elevation, etc. as proxies. In Stevens et al. [59], population data was disaggregated from the national level to $\backsim$100 m^2 grids, using several proxies including LULC types, slope, distance to roads, and temperature.

A significant challenge with machine learning methods is their dependency on large sample sizes for training. When the number of source zones is limited, these methods may not perform optimally. Arumugam et al. [60] tackled this issue in their efforts to disaggregate rice yields from district level down to 500m^2 grids. They proposed training a separate gradient boosting model for each district but recognized the issue with small sample sizes at the district level. To overcome this, they

first disaggregated the district-level rice yield to 5 km^2 grids, assigning the same district yield to each grid to effectively enlarge the sample size per district. This strategy allowed for the training of individual models in each district. It is important to note that as rice yield is typically measured in tons per hectare and is not an absolute measure, mass-preservation was not a concern in this methodology.

Kolluru et al. [61] disaggregated livestock population from the district level to 1 km^2 grid one. The number of districts (200) was deemed a small sample size for random forest model training. Therefore, an initial disaggregation from the district to sub-district level was carried out using areal weighting. A random forest model was then trained with 2000 sub-district-level data as the input. The model was then used to predict data at the 1 km^2 grid level.

Murakami et al. [62] disaggregated Gross Domestic Product (GDP) projections from the national level to 1/12 °grids. Urbanization potential, population projections, agricultural area, distance to major roads and oceans, etc. were all used as proxies. A gradient boosting model was trained for each future year based on the data in that year and some temporally-stationary proxies considered such as agricultural land, distance to the ocean, etc. The model was trained to minimize the mean squared error of the observed GDP and the aggregate of the predicted values.

Aside from random forest and gradient boosting algorithms, the literature has featured increasingly sophisticated machine learning techniques. For example, Monteiro et al. [22] disaggregated automated teller machines (ATM) withdrawals from the administrative level to 200 m^2 grids. They obtained an adequate sample size based on an initial disaggregation using population as proxy data. They proposed the co-training of two different machine learning models, namely random forest and a convolutional neural network [63]. The intention was to use the estimates from one model to fine-tune the estimates of the other, and vice versa.

Yang et al. [64] disaggregated PM2.5 concentration from 10 km^2 to 300 m^2 grids, using latitude, longitude, aerosol optical depth, elevation, and land cover, etc. The latitude and longitude variables were included to capture the spatial autocorrelation present in the target data. A cascade random forest [65] model was employed here. This model essentially consists of several random forest models. Predictions from earlier models were used as additional features for the subsequent ones, thereby iteratively improving the accuracy.

Zhao et al. [66] aimed to disaggregate population from the provincial level to 100 m^2 grids. Various points-of-interest such as catering, residential communities, financial services, educational centers, etc., were considered as relevant proxies. The frequent pattern growth algorithm [67], a type of association rule mining algorithm, was then used to obtain the spatial association between population hotspots and the different points-of-interest. The ones with strong associations were used as proxies. A random forest model was then employed using these as predictors.

Zhao et al. [68] disaggregated emissions from building energy consumption from the provincial level to 1 km^2 grids using GDP, population, temperature, heating degree days, and cooling degree days as proxies. They began by grouping the provinces into three groups based on climate. The disaggregation was then performed for each group separately. In this case as well, the problem with fewer observations arose. They employed a partial least squares regression [69] to disaggregate emissions from the provincial level to an intermediate prefecture one. This regression is well-suited when the number of predictors is more than the number of observations, and when multi-collinearity exists among predictors. The results at this intermediate level serve as input for a cubist regression model [70], which was used to predict the final data at the 1 km^2 grid level.

Georganos et al. [71] proposed a modification of the random forest model to make it spatial in nature, calling it a *geographical random forest*. The authors trained one random forest model per source zone, which included only the neighboring source zones as the input data and another global random forest model. For prediction, the source zone that was closest to a target zone was identified. The predictions from the model trained for this source zone and the global model were averaged to obtain the final prediction. The authors demonstrated the method by disaggregating population data from administrative units to 0.5 m^2 grids.

Verstraete [27] argued that although proxy data is useful for spatial disaggregation, it is not the only explanation for the spatial distribution of the target data. Further challenges, such as the mismatch in the time of collection of proxy and target data, introduce some uncertainties. To address these, a fuzzy inference system was proposed. This system uses a set of linguistic rules, in *IF-THEN* format, to map the relationship between input and output data [72]. This approach allows for the incorporation of the relationship between the proxy and target data but does not follow it too strictly. The proposed method was demonstrated by disaggregating artificial data from gridded source zones to gridded target zones. The results were then compared with those obtained using areal weighting, using the ideal solution as the benchmark.

