



OPEN

DATA DESCRIPTOR

All urban areas' energy use data across 640 districts in India for the year 2011

Kangkang Tong ¹, Ajay Singh Nagpure² & Anu Ramaswami ^{1,2,3,4} 

India is the third-largest contributor to global energy-use and anthropogenic carbon emissions. India's urban energy transitions are critical to meet its climate goals due to the country's rapid urbanization. However, no baseline urban energy-use dataset covers all Indian urban districts in ways that align with national totals and integrate social-economic-infrastructure attributes to inform such transitions. This paper develops a novel bottom-up plus top-down approach, comprehensively integrating multiple field surveys and utilizing machine learning, to model All Urban areas' Energy-use (AllUrE) across all 640 districts in India, merged with social-economic-infrastructure data. Energy use estimates in this AllUrE-India dataset are evaluated by comparing with reported energy-use at three scales: nation-wide, state-wide, and city-level. Spatially granular AllUrE data aggregated nationally show good agreement with national totals (<2% difference). The goodness-of-fit ranged from 0.78–0.95 for comparison with state-level totals, and 0.90–0.99 with city-level data for different sectors. The relatively strong alignment at all three spatial scales demonstrates the value of AllUrE-India data for modelling urban energy transitions consistent with national energy and climate goals.

Background & Summary

India is the third-largest energy user in the world and contributed ~7% of global anthropogenic greenhouse gas (GHG) emissions in 2018¹. Approximately 416 million people (a 90% increase) will be added to the Indian urban population from 2018 to 2050². This rapid urbanization will drive up India's energy use and GHG emissions as almost all urban centers expand, and towns become more urban. GHG emissions per capita in India were 1.94 metric tonnes in 2018 (the 131st in the world)¹, with an annual 6% increase over the past decade³. India's urbanization thus represents an important opportunity to slow energy demand and associated carbon emissions if low-carbon city planning is undertaken as the country urbanizes. Low-carbon urbanization has implications for meeting both India's and global commitments to the Paris Agreement on climate change⁴. To achieve low-carbon urbanization, we first need data on current energy use and carbon footprints in cities to establish a baseline for tracking progress. While the Indian government has conducted district-level surveys on various social-economic-infrastructure attributes, there is no baseline energy-use dataset for all urban areas in ways that: (a) capture key local energy-use features, (b) align with national totals, and (c) integrate social-infrastructure-urban form variables⁵. Such baseline data are essential for India and its cities to develop and evaluate low-carbon policies that align from local to state and national scales.

India has pledged to reduce GHG emissions per gross domestic output by 33%–35% by 2030 (based on 2005 levels) in its Intended Nationally Determined Contribution (INDC) under the Paris Agreement⁶. India's INDC emphasizes transitions to clean and highly-efficient energy systems, while no numeric targets are detailed for controlling urban emissions⁶. Part of the reason is the lack of relevant data with urban-to-nation linkage. Moreover, the method to quantify urban carbon emissions is different from the national carbon inventory. Typically, countries report direct emissions from fuel combustion using a source-based accounting method, which is the foundation for spatially gridded datasets, such as EDGAR⁷ and ODIAC⁸. Both inventories focus on sources of anthropogenic GHG (e.g., fossil fuel combustion in power plants). Source-based accounting for cities can miss mitigation opportunities because urban areas provide unique avenues for deep-decarbonization through demand

¹Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA. ²Hubert H. Humphrey School of Public Affairs, University of Minnesota Twin Cities, Minneapolis, MN, USA. ³M.S. Chadha Center for Global India, Princeton University, Princeton, NJ, USA. ⁴High Meadows Environmental Institute, Princeton University, Princeton, NJ, USA.  e-mail: anu.ramaswami@princeton.edu

reduction when taking advantage of spatial efficiency (i.e., the co-location of various activities)^{9–11}. Indeed, urban GHG accounting should focus on energy use by multiple users (e.g., households, manufacturing, and businesses), tracking energy use activity with trans-boundary supply of energy (e.g., power plants, refineries, etc.)¹². The inclusion of multiple users in an urban dataset is essential to inform cross-sectoral actions (e.g., energy and material exchange across sectors), in addition to single sectoral policies (e.g., efficiency improvement and electrification). A study demonstrated Chinese urban areas supporting energy and material exchange across sectors can additionally reduce 40% of total carbon mitigated by single sectoral efficiency policies¹³. Therefore, urban areas can contribute significantly to national deep decarbonization goals. To do so effectively, both cities and nations need datasets addressing energy use across all urban areas to scale up the impact of urban actions to the national level.

Such local-to-national aligned energy use datasets covering all urban areas have been developed in the U.S.¹⁴ and China^{13,15}. More recently, efforts have been made to develop a fine-spatial scale socio-demographic and economic database in India¹⁶, but this dataset does not include energy use. Previous studies on energy use in India either focused on individual cities¹⁷ or covered individual end-use sectors across multiple cities^{18–20}. These data can be used to inform energy transitions for individual or a subset of cities, or one end-use sector. However, these datasets do not address energy use in multiple sectors across all urban areas, and do not address whether the sum of modelled energy use aligns with national total energy use. Moreover, existing data are fragmented by sectors, and social-economic employment data are not often integrated with infrastructure attributes.

Our paper addresses this gap by developing and implementing a novel bottom-up plus top-down approach to quantify energy use in four end-use sectors. This bottom-up plus top-down approach modifies a method recently used to develop data for Chinese cities^{13,15}. We integrate the available social-demographics, economic, infrastructural, and urban form variables based on the social-ecological-infrastructural-urban systems framework^{5,21} to develop the *All Urban areas' Energy-use* (AllUrE) dataset covering all 640 Indian districts for the year 2011. This novel dataset development method can be translated to other developing countries with relatively sparse data to develop their all-cities' energy use and carbon footprinting databases for supporting low-carbon policies and tracking progress. This dataset has many applications: exploring carbon emissions patterns across cities of different types, evaluating the impact of urbanization levels on emissions in different sectors, and quantifying collective impact of urban decarbonization strategies on national INDCs. These specific applications are beyond the scope of this paper, which focuses on describing the dataset development.

Methods

Method overview. The urban units in our dataset correspond to urban portions of all 640 districts specified by the Government of India, for which this AllUrE dataset provides baseline energy use consistent with the national total for the first time. We chose the year 2011 because it provided the latest publicly available population census, with detailed demographics, employment, housing conditions, residential energy use structure, and water sanitation infrastructure data covering both rural and urban areas in each district.

In the Census of India 2011, 27 states and 8 union territories representing the whole of India are divided into 640 districts. Within a district, there are towns and villages. The Census of India 2011 uses the phrase “urban area” to refer to statutory towns or census towns. Statutory towns are administrative units defined as having urban status, including Municipal Corporations, Municipalities, Cantonment Board, Notified Town Area Committee, Town Panchayat, etc.²². Census towns are defined by three criteria, which are (a) over 5,000 persons, (b) 75% of male workers engaging in non-agricultural activities, and (c) population density higher than 400 persons/sq. km²². The sum of statutory and census towns in a district is the urban area within the district. Remaining geographic areas, excluding statutory and census towns, are classified as rural areas. According to the Census of India 2011, about 31% of Indian total population lives in urban areas. Of the 640 districts, three districts have no defined urban areas (meaning no statutory or census towns are in these districts). Twenty-four districts have over 80% of their population living in defined urban areas, and nine of these 24 districts have 100% urban population (meaning these districts are the sum of statutory or census towns). While our focus was on the energy use in urban areas, the rural portion was also estimated to ensure the addition of energy use across all districts aligned with national totals.

