# Estimating local-scale domestic electricity energy consumption using demographic, nighttime light imagery and Twitter data

Yeran Sun [a, b], Shaohua Wang [c *], Xucai Zhang [b], Ting On Chan [b], Wenjie Wu [d *]

[a] Department of Geography, College of Science, Swansea University, Swansea SA2 8PP, United Kingdom

[b] Guangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China

[c] CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[d] College of Economics, Jinan University, Guangzhou, 510632, China

[*]Correspondence:

Corresponding author: Shaohua Wang

Affiliation: CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign

E-mail: shaohua@illinois.edu

Address: 1301 W Green St, Urbana, IL 61801, United States

Corresponding author: Wenjie Wu

Affiliation: College of Economics, Jinan University

E-mail: wenjiewu@jnu.edu.cn

Address: No. 601 Huangpu Road West, Guangzhou, 510632, China

**Acknowledgments:** This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. 37000-31610453), China.

# Estimating England-wide district-level electricity energy consumption by built-up area and nighttime light intensity

**Abstract:**

To implement a new mixed approach for electricity energy consumption estimates, this study aimed to estimate country-wide local-scale electricity consumption by combining demographic, remote sensing, and social sensing data. Specifically, England-wide local-scale electricity energy consumption, including domestic and non-domestic ones, was estimated based on population in combination with nighttime light intensity or/and tweet volume. Moreover, to improve the explanatory power of statistical regression models, this study applied a newly developed spatial regression model (i.e., the 'random effects eigenvector spatial filtering' model) to the estimation of electricity energy consumption in comparison with conventional spatial regression models used in relevant studies. The spatial regression model used was further compared with machine learning and deep learning models (i.e., random forest and long short-term memory models). The empirical results uncover that: 1) the electricity energy consumption can be best explained by population in combination with both the nighttime light intensity and tweet volume; 2) the domestic electricity energy consumption can be better explained than its non-domestic counterpart; 3) the 'random effects eigenvector spatial filtering' models appear to outperform the conventional spatial regression models; and 4) the performance of the 'random effects eigenvector spatial filtering' models is similar to that of the random forest models and is lower than that of the long short-term memory models.

**Keywords:** electricity energy consumption; Twitter data; nighttime light imagery; SNPP-VIIRS; random effects eigenvector spatial filtering

## Nomenclature

| | | | |
|---|---|---|---|
| DMSP/OLS | Defense Meteorological Satellite Program / Operational Linescan System | | |
| SNPP-VIIRS | Suomi National Polar-orbiting Partnership - Visible and Infrared Imager/Radiometer Suite | | |
| ADEEC | annual domestic electricity energy consumption | | |
| ANEEC | annual non-domestic electricity energy consumption | | |
| ADEECPH | annual domestic electricity energy consumption per household | | |
| GTWR | geographically and temporally weighted regression | | |
| POP | population | LAD | local authority district |
| ANTLI | annual nighttime light intensity | OWTC | one-week tweet count |
| SL | spatial lag | SD | spatial Durbin |
| SE | spatial error | SDE | spatial Durbin error |
| SARMA | spatially autoregressive moving average | SAC | spatial autoregressive combined |

| | | | |
|---|---|---|---|
| SACD | spatial autoregressive combined Durbin | ESF | eigenvector spatial filtering |
| REESF | random effects eigenvector spatial filtering | LM | Lagrange Multiplier |
| VIF | variance inflation factor | OA | output area |
| LSOA | lower layer super output area | MSOA | middle layer super output area |
| RF | random forest | RNN | recurrent neural network |
| CNN | convolutional neural network | LSTM | long short-term memory |
| NMAE | normalized mean absolute error | NRMSE | normalized root mean square error |

## 1. Introduction

Electricity consumption trend differs from one country to another. In some developing countries, like China and India, electricity consumption increases more than two times from 2000 to 2015; whilst in some developed countries, like UK and Japan, electricity consumption is declining slightly [1]. From a geographic perspective, understanding how electricity consumption varies over space and time could contribute to trend analyses of electricity demand/supply at a city, region, or nation scale. A less costly and time-consuming approach for estimation of the energy consumption over space and time is urgently needed for not only researchers but also policymakers. Some policies have been implemented for a few years and thus needs to be evaluated for modifications in the future. A low-cost approach for rapid monitoring of electricity consumption can assist in evaluating policy implementations.