3.5. Geostatistical model-based

Geostatistical models capture the spatial relationship between data points that has largely been ignored in the methods discussed thus far. These methods are well-suited for the disaggregation of data that exhibits spatial autocorrelation. Three geostatistical models can be found in the literature —Geographically Weighted Regression (GWR) [73], the Conditional Autoregressive (CAR) model [74], and variants of kriging [75].

GWR is a modified version of a simple linear regression model. It incorporates the spatial dependency of variables, i.e., the value in a region not only depends on its proxy data but also the neighborhood regions' proxy data. During fitting of a simple regression model, the sum of the squared differences between the predicted and observed data is minimized. However, in GWR, a weighting factor is imposed on each squared differences. This weight could be the geographical proximity of these regions such that the predictors closer to the region carry more penalties than others, which results in a set of local linear equations, with each specific to a source region. This equation is then used to predict values at the target regions.

Zhang et al. [76] employed GWR to disaggregate the CO_2 emissions of district heating systems from the city level to 3 km^2 grids. Night-time lights and temperature-humidity-wind index were used as proxy data. It is noteworthy that the predictions are merely used as weights and the data from the source region is subsequently distributed to the target regions based on the weights, thereby achieving mass-preservation.

CAR works on a similar principle as GWR. It is mathematically represented as:

$$Y_i = \beta + \sum_{j \in neighbors(i)} w_{ij}(Y_j - \mu) + \epsilon \tag{3}$$

where, Y_i is the target value in region i and w_{ij} the weight term representing the influence of a neighborhood region j on region i. These weights typically represent the geographical proximity of regions i and j. μ is the global mean of the data and provides a baseline around which the local variations are modeled. β and ϵ are the slope and error, respectively.

Within the context of spatial disaggregation, CAR is used in an iterative fashion. The process begins with determining initial estimates of target data at the target zone level using CAR. As the neighboring target values Y_j are unknown at this stage, simple disaggregation is performed either based on areal weighting or the proxy data-based method. The resulting estimates form the initial estimates.

Iteratively, the target values are recalculated based on the weights and neighboring target values. The new estimates are then compared to those from the previous iteration. If the changes between iterations fall below a certain threshold (indicating that further iterations do

not significantly alter the estimates), the process is deemed to have converged.

Charkovska et al. [77] used CAR to disaggregate the livestock population from the municipal level to 100 m^2 grids. They used population density and LULC, specifically "arable land", "pastures and agricultural areas" as proxy data to incorporate local effects (LE) on livestock population. Additionally, spatial effects (SE) were incorporated using CAR. The final model can be represented as:

$$Y_i = (LE_i) + (SE_i) \tag{4}$$

LE is determined by a simple linear regression model based on the proxy data and SE is determined based on CAR.

Kriging and its variants, namely, area-to-point kriging and cokriging, are geostatistical methods used for data interpolation. Kriging is a method that predicts values at unknown points based on the spatial correlation of known data points. Area-to-point kriging extends this concept to disaggregate data from source regions to target regions (typically small grids). Cokriging involves using proxy data (referred to as covariates in the context of kriging) to enhance the accuracy of disaggregated data. The basics of kriging are presented in Appendix. For an explanation of area-to-point kriging and cokriging, please refer to Kyriakidis [75] and the studies discussed below.

Kriging methods are not mass-preserving in nature. They are typically employed to disaggregate non-additive data such as GDP (expressed in purchasing power standards), average crop yield (expressed in tons per hectare), and population density. For instance, Triantakonstantis and Stathakis [78] employed cokriging to disaggregate GDP from administrative to municipal regions using night-time lights as a proxy.

Meanwhile, Brus et al. [79] employed area-to-point kriging in their attempt to disaggregate average crop yield from the provincial level to 1 hectare grids. Here, vegetation, precipitation, temperature, and soil data were used as proxies. A linear regression of proxy data was fit. Area-to-point kriging was used to disaggregate the residuals of the linear regression. Pittiglio et al. [80] disaggregated wild boar population density from the administrative level to 5 km^2 grids, following a similar approach, utilizing used temperature, precipitation, vegetation cover, and topography variables such a slope and elevation as proxies.