The overarching process for estimating All Urban Areas' Energy-use (AllUrE) across 640 districts of India is shown in Fig. 1. Energy end-use sectors included households, industrial manufacturing, transportation, and commercial and agricultural sectors. End-use energy included coal, petroleum fuel (e.g., liquefied petroleum gas (LPG), kerosene, gasoline, and diesel), electricity, and firewood (only in the residential sector). The bottom-up approach estimated energy use in households and industrial manufacturing (coal and electricity use only) based on three data sources (i.e., the Annual Survey of Industry 2012²³, the National Sample Survey 68th Round 2011/2012²⁴, and the Census of India 2011²⁵) that reflected local social, economic, and demographic attributes. Fuel use for on-road transportation was quantified based on the number of vehicles and their traveling distance (see Fig. 1). For activities where surveys were not available, including the commercial sector, agricultural activity, and petroleum fuel use in industrial manufacturing, we applied a top-down approach to downscale national energy use reported by the Government of India to the district level based on the number of workers in these sectors. The main data sources for the top-down estimate of energy use by sectors and fuel types included the National Energy Statistical Yearbook²⁶ and Indian Petroleum and Natural Gas Statistics²⁷. The modelled energy use data by end-use sectors in all urban areas were merged with publicly available social-demographics, economic, infrastructural, and urban form variables, which were primarily collected from the Census of India 2011 (see Data Records section for details).

The quality of modelled energy use in AllUrE-India was evaluated at three spatial aggregation scales (i.e., national, state, and city scales) (see Fig. 2). First, the sum of energy use in AllUrE in urban and rural areas across the 640 districts nation-wide was compared against national total energy use reported in Indian statistical yearbooks. The alignment of estimated data with national totals is a key criterion, because a high alignment ensures that data can be used to quantify the carbon mitigation potential of collective city-level actions on national outcomes. Second, the sum of energy use in urban and rural areas across the 640 districts was compared against the available

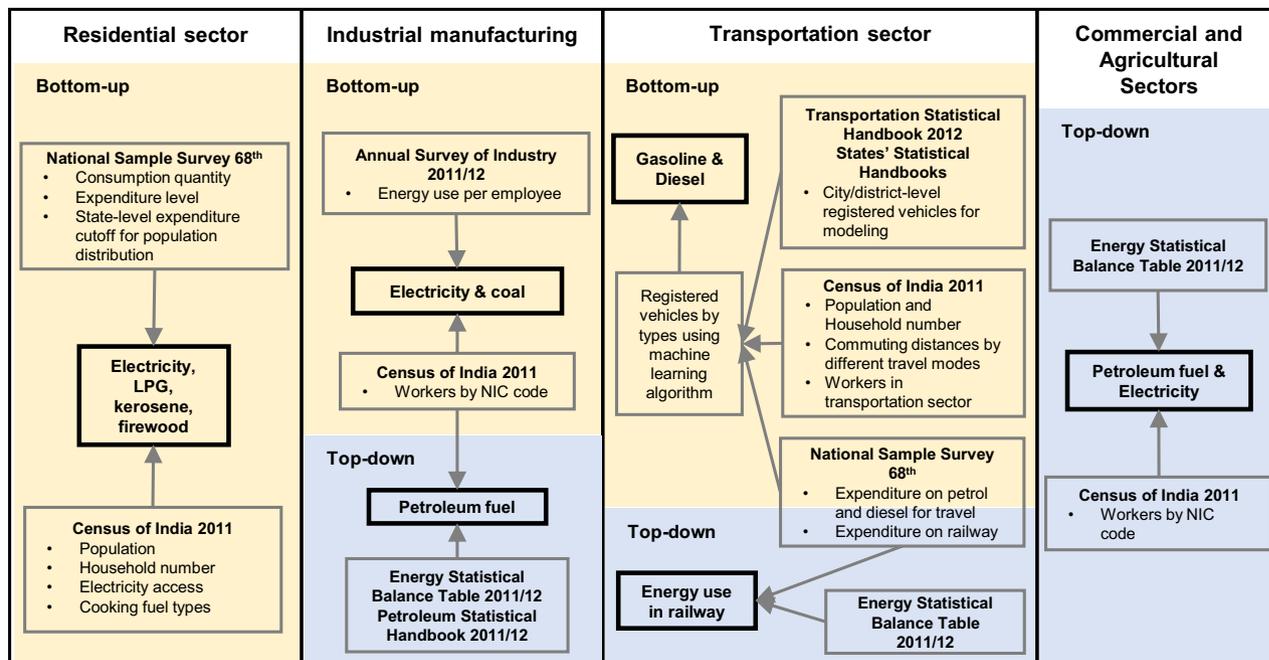


Fig. 1 Diagram showing the data sources to estimate All Urban areas' Energy-use across 640 districts' urban areas in India by end-use sectors (i.e., residential, industrial, transportation, and commercial & agriculture sector).

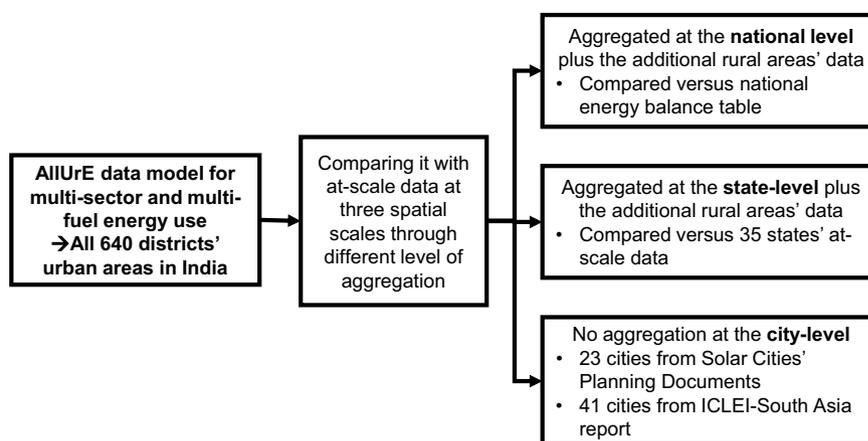


Fig. 2 Schematic illustration of technical validation of energy use data in AllUrE-India at three spatial aggregation scales: (1) national-level aggregation; (2) state-level aggregation; (3) un-aggregated data at the city-level with aligning of appropriate boundary.

at-scale data for all states (27 states and 8 union territories). Lastly, the estimated urban energy use data was compared against the city-level data collected from literature (ICLEI-South Asia report provided 41 cities energy use for 2007/2008¹⁸) and urban planning documents (e.g., solar cities' planning documents yielding 23 city-level comparison points). This city-level comparison reveals the effectiveness of the bottom-up approach in estimating cities' energy use.

Residential sector. Energy use in the residential sector includes electricity, LPG, kerosene, and firewood used for cooking, lighting, and heating/cooling. Gasoline and diesel used for personal travel are included in the transportation sector and not included in the residential sector.

Each urban area's residential energy use was calculated as the sum of average energy use per person within five quintiles (i.e., bottom 20%, 20–40%, 40–60%, 60–80%, and 80–100%) obtained from the National Sample Survey (NSS) multiplied by the population in these quintiles (see Eqs. 1 and 2 for electricity and LPG as examples). We differentiated between these quintile strata because they can better represent the variation of residential energy use per household²⁸. For example, Nagpure *et al.* (2018) showed that per capita electricity use in the highest 20% income stratum is 5.5 times that of the lowest income stratum²⁹. The effectiveness of this approach is evaluated by comparing against utility reported data at the city level (see Technical Validation).

The NSS collected 101,651 households' data from 7,469 villages and 5,268 urban blocks across India in the 68th Round (2011/12)²⁴. Each surveyed household reported the amount of monthly energy use in physical quantities (e.g., kilogram, kWh, or litre). The number of households surveyed in each district was used to calculate the error in the mean by monthly expenditure strata. Both total population and the access to energy data for each district's urban area were extracted from the Census of India 2011³⁰.

The average energy use per person (e.g., $EPC_{ele,i,j}$ and $EPC_{LPG,i,j}$ for electricity and LPG use per person, respectively) in a district's urban area was calculated in five population strata, divided based on households' monthly expenditure levels. Each stratum accounted for 20% of the urban district's population and ranked from the lowest to the highest expenditure levels. The expenditure cut-offs for the population strata were summarized in the NSS Household Expenditure Report³¹ for each state's urban and rural areas. When a district's urban area had less than ten samples within an expenditure range, household samples from the state's urban area with the same expenditure level were used in the estimation. The physical quantity of energy use per person in the wealthiest 20% of the population was capped at the third quantile value of this sample to remove outliers.