A conventional approach for estimation of the energy consumption is using demographic data [2-10]. Specifically, the energy consumption is estimated based on population and economic characteristics [2-10]. A few studies estimated the energy consumption in China according to the population and the gross domestic product (GDP) [2, 4, 6]. Apart from GDP, income [5, 10], energy price [6, 7], foreign investment [10], and tariff [9] were also selected to estimate the energy consumption. Nevertheless, it is extremely costly and time-consuming to update demographic data frequently, and thus demographic data cannot well support exploration of the spatiotemporal variations of the energy consumption. Besides, as most of the economic characteristics are available at the regional scale, the conventional approach based on demographic data is not applicable to energy consumption estimation at a finer scale. For instance, a few studies estimated China's provincial energy usage based on economic characteristics [2, 4, 6]

did not fit to the conventional approach. Apart from the energy consumption estimation in China, those in Brazil and Spain were used as the case studies-at a regional scale [9, 10].

Alternatively, some studies attempted to use other data sources to estimate energy consumption through tracking human activities over space and time. As a low-cost data source, remote sensing data were widely used as the proxy for human activity. Typically, in the last decade, nighttime light satellite imageries, such as the Defense Meteorological Satellite Program / Operational Linescan System (DMSP/OLS) and the Suomi National Polar-orbiting Partnership - Visible and Infrared Imager/Radiometer Suite (SNPP-VIIRS) imageries, have been used to estimate the energy consumption at multiple scales [11-20]. Relevant studies have revealed that a high correlation exists between energy consumption and nighttime light intensity at the country level [12, 15], regional level [11, 13, 16], or city level [19, 20]. For instance, a study assessed the spatiotemporal dynamics of the electricity consumption in core urban areas and suburban areas of China from 2000 to 2012 by using the nighttime light imageries [20]. The findings of the study suggested that the energy consumption in the suburban areas was more crucial for sustainable energy development in China [20]. Furthermore, some studies attempted to estimate energy consumption at a finer scale [14, 17, 18]. For example, the global electricity energy consumption at 1 km resolution was modeled using the intercalibrated nighttime light data obtained from 1992 to 2013 to assess spatiotemporal dynamics of the energy consumption from a global scale down to continental and national scales [17]. Another study applied geographically and temporally weighted regression (GTWR) models to the estimation of province-level energy consumption in China based on the DMSP/OLS global stable nighttime light data [18]. However, the reliability and feasibility of the models estimated in those studies still have a certain level of uncertainty owing to the absence of high-resolution observed data (e.g., historical records of household electricity meters or electricity sales). For instance, although there are a few studies estimated the energy consumption at 1 km resolution [14, 17, 18], their reference data used to validate the estimates are

interpolated or simulated rather than observed directly.

Apart from remote sensing data, social sensing data have been used as the proxy for human activity in very recent years [21, 22]. Like demographic and remote sensing data, social sensing data might have the potential to indicate different level of the energy consumption over space as human activity volume is positively correlated with the energy consumption. Popular sources of the social sensing data include mobile phones, social media, and mobile apps. Compared with mobile phone record data, social media data (e.g., the Twitter data) are highly accessible and spatially fine-grained. Therefore, the Twitter data seem to have the potential for the estimation of the energy consumption over space, even though they have some inherent disadvantages, e.g., heterogeneity issues in sample representativeness and sampling frequency [21, 22]. Particularly, since high densities of the tweets are geotagged around the eastern U.S. and the western Europe [23], the Twitter data are likely to be used in a variety of activity-related applications (e.g., the energy estimation) around those regions [21, 22].

Based on combination of multi-source geospatial data for the energy consumption estimates, this study aimed to propose a mixed approach for the estimation of the electricity consumption. Specifically, this study aimed to integrate three approaches (demographic, remote sensing, and social sensing ones) into one by combining the population data, the satellite imageries and the social media data. Moreover, to improve estimation results, we applied emerging statistical models to the estimation focusing on England with the population, the nighttime light intensity, and the tweet volume data. Specifically, the annual SNPP-VIIRS data was used to obtain the average nighttime light intensity; and one-week geotagged tweet data was used to extract the tweet volume. Using publicly available multi-sourced data, this study can pave a new way for estimation of the electricity consumption in some other countries where direct capture of the realistic consumption data is unavailable.