3.6. Hybrid techniques

Some studies have proposed an amalgamation of the different techniques discussed thus far. For example, Roni and Jia [81] disaggregated population from the ward level to 5 m^2 grids using building data, specifically building area and its associated building type such as residential, commercial, etc. as a proxy. Here, dasymetric mapping using GWR was employed to assign population values in each ward to individual buildings.

Jin et al. [82] disaggregated GDP from the provincial level to 1 km^2 grids using night-time lights, the vegetation index, population, etc. An amalgamation of the random forest model and area-to-area kriging is seen here. A random forest model was trained at the provincial level using proxy data as predictors. The residuals of the predictions were then distributed to target grids using area-to-area kriging.

Highfield et al. [83] disaggregated deer population density from county to LULC polygons using an amalgamation of dasymetric mapping and kriging. First, kriging was performed using the centroid of each county as the associated location for the value therein. The results were then used to obtain the deer density in each of the LULC polygons. Next, the density value was multiplied with a manual weighting assigned to each LULC category (1.2- shrubland, 1.0 - forests, 0.8 - grasslands, rest- 0). The final values were used as weights to allocate the deer population to each LULC polygon.

4. Discussion

From the reviewed studies, it is clear that the spatial disaggregation is applied to different target data —spatial disaggregation method category combinations. Fig. 3 shows these plotted against the year of publication. Drawing on the figure, the following observations can be made:

1. The proxy data-based methods are the most popular, followed by machine learning-based ones.
2. Emissions data is the most disaggregated target data, followed by population data
3. For the disaggregation of emissions data, the popular choice is proxy data-based methods. The reason for this is perhaps that the domain knowledge regarding emissions and the readily available proxy data makes it an obvious choice.
4. In recent years, machine learning-based methods have replaced dasymetric mapping as a popular choice for population data disaggregation, perhaps owing to the cumbersome manual steps involved in dasymetric mapping. However, some hybrid techniques involving dasymetric mapping have been seen in recent publications.

The national climate action plans consist of data related to various topics such as emissions and energy demand per sector, building assets, district heating, and renewable energy capacity, etc. Therefore, a single disaggregation technique might not be well-suited across all the target variables. The choice of a method depends on the following:

1. The target data.
2. The number of source zones.
3. The availability of proxy data in the target zones.
4. Domain knowledge.
5. Existence of spatial autocorrelation in the target data.

For a particular target dataset, if the user has the required domain knowledge to choose appropriate proxies and combine them with appropriate weights, the proxy data-based methods are well-suited. This further depends on the availability of the proxy data in the target zones. If the proxy data is missing in some target zones, data imputation techniques must be employed. If the proxy data is not readily available in the target zones, the user must: (a) synthesize the proxy data by fusing different datasets; or (b) first, disaggregate the proxy data that might be present at an intermediate spatial level.

If the user can identify a set of relevant proxies but, finds it challenging to combine them, machine learning-based methods are well-suited. The machine learning-based methods require a high number of source zones and associated target data to be effectively applied. Therefore, if the user is working with a single national climate action plan, the target data must first be disaggregated to an intermediate spatial level by employing proxy data-based methods. This can be an effective approach because as one ascends the administrative ladder, data availability grows, allowing for a plethora of proxy data options. However, the number of source zones even at the intermediate spatial resolution might not be sufficient to employ machine learning models. In such a case, techniques such as partial least squares regression must be employed, as it is known to handle fewer data points.

The choice of a particular machine learning approach depends on the target data, the required degree of accuracy, and available computational resources. A simple random forest or gradient boosting algorithm might be sufficient in most cases. Although complex models such as the co-training model or the cascade random forest one might render more accurate results, they are computationally-expensive.

The existence of spatial autocorrelation in target data must be considered during spatial disaggregation. Data such as emissions and district heating exhibit spatial autocorrelation at levels of fine spatial resolution such as municipalities. However, this might not be the case