Total electricity use in each district's urban area was calculated based on Eq. (1):

$$RES_{ele,j} = \sum_i POP_{i,j} * a_{ele,i,j} * EPC_{ele,i,j} \quad (1)$$

in which $POP_{i,j}$ is the population in each j^{th} district's urban area in the i^{th} population stratum (e.g., bottom 20%, 20~40% of the total population) based on monthly per person expenditure; $a_{ele,i,j}$ is the percentage of households having access to electricity in the i^{th} population stratum based on expenditure levels in the j^{th} district's urban area. Households that did not have access to electricity were assumed to have lower monthly expenditure level. $EPC_{ele,i,j}$ is the average monthly electricity consumption per person (in the unit of kWh/person) in the j^{th} district's urban area for the i^{th} population stratum.

Indian households can use multiple fuel types for lighting and cooking. For example, LPG, kerosene, and firewood can all be used for cooking³¹. Each surveyed household in NSS specified what their primary cooking fuel was, although the reported consumption of each type of energy was used for all end-use purposes in the households. In the Census of India 2011, each district reported the number of households using each type of primary energy for cooking. The process of calculating residential LPG use was similar to the process of modelling electricity use, with additional consideration of the proportion of households using LPG as primary cooking fuel (see Eq. 2).

$$RES_{LPG,j} = \sum_i POP_{i,j} * a_{LPG,i,j} * EPC_{LPG,i,j}^{1st} + \sum_i POP_{i,j} * p_{LPG,i,j}^{2nd} * EPC_{LPG,i,j}^{2nd} + \sum_i POP_{i,j} * p_{LPG,i,j}^{3rd} * EPC_{LPG,i,j}^{3rd} \quad (2)$$

in which, $a_{LPG,i,j}$ is the percentage of households using LPG as primary cooking fuel in the i^{th} population stratum by expenditure in the j^{th} district's urban area. The proportion of households using LPG as the primary cooking fuel was extracted from the Census of India 2011³⁰. $EPC_{LPG,i,j}^{1st}$ (in the unit of kg LPG/person) was calculated from households whose primary cooking fuel was LPG in the i^{th} population stratum in the j^{th} district's urban area. $p_{LPG,i,j}^{2nd}$ and $p_{LPG,i,j}^{3rd}$ are the proportion of households using LPG as the secondary fuel or tertiary fuel. When $POP_{i,j}$ is multiplied with $a_{LPG,i,j}$ or $p_{LPG,i,j}^{2nd}$ or $p_{LPG,i,j}^{3rd}$, the actual population in each stratum using LPG as primary, secondary, and tertiary fuel are different and represent the fact that there is multi-fuel use in India with some households using LPG as primary, while others may use LPG as secondary or tertiary fuel for which the data were derived from the NSS. $EPC_{LPG,i,j}^{2nd}$ and $EPC_{LPG,i,j}^{3rd}$ is the average monthly per person LPG used as the secondary and tertiary fuel for the i^{th} population stratum in the j^{th} district's urban area.

The amount of kerosene consumed in households was calculated using a similar process as described above for calculating LPG use for each district's urban area. In contrast to LPG, kerosene can be used for both cooking and lighting. Thus, the percentage of households using kerosene as the primary cooking fuel was not used when quantifying the amount of kerosene use in each urban area. It was assumed that people had equal access to kerosene.

The AllUrE-India dataset also includes the amount of firewood used in households. The population scaling-up process for firewood use followed the same steps as Eq. 2. Our model, derived entirely from the NSS data, shows that the firewood, like other fuels, varies in its usage by quintile across different districts.

Industrial manufacturing. Coal and electricity use in industrial manufacturing were estimated using a hybrid approach. In this hybrid approach, we calculated average energy intensity by employee for coal and electricity use for each industry sector from the national Annual Industrial Survey (ASI) dataset²³, and multiplied it by the number of employees in each industrial sector in a district's urban area. The ASI is an annual factory-level survey conducted by the Minister of Statistics and Programme Implementation, Government of India²³. The location of each surveyed factory is reported at the state level, and whether the factory is in an urban area. The ASI uses the national industry classification (NIC) code released in 2008 to categorize each factory at the 5-digit level. The number of employees working in different sectors in each urban area is extracted from the Census of India 2011²⁵.

The average coal and electricity use intensity by employee by industrial sectors ($Energy\ Intensity_{employee,n}$ in Eq. 3) was calculated using a composite average at the 3-digit NIC code level. Aggregating all surveyed firms nationwide at the 3-digit NIC code ensured that most industrial manufacturing sectors had more than ten firms sampled. We did not separate plants in urban and rural areas, because this separation did not reduce the coefficient of variation in energy use per employee. The average coefficient of variation across 3-digit NIC level was 1.81 for surveyed plants in urban areas, which was not statistically different from both the coefficient in rural areas (1.56) and nation-wide (1.81). Coal use per employee also showed the same pattern. Therefore, we used the national average, because we had more samples across multiple industrial types. Energy use per employee

was calculated by dividing all plants' energy input with all employees in each 3-digit NIC code because the ASI samples were stratified based on 3-digit NIC code.

Data cleaning steps were conducted to remove outliers in the ASI samples when calculating energy use intensity by employee. The total number of employees in each factory was compared with the projected number of workers (as calculated using total working days and total working man-days provided in the ASI). The k-nearest neighbours algorithm was applied to replace the number of workers in factories showing internal misalignment with their own data. ASI reported both the physical quantity and monetary expenditures on coal and electricity, which was used to calculate the expenditure per physical unit of these two energy sources to affirm that the data were overall reasonable.

The Census of India 2011 reports the number of workers in each district by 3-digit NIC code with urban or rural areas separated. Workers have been categorized into the main workers (who worked more than half a year) and the marginal workers (who worked less than half a year). The ASI reports workers based on how many employees work in a factory. A weighting factor was needed to match the number of workers in these two data sources. The Economic Census 2013/14³² was used to calculate the weighting factors because it covered all factories in India and was used the same way to report the number of employees as in ASI. The comparison between the number of workers in the Census of India and employees in the Economic Census yielded the weighting factor for the main workers (70%) and the marginal workers (25%). After applying the weighting factors to the number of workers reported by the Census of India, we got 30,243,052 workers, which was comparable to the 30,357,249 reported in the Economic Census. The comparison of total full-time workers at the 3-digit NIC level demonstrated a goodness-of-fit of 0.82 (in Supplementary Fig. 1), indicating the weighting factors worked for different industries as well.

Equation 3 was used to calculate the electricity and coal use:

$$IND_{k,j} = \sum_{n=100,j}^{n=330,j} (Mainworkers_{n,j} * 0.7 + Marginalworkers_{n,j} * 0.25) \times Energy\ Intensity_{employee,n} \quad (3)$$

in which $IND_{k,j}$ is total k^{th} energy use (purchased electricity or coal) in the j^{th} district's urban area. And n refers to the n^{th} 3-digit NIC code in manufacturing sectors. The main and marginal workers in each district's urban area are from Census of India 2011. And $Energy\ Intensity_{employee,n}$ is the average energy use intensity by employee in the n^{th} industrial sector.

In contrast to coal and electricity, which were reported in physical quantity in the ASI, petroleum data did not specify the types of fuel and were only reported in monetary units in the ASI. There were many different types of petroleum products, which were not disaggregated by type, and only expenditure on all petroleum fuel was given, making it difficult to use the same approach. Therefore, petroleum fuel use in the industrial sector was downscaled from the national level to each urban area using total workers in that sector. Petroleum fuel use for industrial manufacturing at the national level was collected from the Petroleum Statistical Yearbook²⁷. The number of workers by industrial sectors was extracted from the Census of India²⁵.