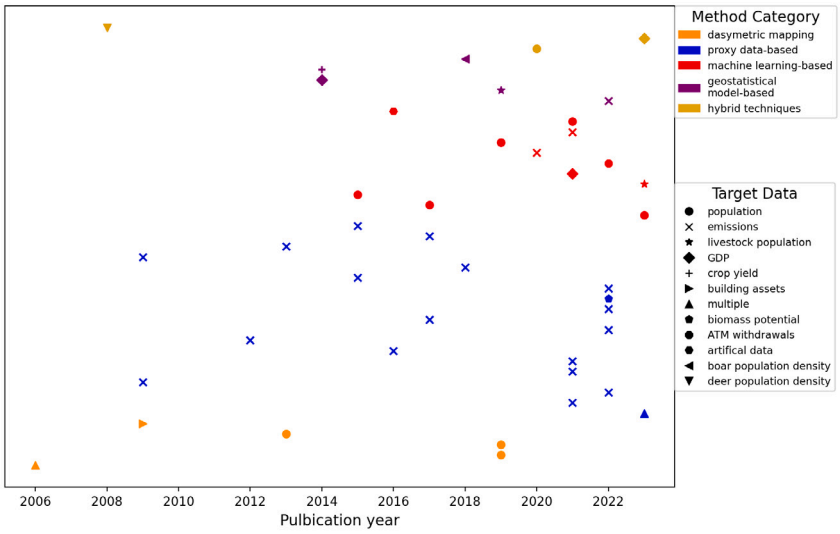

Fig. 3. A representation of the target data-spatial disaggregation method category combinations seen in the publication reviewed in this study. These combinations are plotted against the year of publication.

at coarse spatial resolutions such as federal states. One must therefore test for spatial autocorrelation in the data before choosing a disaggregation method. Measures such a Moran's index [84] can be employed for this purpose. If the target data is spatially-autocorrelated, geostatistical model-based methods must be employed. Alternatively, machine learning approaches like geographical random forest or incorporating the latitude and longitude coordinates of source zone centroids as additional predictors in a machine learning model can also be effective.

The national climate action plans not only include current data but also future targets. Therefore, when disaggregating future targets, the user must incorporate the regional changes that are expected in future years. For example, if the national climate plans include X GW photovoltaics capacity expansion by 2030, the photovoltaics potentials in each municipality must be considered as a proxy to disaggregate this information.

It is noteworthy that spatial disaggregation is not error-free. Nieves et al. [85] note some key issues that contribute to the inaccuracy of the results:

1. The data quality of the proxy data is often poor, thus affecting the quality of disaggregation.
2. The year of data collection of the proxy variable and the variable to be disaggregated might not be the same.
3. The relationship between a proxy variable and the target data might not hold true in all countries.
4. The relationship between a proxy variable and the target data might not hold true at all spatial resolutions.

The validation of spatially-disaggregated data is a challenge irrespective of the method applied. In previous works, the disaggregation results were validated against:

1. The target data available at either some or all target zones.
2. The target data available at an intermediate spatial resolution for either the entire area of interest or a few regions.

In terms of energy and climate plans, one could compare the disaggregated plans to the regional ones of certain cities, as seen in Muñoz et al. [4]. Additionally, reaching out to local authorities might be beneficial for refining any figures as needed.

5. Conclusions

In this study, we conducted a comprehensive review of the literature on techniques for disaggregating spatial data. Our goal was to provide a detailed overview of existing disaggregation methods. Furthermore, we aimed to recommend appropriate strategies for the spatial disaggregation of national climate action plan. It is essential to adapt national energy and climate strategies to local levels like municipalities to ensure effective implementation. The benefits of this approach include:

1. Engaging every region in mitigating climate change, rather than only those with the resources to create their own climate strategies.
2. Ensuring that local initiatives align with national climate targets.

We observed that no single spatial disaggregation method is universally applicable to national energy and climate strategies due to the diversity of target datasets and the various challenges associated with proxy data. This study narrows down the relevant methods to proxy data-based, machine learning-based and geostatistical-based methods. Our study identifies three relevant methodologies: proxy data-based, machine learning-based, and geostatistical-based approaches. Additionally, we provide guidelines for selecting the most suitable method based on factors like the presence of spatial autocorrelation in the data and the availability of relevant proxy data.

Future research will focus on applying these methods to disaggregate strategies and assessing the accuracy of the results. Despite the chosen method, validating these results remains challenging. Effective climate action planning requires close collaboration with local authorities to refine the outcomes.

CRediT authorship contribution statement

Shruthi Patil: Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Noah Pflugradt:** Writing – review & editing, Supervision, Resources, Conceptualization. **Jann M. Weinand:** Writing – review & editing, Supervision, Resources. **Detlef Stolten:** Supervision, Resources, Project administration. **Jürgen Kropp:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the language and readability of the work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work was developed as part of the project *LOCALISED —Localised decarbonization pathways for citizens, local administrations and businesses to inform for mitigation and adaptation action*. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101036458. We extend our sincere gratitude to the entire LOCALISED project team for their invaluable contributions. We also wish to acknowledge the financial support that has made this project possible. Special thanks are due to our colleagues at the Forschungszentrum Jülich for their diligent proofreading and insightful feedback, which have significantly enhanced the quality of our work.