Transportation sector. Estimating energy use for on-road transportation is complex and has been done in different cities and countries using different approaches. One approach collects local gasoline sales data (e.g., ICLEI-South Asia has collected cities' gasoline data from gas stations¹⁸). However, this approach is too place-specific to be applied to all 640 cities and may not accurately capture refuelling events outside the city. The second approach is to calculate energy use based on the number of vehicles in a city and the use of vehicles (e.g., vehicle kilometers travelled)³³, which is called the vehicle-activity approach. This approach can be used to inform low-carbon mobility transitions³⁴ and behavioural interventions. Some US cities collect trips' origin and destination details in travel behaviour surveys to better capture local travel demand;³⁵ however this type of survey is not available for Indian cities. The third approach uses household expenditure on gasoline or diesel for personal travel to calculate fossil fuel use along with fuel price³⁶. The household expenditure-based approach is only suitable to quantify fuel used for personal travel, but not for fuel used by buses or taxis due to a lack of relationship between expenditure on these modes and fuel use.

Based on data availability in India (Supplementary Table 1), we can use either the vehicle-activity approach or the household expenditure-based approach (detailed in Supplementary Part 1) to quantify fossil fuel used for personal travels. We tested the results from two approaches and found that the fossil fuel estimate of personal private mobility from the vehicle-activity approach was highly aligned with the expenditure-based approach (Supplementary Fig. 2 shows the goodness-of-fit is 0.8). Considering these two approaches' pros and cons, we adopted the vehicle-activity approach to estimate gasoline and diesel used for on-road travel (see details below).

The vehicle activity-based approach calculated the fuel use for on-road transportation based on Eq. (4):

$$TRA_{k,j} = \sum_v \frac{((Veh_{reg,v,j} \times \mu_v) \times VKT_v)}{FE_v} \times \tau_{k,v} \quad (4)$$

in which:

- $Veh_{reg,v,j}$ is the number of registered vehicle (the v^{th} type) in the j^{th} district's urban area. Vehicles are categorized into five types, i.e., two-wheelers, three-wheelers, cars, taxis, buses, and light- & heavy-duty vehicles. $Veh_{reg,v,j}$ is not available across all 640 districts. We used machine learning to get this value for all districts (method is detailed below).
- In India, registered vehicles are not deregistered even after they are no longer in use³³. The number of registered vehicles was transferred to the number of on-road vehicles by multiplying a ratio, μ_v . This ratio, μ_v , was extracted from literature for different vehicle types^{33,37,38} (Table 1).

	Two wheelers	Three wheelers	Cars & Jeeps	Taxis	Buses	Light- & heavy-duty vehicles
Annual VKT per year for cities with ≥ 6 million population (km/vehicle) ⁴¹	12,800	43,800	12,200	12,600	49,400	37,500
Annual VKT per year for cities with < 6 million population ³⁸ (km/vehicle)	6,750	29,500	9,000	10,000	36,400	
Fuel economy (km/L) ^a	49.9	26.9	15.2	15.2	5.8	4.3
% of vehicle in use (μ_v) ^b	89%	87%	97%	97%	52%	52%

Table 1. Annual VKT, fuel economy, and percentage of vehicles on road for estimating fuel use in on-road transportation sector by vehicle types. ^aFuel economy of two-wheelers, three-wheelers, cars and taxis was extracted from^{37,38}. Fuel economy for buses and light- & heavy-duty vehicles was extracted from⁴². ^bBased on comparing the age structure of different vehicle types^{37,38} and the survival rate from³³.

- VKT_v is vehicle kilometers travelled for the v^{th} type of vehicle. Prior work in India showed that the VKT was similar in smaller cities³⁹, while cities with more than 6 million had a significant increase in travel demand⁴⁰. Considering the differences between large and small cities, we used Delhi's VKT from⁴¹ for bigger cities (larger than 6 million) and VKT from³⁸ for smaller cities (Table 1).
- FE_v is fuel economy (in the unit of km/L) of the v^{th} vehicle type. FE_v was extracted from literature^{37,38,42} (Table 1).
- $\tau_{k,v}$ is the proportion of vehicle running by gasoline or diesel. This value was extracted from⁴³.

This vehicle-activity based approach captures inter-urban differences in fuel use for on-road transportation, because the number of registered vehicles demonstrates the highest variation across cities, compared to both VKT and fuel economy (Supplementary Table 2). The number of registered vehicles demonstrates variation at a factor of \sim thousand, introducing the largest variation across cities (Supplementary Table 2). The variation in VKT was much smaller, with maximum of 3^{33,37}, and other studies have shown even more minor variation of fuel economy of vehicles. The high agreement of vehicle-based approach versus expenditure-based approach for personal mobility across cities (Supplementary Fig. 2) indicated that the vehicle-activity approach can capture the variation of local fossil fuel use for personal mobility by focusing on the parameters contributing to the largest inter-city variation.

A machine learning approach was used to model registered vehicles (Veh_{reg}) for all 640 Indian districts. First, we collected the number of registered vehicles from 109 districts (in Assam⁴⁴, Haryana⁴⁵, Odisha⁴⁶, and Rajasthan⁴⁷) and 44 large cities (with more than 1 million population)⁴⁸. Previous local surveys found that households owning cars have on average one car per household⁴⁹. Therefore, we used the number of households owning cars from the Census of India to represent the number of registered cars. Second, using all districts'/cities' data, we explored the relationship between the number of registered vehicles and potential exploratory variables (see Supplementary Table 3), using a linear regression algorithm. We chose variables based on the demand for traveling by model (population size, workforce size, affluence level, education level, population density, commuting distance) and local supply level (e.g., number of workers in commercial sector providing transportation services). Specifically, the predictor variables included household expenditure on travel by different modes (estimated from the NSS²⁴ using a similar approach to calculate districts' electricity use), the total number of workers (main plus marginal workers)²⁵, the number of workers working in transportation sector²⁵, commuting distance by different traveling modes³⁰, and vehicle ownership from the Census of India 2011 (Supplementary Table 3). We applied the stepwise variable selection procedure to select the most significant variables in models for different vehicle types. The goodness-of-fit of models for different vehicle types ranges from 0.61 to 0.90 (Supplementary Table 3). For two-wheelers, motorcycle ownership is the most important variable, explaining about 85% of variation in registered two-wheelers. For three-wheelers, buses, taxis, and light-/heavy-duty vehicles, the number of cars is the most important variable, explaining 64%, 75%, 57%, and 64% of variation, respectively. The remaining variables (expenditure level, literacy rate, commuting distance etc.) together explained an additional 4% to 10% of variation. Last, these models were the foundation to conduct supervised machine learning to improve the prediction quality. These at-scale data were divided into a training dataset (80% of all cities/districts) and a testing dataset (20% of all cities/districts). We ran algorithms using tenfold cross-validation in the training dataset. The trained model was used to predict the number of vehicles across all districts' urban and rural areas. The predicted number of registered vehicles was used as a weighting factor to allocate the state-level registered vehicles to each district's urban area.

Diesel and electricity used in the railway sector were downscaled from the national total reported by the Government of India²⁶ based on households' expenditure on railway travel. The expenditure on the railway at the district level was calculated from NSS²⁴ based on households' monthly expenditure level.

Remaining sectors: commercial activities, agricultural sector, and informal businesses. The top-down approach was applied to estimate energy use in the remaining end-use sectors that could not be calculated using the bottom-up approach. These end-use sectors included the commercial activities (e.g., retail, hotels, etc.), agricultural activities, and households having informal businesses (e.g., selling street food). Petroleum and electricity use in these sectors were downscaled from the national statistics to each district's urban area based on the number

All Urban areas' Energy-use (AllUrE) across 640 Indian Districts for 2011	
SEIU categories	Examples of variables
Social demographics	• Total population
	• Total population by gender
	• Number of households
	• Infrastructural access
	• Electricity
	• Cooking fuels by types
	• Treated water
• Sanitation	
Economic feature	• Number of workers in different economic sectors
Infrastructure use	• Energy use by end-use sector and by fuel types
	• Residential sector
	• Industrial manufacturing
	• Transportation
Urban form	• Commercial sector
	• Population density

Table 2. Overview of the data structure in All Urban Areas' Energy-use across 640 Indian districts dataset providing selected examples of variables based on the Social-Economic-Infrastructural-Urban form (SEIU) framework^{5,9}.

of workers, with the assumption of similar energy use per employee in the same activity sectors. The number of workers was extracted from the Census of India 2011²⁵, and national total energy use in these sectors was extracted from the National Energy Balance Table²⁶.