Disclaimer

This work reflects the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

Appendix. Fundamentals of kriging

A semivariogram that explains the spatial relationship between the neighborhood target values is at the heart of kriging. One could plot the difference between pairs of target values against the distance between them and fit a theoretical model such as spherical, exponential, Gaussian one, etc. to obtain a semivariogram. Mathematically, this is represented as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \tag{A.1}$$

where, $\gamma(h)$ is the semivariance for a neighborhood distance h

$N(h)$ is the number of data pairs separated by distance h

$Z(x_i)$ is the target value in source region x_i

$Z(x_i + h)$ is the target value in another source region separated by a distance h from region x_i

This fitted semivariogram helps derive the covariance of any two values separated by a distance h, and is given by:

$$Cov(h) = C(0) - \gamma(h) \tag{A.2}$$

where, $C(0)$ is the covariance at zero distance and is often equal to the variance of the known target values.

The kriging estimator equations can be summarized as follows:

$$\sum_{j=1}^{n} \lambda_j Cov(Z_i, Z_j) = Cov(Z_i, Z_0), \forall i = 1, 2, \ldots, n \tag{A.3}$$

where, Z_i is the known target value in source region i

Z_0 is the unknown target value in a target region

λ_j are the weights assigned to the known target values

$Cov(Z_i, Z_j)$ is the covariance between known target values in source regions i and j

$Cov(Z_i, Z_0)$ is the covariance between known target value in source region i and unknown target value in a target region

n is the number of source regions

These equations are solved under the unbiasedness constraint, i.e.,

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{A.4}$$

Solving the above equations provides us with weights λ_j for each source region. These weights are used to obtain values at unknown target regions:

$$Z_0 = \sum_{j=1}^{n} \lambda_j Z_j \tag{A.5}$$

References

[1] Meeus L, Nouicer A. The EU clean energy package. 2018.
[2] Kona A, Bertoldi P, Monforti-Ferrario F, Rivas S, Dallemand JF. Covenant of mayors signatories leading the way towards 1.5 degree global warming pathway. Sustainable Cities Soc 2018;41:568–75.
[3] Reckien D, Salvia M, Heidrich O, Church JM, Pietrapertosa F, De Gregorio-Hurtado S, et al. How are cities planning to respond to climate change? Assessment of local climate plans from 885 cities in the EU-28. J Cleaner Prod 2018;191:207–19.
[4] Muñoz I, Hernández P, Pérez-Iribarren E, García-Gusano D, Arrizabalaga E. How can cities effectively contribute towards decarbonisation targets? a downscaling method to assess the alignment of local energy plans with national strategies. Energy Strategy Rev 2023;49:101137.
[5] European city facility. 2014, https://www.eucityfacility.eu/home.html, (Accessed: 2014-04-29).
[6] Choi M, Hur Y. A microwave-optical/infrared disaggregation for improving spatial representation of soil moisture using AMSR-E and MODIS products. Remote Sens Environ 2012;124:259–69.
[7] Wilby RL, Dawson CW, Barrow EM. SDSM—a decision support tool for the assessment of regional climate change impacts. Environ Model Softw 2002;17(2):145–57.
[8] Chen C, He Q, Li Y. Downscaling and merging multiple satellite precipitation products and gauge observations using random forest with the incorporation of spatial autocorrelation. J Hydrol 2024;130919.
[9] Koutsoyiannis D, Onof C, Wheater HS. Multivariate rainfall disaggregation at a fine timescale. Water Resour Res 2003;39(7).
[10] Chen S, Poll S, Hendricks Franssen H-J, Heinrichs H, Vereecken H, Goergen K. Convection-permitting ICON-LAM simulations for renewable energy potential estimates over southern africa. J Geophys Res: Atmos 2024;129(6). e2023JD039569.
[11] Pardo-Igúzquiza E, Chica-Olmo M, Atkinson PM. Downscaling cokriging for image sharpening. Remote Sens Environ 2006;102(1–2):86–98.
[12] Carvalho AC, Carvalho A, Martins H, Marques C, Rocha A, Borrego C, et al. Fire weather risk assessment under climate change using a dynamical downscaling approach. Environ Model Softw 2011;26(9):1123–33.
[13] Lindberg F, Thorsson S, Rayner D, Lau K. The impact of urban planning strategies on heat stress in a climate-change perspective. Sustainable Cities Soc 2016;25:1–12.
[14] Li X, Du Y, Ling F, Wu S, Feng Q. Using a sub-pixel mapping model to improve the accuracy of landscape pattern indices. Ecol Indic 2011;11(5):1160–70.
[15] Weiss DJ, Lucas TC, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of plasmodium falciparum, 2000–17: a spatial and temporal modelling study. Lancet 2019;394(10195):322–31.
[16] Ekström M, Grose MR, Whetton PH. An appraisal of downscaling methods used in climate change research. Wiley Interdiscip Rev Clim Change 2015;6(3):301–19.
[17] Vaysse K, Lagacherie P. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional 2015;4:20–30.
[18] Werner AT, Cannon AJ. Hydrologic extremes–an intercomparison of multiple gridded statistical downscaling methods. Hydrol Earth Syst Sci 2016;20(4):1483–508.
[19] Orozco EV, Rivera JV, Mata GA. Labor figures for Mexico's municipalities: Small area estimation. Stat J IAOS 2021;37(2):629–40.
[20] Monteiro J, Martins B, Pires JM. A hybrid approach for the spatial disaggregation of socio-economic indicators. Int J Data Sci Anal 2018;5(2–3):189–211.
[21] Karunarathne A, Lee G. Estimating hilly areas population using a dasymetric mapping approach: A case of Sri Lanka's highest mountain range. ISPRS Int J Geo-Inf 2019;8(4):166.
[22] Monteiro J, Martins B, Costa M, Pires JM. A co-training approach for spatial data disaggregation. In: Proceedings of the 30th international conference on advances in geographic information systems. 2022, p. 1–10.
[23] Mei Y, Gui Z, Wu J, Peng D, Li R, Wu H, et al. Population spatialization with pixel-level attribute grading by considering scale mismatch issue in regression modeling. Geo-spat Inf Sci 2022;25(3):365–82.
[24] Wu J, Li Y, Li N, Shi P. Development of an asset value map for disaster risk assessment in China by spatial disaggregation using ancillary remote sensing data. Risk Anal 2018;38(1):17–30.
[25] Kuenen J, Dellaert S, Visschedijk A, Jalkanen J-P, Super I, Denier van der Gon H. CAMS-REG-v4: A state-of-the-art high-resolution European emission inventory for air quality modelling. Earth Syst Sci Data 2022;14(2):491–515.