Data sources for energy use reported at-scale across the national, state, and local levels. **National total** energy use in different end-use sectors was extracted from the energy balance table from National Energy Statistical Handbook 2013²⁶. The petroleum fuel in the national energy balance table only included LPG, kerosene, diesel, and heavy fuel oil. The amount of petroleum used in different end-use sectors was collected from the Petroleum and Natural Gas Statistical Handbook²⁷. In addition, about 85% of diesel use was marked as “end-use not specified” in both the national energy balance table and the petroleum handbook. This research reallocated the amount of diesel use to different vehicle types using data collected by the Petroleum Planning and Analysis Cell⁴³.

State-wide electricity in residential and industrial sectors, and gasoline and diesel use for on-road transportation data, were collected for comparisons at the state level. Both industrial and residential electricity use data at the state-level were collected from TERI Energy & Environment Data Diary and Yearbook⁵⁰. The state-wide total gasoline and diesel use (for transportation and non-transportation purposes) were provided in the Petroleum & Natural Gas Statistical Handbook²⁷. The total state-wide gasoline and diesel use for on-road transportation were extracted based on the regional gasoline and diesel use structure detailed in the All India Study on Sectoral Demand of Diesel and Petrol⁴³.

City-level energy use data are sparse and not available for a large number of cities over time. In this analysis, city-level energy use data were extracted from Solar City Master plans and ICLEI-South Asia's report on GHG patterns of south Asian cities¹⁸. Fuel use in different end-use sectors was extracted from 23 cities' solar city master plans. ICLEI-South Asia reported fossil fuel use by different economic sectors for the year 2007/2008 in 41 cities¹⁸, while there is no such report available for 2011. The reported energy use may be at the municipal corporation level, which can differ from the boundary of a city and a district's urban area. To ensure the same boundary of reported and estimated energy use, the same approaches to evaluate energy use at the urban areas were applied to cities using social-economic data from the Census of India 2011^{25,30}.

Data Records

Because infrastructure use in cities is shaped by social-economic factors (e.g., total population and employment structure), ecological factors, infrastructure access, and urban form (e.g., population density)^{5,51}, the inclusion of these non-energy use variables can help to investigate the drivers of energy use. Variables in the AllUrE-India dataset are therefore organized based on the social-ecological-urban-infrastructure (SEIU) system framework^{5,21} (Table 2).

In the dataset, variables are categorized into five sections: social-demographic variables (including infrastructure access variables), economic activities, energy use in four end-use sectors, and urban form (see Table 2). The ecological variables are not included in the AllUrE-India dataset due to the lack of broad coverage of such city-scale data in India. The Census of India 2011 reported socio-demographics (e.g., population, household number, age of population etc.) and households' access to infrastructural services (e.g., access to electricity, water, and sanitation) information at the district level with urban and rural areas separated. These non-energy use variables in AllUrE-India were all extracted from the Census of India 2011^{25,30} and merged with urban energy use data. An Excel file including the AllUrE-India 2011 datafile (tab name “AllUrE_20210104”) and codebook (tab name “codebook”) can be downloaded from <https://doi.org/10.6084/m9.figshare.12331283.v3>⁵².

Fuel type by end-use sector	National aggregation from AllUrE-India for 640 districts	At-scale national total energy use from GoI	Difference (GoI's data as the base)
a. Industrial manufacturing			
Coal and lignite (kilo tonne)	221,279	238,538	-7.2%
Electricity (GWh)	351,216	331,158	6.1%
b. Residential sector			
LPG (kilo tonne)	10,693	12,365	-13.5%
Kerosene (kilo tonne)	7,909	7,922	-0.2%
Electricity (GWh)*	197,606	170,034	16.2%
c. On-road transportation			
Gasoline (kilo tonne)	14,479	14,932	-3.0%
Diesel (kilo tonne)**	44,796	42,911	4.4%
d. Sum of national total electricity and fossil fuel end-use			
Coal and lignite (kilo tonne)***	221,279	238,538	-7.2%
LPG (kilo tonne)	12,693	14,364	-11.6%
Kerosene (kilo tonne)	8,093	8,106	-0.2%
Gasoline (kilo tonne)	14,573	14,932	-2.6%
Diesel (kilo tonne)	66,086	64,742	2.1%
Heavy fuel oil (kilo tonne)	6,455	6,455	0.0%
Electricity (GWh)	818,232	770,603	6.2%
Sum of energy use all above in the unit of Petajoule (PJ)	12,799	13,026	-1.7%

Table 3. National-scale validation: comparing energy use estimate aggregated nationally across 640 districts versus national total reported by the Government of India (GoI) in (a) industrial manufacturing, (b) residential sector, (c) on-road transportation. The comparison of total end-use energy by types was shown in part (d). Note: *In total coal use, coal used for generating electricity is not included to avoid double counting. **Residential electricity use is higher than officially reported, because households may report the amount of electricity use for informal commercial activities. ***Diesel use for aviation is not included.

Technical Validation

The quality of energy use data in the AllUrE-India dataset and the effectiveness of the bottom-up approach have been evaluated by comparing the sum of energy use from AllUrE-India and the estimates of rural areas' energy use with the national total reported by the Government of India. The same aggregation has been done at the state level to investigate the state-level alignment. A high alignment with at-scale national and state totals ensures that the AllUrE-India dataset can be used to evaluate the energy-saving and carbon mitigation potential of collective urban actions at the national or state level. We compared the effectiveness of a bottom-up approach with respect to the city-level data when such data were available.

National level alignment. Table 3 demonstrates total national energy use aggregated from AllUrE-India data modelled for 640 districts versus the at-scale national energy use reported by the government (Government of India, GoI) in different end-use sectors by fuel types. The total estimated national end-use energy is about 1.7% less than the national total reported by GoI. The differences in aggregated national energy use by types ranges from -11.6% to 6.2%.

State-level data comparison. Energy use aggregated from AllUrE-India plus the estimated rural energy use for 35 states is compared against at-scale state industrial and residential electricity use, as well as gasoline and diesel use for on-road transportation. The goodness-of-fit (R^2) for electricity is 0.95 and 0.78 in the residential sector (Fig. 3a) and the industrial sector (Fig. 3b), respectively. The goodness-of-fit (R^2) for gasoline and diesel use is 0.90 at the state level (Fig. 3c). These comparisons demonstrate that the estimated energy total at the state level had a high alignment with actual energy use at the state level.

City-level comparison. City-level energy use data are sparse and do not have broad coverage for a large number of cities during the same period. At-scale gasoline and electricity-use data reported by utilities can themselves have uncertainty. For example, non-technical electricity loss can be as high as 20% in India⁵³. When comparing AllUrE against the available at-scale data, it is found that the goodness-of-fit for electricity-use at the city-level is 0.99 (Fig. 4a), and is 0.90 for gasoline and diesel use for on-road transport at the city level (Fig. 4b). The normalized root-mean-square error is 3% and 44% for electricity-use and on-road fuel use, respectively.

The total electricity use derived from the NSS data fits well with utility-reported energy use data (at-scale), demonstrating the effectiveness of the bottom-up methodology (Fig. 4a). Because we differentiated household energy use by expenditure quintile, the AllUrE dataset can represent not only the aggregated total, but also the inequality in residential energy use, which is highly skewed to higher income populations.

The at-scale gasoline and diesel use data are for the period of 2007-2008, while the estimated transportation fuel use is for the year 2011. Furthermore, the estimated diesel and gasoline use based on local registered vehicle

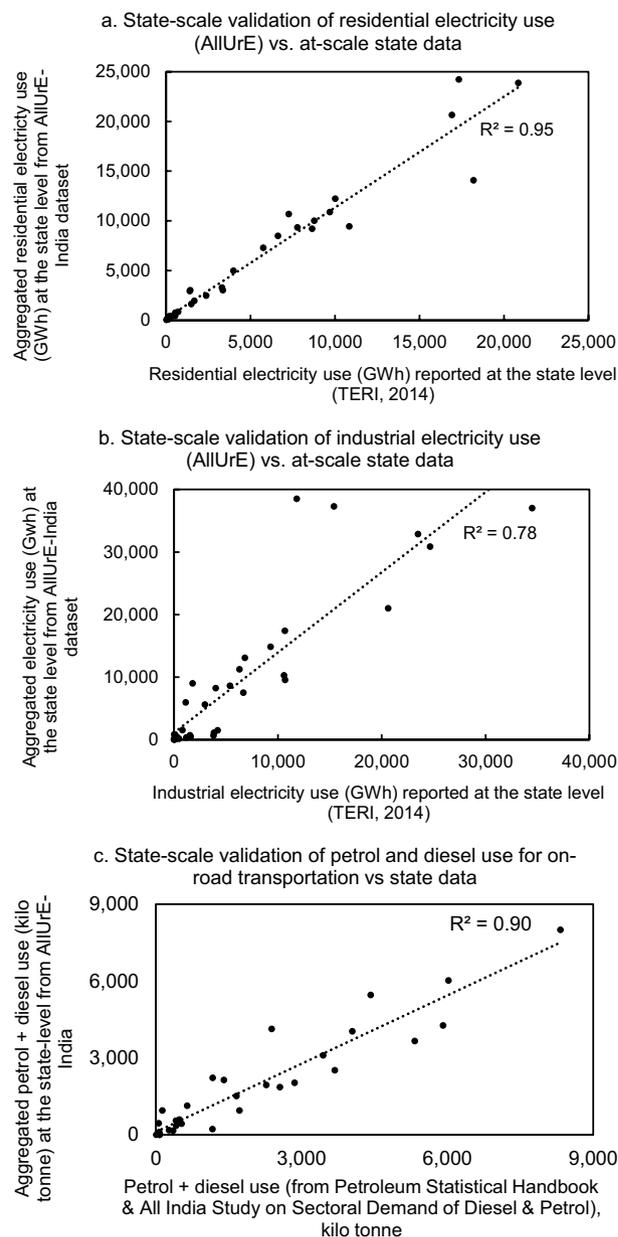


Fig. 3 State-scale validation comparing energy use from AllUrE-India dataset in (a) residential sector, (b) industrial manufacturing, and (c) on-road transportation versus energy use reported at the state level.

number is used for both local and long-distance travel. In comparison, diesel and gasoline use reported at the city-scale is based on the sales data from local gas stations¹⁸, meaning that amount of diesel is sold locally. Thus, different methods of estimating energy use for on-road transportation explain the difference between the at-scale and estimated data.

The validation across spatial scales demonstrates a new aspect of urban data science, where we are gathering data across all urban areas consistent with national and provincial/state levels with the focus on energy use rather than the location of fuel combustion.

Usage Notes

Overall, the database is novel in representing energy use in multiple sectors based on unique features of each urban district, while also aligning total energy use by different fuel types across spatial scales (city, state, and national). Specifically:

- Residential energy use captures both the population and inequality in access to energy, and consumption of energy in each district, using census data for population and access, along with consumer survey data for inequality in consumption.

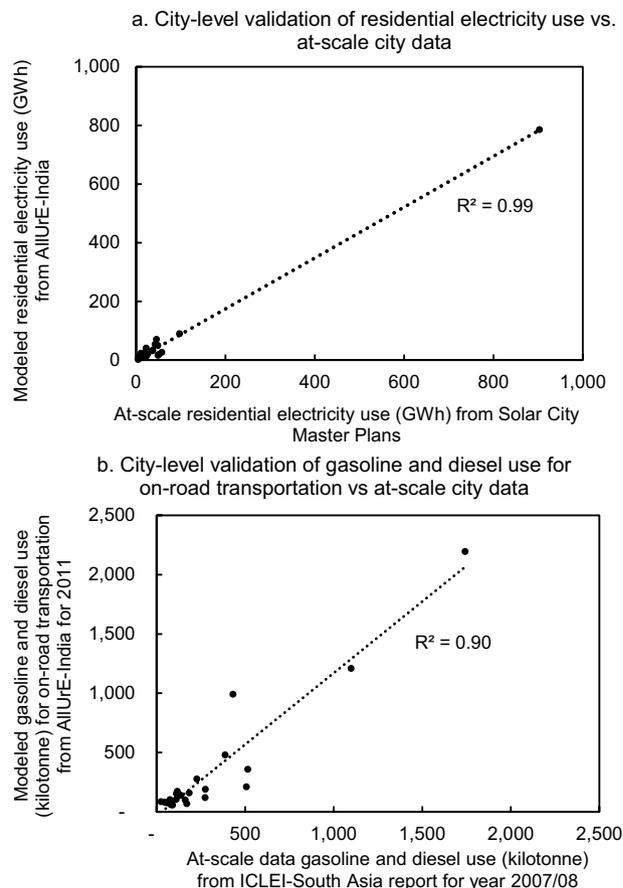


Fig. 4 City-scale validation: **(a)** compare modelled residential electricity use in AllUrE-India versus at-scale city-level electricity use from Solar City Master Plan documents; **(b)** compare modelled gasoline and diesel use for on-road transportation in AllUrE-India versus at-scale city-level data reported by ICLEI-South Asia.

- Industrial energy use mostly follows the number of industrial manufacturing workers by sector in each city from census data with sectoral energy use intensity derived from the Annual Industrial Survey. Commercial energy use reflects the number of workers in commercial sectors in each city. Thus, the employment activity in each district is reflected in energy use.
- Transportation energy use is calculated using modelled registered vehicles at the city level, which in turn is found to be highly correlated with local social-economic parameters in each district (e.g., automobile ownership, education, per capita expenditures, literacy), along with miles travelled represented for city-types in India (larger than 6 million or smaller than 6 million population).

Urban areas in this dataset were the aggregated administrative towns and municipalities defined by the census in each district. They covered all census defined cities, which are not equal to the total of urban agglomerations defined by remote sensing data.

The use of firewood has not been compared against any at-scale data, because biomass use is not reported in governmentally released statistical yearbooks. The effectiveness of this bottom-up method in the residential sector has been demonstrated when evaluating electricity and fossil fuel use. The uncertainty in estimated firewood was assumed to be the same as the electricity estimate in the residential sector.

City-specific freight data are very challenging to collect. In US cities, prior work has used input-output economic data to evaluate long-distance freight costs⁵⁴. In India, limited data are available for long-distance freights at the city level^{55,56}. There are data only for the 20 largest metro areas in India^{55,56}. In general, the bigger the city, the more freight goes into the city⁵⁷. In the analysis, we used 153 districts' at-scale reported trucks to build a model to predict the number of trucks in the rest of the cities, in which the number of workers (another metric for city size) demonstrated significant impacts on the number of registered trucks. The trucks' VKT and fuel economy were extracted from previous studies analysing energy use and air pollutant emissions³³. Although the energy use for freight is not combustion location-specific, fuel use for freights indicates how much energy is needed to transport goods in and out of the city. On-road freight accounts for 6.9% of national total end-use energy. Although there may be uncertainty at the city level, the energy balance at the national level is not influenced.