[26] Horabik-Pyzel J, Nahorski Z. Uncertainty of spatial disaggregation procedures: Conditional autoregressive versus geostatistical models. In: 2016 federated conference on computer science and information systems (fedCSIS). IEEE; 2016, p. 449–57.

[27] Verstraete J. The spatial disaggregation problem: simulating reasoning using a fuzzy inference system. IEEE Trans Fuzzy Syst 2016;25(3):627–41.

[28] Qiu Y, Zhao X, Fan D, Li S, Zhao Y. Disaggregating population data for assessing progress of SDGs: methods and applications. Int J Digit Earth 2022;15(1):2–29.

[29] Mennis J. Dasymetric mapping for estimating population in small areas. Geogr Compass 2009;3(2):727–45.

[30] Goodchild MF, Lam NS-N. Areal interpolation: A variant of the traditional spatial problem. Geo-processing 1980;1(3):297–312.

[31] Büttner G. CORINE land cover and land cover change products. In: Land use and land cover mapping in europe: practices & trends. Springer; 2014, p. 55–74.

[32] Mennis J, Hultgren T. Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci 2006;33(3):179–94.

[33] Maroko A, Maantay J, Pérez Machado RP, Barrozo LV. Improving population mapping and exposure assessment: three-dimensional dasymetric disaggregation in New York City and São Paulo, Brazil. Pap Appl Geogr 2019;5(1–2):45–57.

[34] Bajat B, Krunić N, Samardžić-Petrović M, Kilibarda M. Dasymetric modelling of population dynamics in urban areas [Dasimetrično modeliranje dinamike prebivalstva na urbanih območjih]. Geodetski vestnik 2013;57(4):777–92.

[35] Wünsch A, Herrmann U, Kreibich H, Thieken AH. The role of disaggregation of asset values in flood loss estimation: a comparison of different modeling approaches at the Mulde River, Germany. Environ Manag 2009;44:524–41.

[36] Moran D, Pichler P-P, Zheng H, Muri H, Klenner J, Kramel D, et al. Estimating CO 2 emissions for 108,000 European cities. Earth Syst Sci Data Discussions 2021;2021:1–23.

[37] Valencia VH, Levin G, Ketzel M. Downscaling global anthropogenic emissions for high-resolution urban air quality studies. Atmos Pollut Res 2022;13(10):101516.