End-use energy in commercial and agricultural sectors was downscaled using the number of employees as a proxy. The top-down approach used for these two sectors in this research only ensures the alignment with the national total, while local activity features are not captured. Thus, the use of data to inform local energy

End-use sector	New data sources	New method for data estimation	Examples
Households	Street imagery data Phone tower data	Artificial intelligence or deep-learning approaches to process street imagery data to extract social-demographic, income data, rooftop solar PV.	<ul style="list-style-type: none"> • Suel <i>et al.</i> (2019) processed street imagery to extract detailed social-demographic data, i.e. income, education, unemployment, housing, living environment, health and crime⁵⁸. • Gebru <i>et al.</i> (2017) analyzed Google Street View images of cars to extract neighborhoods' socioeconomic attributes, e.g. income, race, education, and voting patterns⁵⁹. • Yu <i>et al.</i> (2018) processed imagery data to get all roof-top solar PV panels across the US⁶⁰. • Barbour <i>et al.</i> (2020) applied phone tower data to estimate building occupancy for estimating energy use⁶¹.
Industrial manufacturing	Satellite data, imagery data, global production data and production index to extract fossil fuel use.	Satellite data using short-lived air pollutants and local survey data to identify plants. Using the production data to allocate energy use to fine temporal scale.	<ul style="list-style-type: none"> • Van Damme <i>et al.</i> (2018) identified the ammonia point sources in industrial and agricultural sectors using satellite NH₃ measurement data⁶². • Liu <i>et al.</i> (2020) applied industrial production data to allocate fossil fuel associated CO₂ emissions to daily scale⁶³.
Transportation	GPS data for traffic congestion, etc. Phone tower data	Using big data, e.g., machine learning, and artificial intelligence techniques	<ul style="list-style-type: none"> • tomtom.com provides traffic congestion index for 416 cities across the globe⁶⁴. • streetlightdata.com employs high spatial-temporal solution data and survey data to extract travel demand.

Table 4. Emerging data sources and methods that can be used for developing all urban areas' energy use databases.

transitions in these two sectors is limited. In the future, new approaches can be used to quantify commercial buildings' energy use. For example, an alternative approach is to quantify energy use based on energy use intensity per floor area instead of energy use per employee. An alternative approach to estimate energy use in agricultural sectors is using the amount of food production and energy used per unit of food. Future research can explore new data analytical techniques and new data sources to support the development of these alternative bottom-up approaches to quantify energy use in these two end-use sectors.

Nowadays, the emerging urban data revolution provides more data and analytical approaches to model urban energy use that are being tested in other countries, which can be translated in the future to India and other countries. Table 4 presents the representative data sources and methods to generate such data at the urban level. Translating this to India will require more on-the-ground data and fieldwork, combining satellite data with cell phone data and on-the-ground surveys. In addition to mining emerging data, local and national governments can conduct surveys, such as travel demand surveys and commercial buildings energy use surveys, to develop urban energy use databases.

In this research, a novel bottom-up plus top-down approach is developed and adopted to assemble the AllUrE-India database for the year 2011. The database represents energy use in multiple sectors across all 640 urban districts of India, leveraging a number of surveys and machine learning models, representing both the unique features of individual urban areas, while also aligning total energy use by fuel types across city, state, and national totals. The method developed in this research is applicable to other countries with survey data, such as households and industrial/commercial energy use surveys to develop their all-cities' energy use and carbon footprinting databases.

Code availability

Data were processed in R version 3.5. and code was written in RStudio 1.1.463. The sample code to calculate residential electricity and LPG use using the National Sample Survey can be downloaded from⁵².

Received: 29 May 2020; Accepted: 29 January 2021;

Published online: 12 April 2021

References

1. Crippa, M. *et al.* Fossil CO₂ and GHG emissions of all world countries 2019. Report No. EUR 29849 EN, (Publications Office of the European Union, Luxembourg, 2019).
2. United Nations, Department of Economic and Social Affairs & Population Division. World Urbanization Prospects 2018 Highlight. (New York City, United States, 2018).
3. Andrew, R. Why India's CO₂ emissions grew strongly in 2017. *Carbon Brief* <https://www.carbonbrief.org/guest-post-why-indias-co2-emissions-grew-strongly-in-2017> (2018).
4. United Nations. Paris Agreement. (Paris, France, 2015).
5. Ramaswami, A. *et al.* A Social-Ecological-Infrastructural Systems Framework for Interdisciplinary Study of Sustainable City Systems. *J. Ind. Ecol.* **16**, 801–813, <https://doi.org/10.1111/j.1530-9290.2012.00566.x> (2012).
6. Government of India. India's intended nationally determined contribution: working towards climate justice. (Government of India, Ministry of Environment Forest and Climate Change, India, 2015).
7. Crippa, M. *et al.* High resolution temporal profiles in the Emissions Database for Global Atmospheric Research. *Sci. Data* **7**, 121, <https://doi.org/10.1038/s41597-020-0462-2> (2020).