[38] Saide P, Zah R, Osses M, de Eicker MO. Spatial disaggregation of traffic emission inventories in large cities using simplified top–down methods. Atmos Environ 2009;43(32):4914–23.

[39] Ramacher MOP, Kakouri A, Speyer O, Feldner J, Karl M, Timmermans R, et al. The UrbEm hybrid method to derive high-resolution emissions for city-scale air quality modeling. Atmosphere 2021;12(11):1404.

[40] Lam YF, Cheung CC, Zhang X, Fu JS, Fung JCH. Development of a new emission reallocation method for industrial sources in China. Atmos Chem Phys 2021;21(17):12895–908.

[41] Kuik F, Lauer A, Churkina G, Denier van der Gon HA, Fenner D, Mar KA, et al. Air quality modelling in the berlin–brandenburg region using WRF-chem v3. 7.1: sensitivity to resolution of model grid and input data. Geosci Model Dev 2016;9(12):4339–63.

[42] Wang R, Tao S, Wang W, Liu J, Shen H, Shen G, et al. Black carbon emissions in China from 1949 to 2050. Environ Sci Technol 2012;46(14):7595–603.

[43] OpenStreetMap contributors. 2017, Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org.

[44] Jedlička K, Hájek P, Cada V, Martolos J, Šťastný J, Beran D, et al. Open transport map — Routable OpenStreetMap. In: 2016 IST-africa week conference. 2016, p. 1–11. http://dx.doi.org/10.1109/ISTAFRICA.2016.7530657.

[45] Zasina D, Zawadzki J. Spatial surrogate for domestic combustion's air emissions: A case study from Silesian Metropolis, Poland. J Air Waste Manage Assoc 2017;67(9):1012–9.

[46] Hernández KS, Henao JJ, Rendón AM. Dispersion simulations in an andean city: Role of continuous traffic data in the spatio-temporal distribution of traffic emissions. Atmos Pollut Res 2022;13(3):101361.

[47] Chakraborty A, Biswal A, Pandey V, Shadab S, Kalyandeep K, Murthy C, et al. Developing a spatial information system of biomass potential from crop residues over India: A decision support for planning and establishment of biofuel/biomass power plant. Renew Sustain Energy Rev 2022;165:112575.

[48] Mu Q, Denby BR, Wærsted EG, Fagerli H. Downscaling of air pollutants in Europe using uEMEP_v6. Geosci Model Dev 2022;15(2):449–65.

[49] Gately CK, Hutyra LR, Sue Wing I. Cities, traffic, and CO2: A multidecadal assessment of trends, drivers, and scaling relationships. Proc Natl Acad Sci 2015;112(16):4999–5004.

[50] Maes J, Vliegen J, Van de Vel K, Janssen S, Deutsch F, De Ridder K, et al. Spatial surrogates for the disaggregation of CORINAIR emission inventories. Atmos Environ 2009;43(6):1246–54.

[51] Alam MS, Duffy P, Hyde B, McNabola A. Downscaling national road transport emission to street level: A case study in Dublin, Ireland. J Cleaner Prod 2018;183:797–809.

[52] Righi S, Farina F, Marinello S, Andretta M, Lucialli P, Pollini E. Development and evaluation of emission disaggregation models for the spatial distribution of non-industrial combustion atmospheric pollutants. Atmos Environ 2013;79:85–92.

[53] Guevara M, Tena C, Soret A, Serradell K, Guzmán D, Retama A, et al. An emission processing system for air quality modelling in the Mexico city metropolitan area: Evaluation and comparison of the MOBILE6. 2-Mexico and MOVES-Mexico traffic emissions. Sci Total Environ 2017;584:882–900.

[54] Ran L, Loughlin D, Yang D, Adelman Z, Baek B, Nolte C. ESP v2. 0: enhanced method for exploring emission impacts of future scenarios in the United States–addressing spatial allocation. Geosci Model Dev Discuss 2015;8(1):263–300.

[55] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[56] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;1189–232.

[57] Wan H, Yoon J, Srikrishnan V, Daniel B, Judi D. Landscape metrics regularly outperform other traditionally-used ancillary datasets in dasymetric mapping of population. Comput Environ Urban Syst 2023;99:101899.

[58] Patel NN, Stevens FR, Huang Z, Gaughan AE, Elyazar I, Tatem AJ. Improving large area population mapping using geotweet densities. Trans GIS 2017;21(2):317–31.

[59] Stevens FR, Gaughan AE, Linard C, Tatem AJ. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS One 2015;10(2):e0107042.