8. Oda, T. & Maksyutov, S. A very high-resolution (1 km × 1 km) global fossil fuel CO₂ emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmos. Chem. Phys.* **11**, 543–556, <https://doi.org/10.5194/acp-11-543-2011> (2011).
9. Ramaswami, A., Russell, A. G., Culligan, P. J., Sharma, K. R. & Kumar, E. Meta-principles for developing smart, sustainable, and healthy cities. *Science* **352**, 940–943, <https://doi.org/10.1126/science.aaf7160> (2016).
10. Chertow, M. R. “Uncovering” Industrial Symbiosis. *J. Ind. Ecol.* **11**, 11–30, <https://doi.org/10.1162/jiec.2007.1110> (2007).
11. Chertow, M. R. & Lombardi, D. R. Quantifying Economic and Environmental Benefits of Co-Located Firms. *Environ. Sci. Technol.* **39**, 6535–6541, <https://doi.org/10.1021/es050050+> (2005).
12. ICLEI-USA. *U.S. Community Protocol for Accounting and Reporting of Greenhouse Gas Emissions*, <http://www.icleiusa.org/tools/ghg-protocol/community-protocol> (2012).
13. Ramaswami, A. *et al.* Urban Cross-Sector Actions for Carbon Mitigation with Local Health Co-Benefits in China. *Nature Clim. Change* **7**, 736–742, <https://doi.org/10.1038/nclimate3373> (2017).
14. Department of Energy & National Renewable Energy Laboratory. *State and Local Energy Profile*, <https://apps1.eere.energy.gov/sled/> (2015).
15. Tong, K. *et al.* The collective contribution of Chinese cities to territorial and electricity-related CO₂ emissions. *J. Clean Prod.* **189**, 910–921, <https://doi.org/10.1016/j.jclepro.2018.04.037> (2018).
16. Asher, S. & Novosad, P. Socioeconomic High-resolution Rural-Urban Geographic Dataset for India (SHRUG). *Harvard Dataverse* <https://doi.org/10.7910/DVN/DPESAK> (2019).
17. Chavez, A., Ramaswami, A., Nath, D., Guru, R. & Kumar, E. Implementing Trans-Boundary Infrastructure-Based Greenhouse Gas Accounting for Delhi, India. *J. Ind. Ecol.* **16**, 814–828, <https://doi.org/10.1111/j.1530-9290.2012.00546.x> (2012).
18. ICLEI-South Asia. Energy and Carbon Emissions Profile of 54 South Asian Cities. (2009).
19. Ahmad, S., Baiocchi, G. & Creutzig, F. CO₂ Emissions from Direct Energy Use of Urban Households in India. *Environ. Sci. Technol.* **49**, 11312–11320, <https://doi.org/10.1021/es505814g> (2015).
20. Ahmad, S. & Creutzig, F. Spatially contextualized analysis of energy use for commuting in India. *Environ. Res. Lett.* **14**, 045007, <https://doi.org/10.1088/1748-9326/ab011f> (2019).
21. Kammen, D. M. & Sunter, D. A. City-integrated renewable energy for urban sustainability. *Science* **352**, 922–928, <https://doi.org/10.1126/science.aad9302> (2016).
22. Office of the Registrar General & Census Commissioner. Census of India 2011: Meta Data. (2012).
23. Ministry of Statistics and Programme, Central Statistics Office & Government of India. Annual Survey of Industries 2011/2012. (Delhi, India, 2013).
24. MOSPI. in July 2011 - June 2012 (ed Schedule 1.0 NSS 68th Round, July 2011 - June 2012 (type 1) - Household Consumer Expenditure: National Sample Survey Office, NSSO) (India, 2012).
25. POMPI. Census of India 2011 Provisional Population Totals. (2011).
26. Central Statistics Office, National Statistical Organisation, Ministry of Statistics Programme Implementation & Government of India. Energy Statistics 2013. (2014).
27. Ministry of Petroleum & Natural Gas, E. D. & Government of India. Indian Petroleum & Natural Gas Statistics. (New Delhi, India, 2013).
28. Kuusela, P., Norros, I., Weiss, R. & Sorasalmi, T. Practical lognormal framework for household energy consumption modeling. *Energy Build.* **108**, 223–235, <https://doi.org/10.1016/j.enbuild.2015.09.008> (2015).
29. Nagpure, A. S., Reiner, M. & Ramaswami, A. Resource requirements of inclusive urban development in India: insights from ten cities. *Environ. Res. Lett.* **13**, 025010, <https://doi.org/10.1088/1748-9326/aaa4fc> (2018).
30. Ministry of Home Affairs, G. o. I. Houselisting and Housing Census Data - 2011. (2011).
31. National Sample Survey Office. Key Indicators of Household Consumer Expenditure in India. (New Delhi, India, 2013).
32. MoSPI. Sixth Economic Census 2013-14. (India, 2017).
33. Pandey, A. & Venkataraman, C. Estimating emissions from the Indian transport sector with on-road fleet composition and traffic volume. *Atmos. Environ.* **98**, 123–133, <https://doi.org/10.1016/j.atmosenv.2014.08.039> (2014).
34. Guttikunda, S. K. & Mohan, D. Re-fueling road transport for better air quality in India. *Energy Policy* **68**, 556–561, <https://doi.org/10.1016/j.enpol.2013.12.067> (2014).
35. Hillman, T., Janson, B. & Ramaswami, A. Spatial Allocation of Transportation Greenhouse Gas Emissions at the City Scale. *J. Transp. Eng.* **137**, 416–425, [https://doi.org/10.1061/\(ASCE\)TE.1943-5436.0000136](https://doi.org/10.1061/(ASCE)TE.1943-5436.0000136) (2011).
36. Noronha, P. No Change In Petrol, Diesel Prices On Thursday, <https://www.ndtv.com/business/petrol-price-today-petrol-diesel-prices-see-no-change-on-thursday-october-29-check-latest-rates-here-2317271> (2020).
37. Mohan, D., Goel, R. & Tiwari, G. Assessment of Motor Vehicle Use Characteristics in Three Indian Cities. Magnum Custom Publishing (2014).
38. Goel, R., Mohan, D., Guttikunda, S. K. & Tiwari, G. Assessment of motor vehicle use characteristics in three Indian cities. *Transp. Res. D Transp. Environ.* **44**, 254–265, <https://doi.org/10.1016/j.trd.2015.05.006> (2016).
39. Ministry of Urban Development. Study on Traffic and Transportation Policies And Strategies in Urban Areas in India. (2008).
40. IHS. Urban Transport in India Challenges and Recommendations. (Indian Institute for Human Settlements, 2015).
41. Goel, R., Guttikunda, S. K., Mohan, D. & Tiwari, G. Benchmarking vehicle and passenger travel characteristics in Delhi for on-road emissions analysis. *Travel Behav. Soc.* **2**, 88–101, <https://doi.org/10.1016/j.tbs.2014.10.001> (2015).
42. International Council on Clean Transportation. Fuel Consumption Standards for Heavy-duty Vehicles in India. (2017).
43. Petroleum Planning and Analysis Cell. All India Study on Sectoral Demand of Diesel & Petrol. (Ministry of Petroleum and Natural Gas, Government of India, New Delhi, India, 2013).
44. Director of Economics and Statistics & Government of Assam. Statistical Handbook Assam 2012. (Guwahati, Assam, 2013).
45. Department of Economic and Statistical Analysis. Statistical Abstract Haryana 2012-13. (Haryana State Government, Chandigarh, Haryana, 2014).
46. Directorate of Economics and Statistics Odisha. Statistical Abstract of Odisha 2012. (Bhubaneswar, Odisha, 2013).
47. Directorate of Economics and Statistics Rajasthan. Statistical Abstract of Rajasthan 2012. (Jaipur, Rajasthan, 2013).
48. Transport Research Wing, Ministry of Road Transport & Highways & Government of India. Road Transport Yearbook 2011–2012. (New Delhi, India, 2013).
49. Bansal, P., Kockelman, K. M., Schievelbein, W. & Schauer-West, S. Indian vehicle ownership and travel behavior: A case study of Bengaluru, Delhi and Kolkata. *Res. Transp. Econ* **71**, 2–8, <https://doi.org/10.1016/j.retrec.2018.07.025> (2018).
50. TERI. *TERI Energy & Environment Data Diary and Yearbook (TEDDY 2014/15)*. (2016).
51. Ramaswami, A., Jiang, D., Tong, K. & Zhao, J. Impact of the Economic Structure of Cities on Urban Scaling Factors: Implications for Urban Material and Energy Flows in China. *J. Ind. Ecol.* **22**, 392–405, <https://doi.org/10.1111/jiec.12563> (2018).
52. Tong, K., Nagpure, A. S. & Ramaswami, A. All Urban Areas’ Energy Use Data Across 640 Indian Districts: For Year 2011. *figshare* <https://doi.org/10.6084/m9.figshare.12331283.v3> (2021).
53. Gaur, V. & Gupta, E. The determinants of electricity theft: An empirical analysis of Indian states. *Energy Policy* **93**, 127–136, <https://doi.org/10.1016/j.enpol.2016.02.048> (2016).
54. Hillman, T. & Ramaswami, A. Greenhouse gas emission footprints and energy use benchmarks for eight US cities. *Environ. Sci. Technol.* **44**, 1902–1910, <https://doi.org/10.1021/es9024194> (2010).

55. Planning Commission. Total transport system study on traffic flows and modal costs., (Government of India, Delhi, India, 2008).
56. Planning Commission. Total transport system study on traffic flows and modal costs., (Government of India, Delhi, India, 2011).
57. Goyal, V. Urban Freights and Logistics: The State of Practices in India. (Deutsche Gesellschaft fur, 2014).
58. Suel, E., Polak, J. W., Bennett, J. E. & Ezzati, M. Measuring social, environmental and health inequalities using deep learning and street imagery. *Sci. Rep.* **9**, 6229, <https://doi.org/10.1038/s41598-019-42036-w> (2019).
59. Gebru, T. *et al.* Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl. Acad. Sci. USA* **114**, 13108–13113, <https://doi.org/10.1073/pnas.1700035114> (2017).
60. Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule* **2**, 2605–2617, <https://doi.org/10.1016/j.joule.2018.11.021> (2018).
61. Barbour, E. *et al.* Planning for sustainable cities by estimating building occupancy with mobile phones. *Nat. Commun.* **10**, 3736, <https://doi.org/10.1038/s41467-019-11685-w> (2019).
62. Van Damme, M. *et al.* Industrial and agricultural ammonia point sources exposed. *Nature* **564**, 99–103, <https://doi.org/10.1038/s41586-018-0747-1> (2018).
63. Liu, Z. *et al.* Carbon Monitor, a near-real-time daily dataset of global CO₂ emission from fossil fuel and cement production. *Sci. Data* **7**, 392, <https://doi.org/10.1038/s41597-020-00708-7> (2020).

Acknowledgements

This research is funded by NSF-SRN: Sustainable Healthy Cities [#1444745]. We thank Bhavin Pradhan and Graham Ambrose for assistance downloading select datasets, and Dana Boyer and Emily Eckart for editorial comments.

Author contributions

A.R. initiated the research ideas and supervised the development of the dataset. A.N. collected the raw data. K.T. and A.N. conducted data analysis. K.T. prepared and finalized the dataset and codes for sharing. K.T. and A.R. wrote the manuscript and responded to reviewers' comments.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-00853-7>.

Correspondence and requests for materials should be addressed to A.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021