[60] Arumugam P, Chemura A, Schauberger B, Gornott C. Remote sensing based yield estimation of rice (Oryza sativa l.) using gradient boosted regression in India. Remote Sens 2021;13(12):2379.

[61] Kolluru V, John R, Saraf S, Chen J, Hankerson B, Robinson S, et al. Gridded livestock density database and spatial trends for Kazakhstan. Sci Data 2023;10(1):839.

[62] Murakami D, Yoshida T, Yamagata Y. Gridded GDP projections compatible with the five SSPs (shared socioeconomic pathways). Front Built Environ 2021;7:760306.

[63] LeCun Y, Bengio Y, Hinton G. Deep learning. nature 2015;521(7553):436–44.

[64] Yang Q, Yuan Q, Li T, Yue L. Mapping PM2. 5 concentration at high resolution using a cascade random forest based downscaling model: Evaluation and application. J Cleaner Prod 2020;277:123887.

[65] Zhou Z-H, Feng J. Deep forest. Natl Sci Rev 2019;6(1):74–86.

[66] Zhao Y, Li Q, Zhang Y, Du X. Improving the accuracy of fine-grained population mapping using population-sensitive POIs. Remote Sens 2019;11(21):2502.

[67] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM Sigmod Rec 2000;29(2):1–12.

[68] Zhao Z, Yang X, Yan H, Huang Y, Zhang G, Lin T, et al. Downscaling building energy consumption carbon emissions by machine learning. Remote Sens 2021;13(21):4346.

[69] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr Intell Laboratory Syst 2001;58(2):109–30.

[70] Quinlan J. 5th Australian joint conference on artificial intelligence. World Scientific; 1992.

[71] Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, et al. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int 2021;36(2):121–36.

[72] Dernoncourt F. Introduction to fuzzy logic. Massachusetts Institute of Technology 2013;21:50–6.

[73] Brunsdon C, Fotheringham AS, Charlton ME. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr Anal 1996;28(4):281–98.

[74] Besag J. Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B Stat Methodol 1974;36(2):192–225.

[75] Kyriakidis PC. A geostatistical framework for area-to-point spatial interpolation. Geogr Anal 2004;36(3):259–89.

[76] Zhang W, Wang J, Xu Y, Wang C, Streets DG. Analyzing the spatio-temporal variation of the CO2 emissions from district heating systems with "Coal-to-Gas" transition: Evidence from GTWR model and satellite data in China. Sci Total Environ 2022;803:150083.

[77] Charkovska N, Horabik-Pyzel J, Bun R, Danylo O, Nahorski Z, Jonas M, et al. High-resolution spatial distribution and associated uncertainties of greenhouse gas emissions from the agricultural sector. Mitig Adapt Strateg Glob Change 2019;24:881–905.

[78] Triantakonstantis D, Stathakis D. Cokriging areal interpolation for estimating economic activity using night-time light satellite data. In: Computational science and its applications–ICCSA 2014: 14th international conference, Guimarães, Portugal, June 30–July 3, 2014, proceedings, part IV 14. Springer; 2014, p. 243–52.

[79] Brus D, Boogaard H, Ceccarelli T, Orton T, Traore S, Zhang M. Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging. Eur J Agron 2018;97:48–59.

[80] Pittiglio C, Khomenko S, Beltran-Alcrudo D. Wild boar mapping using population-density statistics: From polygons to high resolution raster maps. PLoS One 2018;13(5):e0193295.

[81] Roni R, Jia P. An optimal population modeling approach using geographically weighted regression based on high-resolution remote sensing data: A case study in Dhaka City, Bangladesh. Remote Sens 2020;12(7):1184.
[82] Jin Y, Ge Y, Fan H, Li Z, Liu Y, Jia Y. Mapping gross domestic product distribution at 1 km resolution across thailand using the Random Forest Area-to-area regression kriging model. ISPRS Int J Geo-Inf 2023;12(12):481.
[83] Highfield L, Ward M, Laffan S. Representation of animal distributions in space: how geostatistical estimates impact simulation modeling of foot-and-mouth disease spread. Veterinary Res 2008;39(2):1–14.
[84] Moran PA. Notes on continuous stochastic phenomena. Biometrika 1950;37(1/2):17–23.
[85] Nieves JJ, Stevens FR, Gaughan AE, Linard C, Sorichetta A, Hornby G, et al. Examining the correlates and drivers of human population distributions across low-and middle-income countries. J R Soc Interface 2017;14(137):20170401.