Most of the studies focusing on the estimation of the energy consumption are based on temporal variations [24, 25]. Temporal models (e.g., machine learning or deep learning models) were widely used to predict time-series of the energy consumption [24, 25].

Recently, the deep learning models, e.g., convolutional neural network (CNN), recurrent neural network (RNN) and long short-term memory (LSTM) models, had been broadly used to predict temporal variations of energy consumption [26-28]. Some other studies focused on spatial or spatiotemporal variations of the energy consumption [29, 30]. Statistical regression models were more widely used in the estimation for the spatial or spatiotemporal variations of the energy consumption than that of the temporal variations [29, 30]. Due to the data availability issue, this study focuses on the spatial variations of the electricity energy consumption rather than the temporal or spatiotemporal variations. Accordingly, the statistical regression models (e.g., spatial regression models) were used to estimate electricity energy consumption over space. Although machine learning or deep learning models are likely to outperform statistical regression models, statistical regression models have a higher level of explanatory ability. Particularly, the contributions of independent variables can be clearly and directedly quantified by the statistical regression models. Moreover, as one type of the explanatory models, the spatial regression models have been widely used to estimate energy consumption over space [2, 3, 6, 8, 10]. Theoretically, replacing the spatial regression models with non-spatial regression models (e.g., OLS models) can reduce estimation errors as the residual spatial autocorrelation in the non-spatial models can be estimated. The residual spatial autocorrelation is likely to undermine the fundamental assumption of residual independence in the regression models, and thereby needs to be reduced by the selection of the spatial regression models dedicated to address this issue. Empirically, ignoring spatial dependence of energy consumption is likely to cause biased estimation results since the spatial dependence of the energy consumption has been reported in the existing studies [3, 6]. Therefore, the spatial regression models are highly recommended for the estimation of the energy consumption over space. The basic types of the spatial regression models, such as the spatial lag model, spatial error model and the spatially autoregressive moving average model (a combination of a spatial lag model and a spatial error model), were early applied to the estimates of energy consumption [2, 3, 10]. One weakness of the basic forms of the spatial regression models is that the dependent variable in the models may

be explained not only by a spatially lagged dependent variable or spatially autocorrelated error term but also by a combination of a spatially lagged dependent variable and some spatially lagged independent variables [6]. Some studies lately proposed a better solution to overcome this disadvantage: the spatial Durbin model (SDM), which includes the spatially lagged dependent variables and the spatially lagged independent variables [3, 6, 8].

In this study, we made some contributions to abridge several research gaps in relevant studies: 1) as the highest geography level among most of the existing studies is the city level [19, 20], this study aimed to estimate the energy consumption at a finer scale; 2) this study uses realistic data (e.g., historical records of household electricity meters or electricity sales) to validate the estimated results at a finer scale, which ensures the reliability and feasibility of the models estimated in comparison with the relevant studies; 3) to improve the explanatory power of models, this study used a newly developed spatial regression model (i.e., random effects eigenvector spatial filtering model) in comparison with the conventional spatial regression models (e.g., spatial lag model, spatial error model, and spatial Durbin model) used in previous studies [2, 3, 6]; and 4) this study further compared the spatial regression models with the machine learning and deep learning models in the prediction of electricity energy consumption.

## 2. Methods

In this section, the research data and the used exploratory spatial data analysis method are firstly introduced. Subsequently, the estimation methods are presented, including those using the explanatory variables and responses as well as regression models (i.e., the ‘random effects eigenvector spatial filtering’ model).

### *2.1. Research Data*

The study area is England, which is experiencing a slight drop in electricity energy consumption in recent years. The areal unit used in this study is the local authority district (LAD). The LAD is a subnational areal unit widely used in the UK

demographics. In England, a LAD is a generic term covering the following regions: London boroughs, metropolitan districts, unitary authorities and non-metropolitan districts [31]. Although some LADs are styled as cities, most LADs are smaller than a city. The LAD is similar to a county for some countries. In other words, the LAD level is likely to be a higher geography level than the city level. According to the latest LAD boundaries data, there are 317 LADs in England [32].

Electricity energy consumption data: The GOV.UK offers data on electricity energy consumption at the local authority district (LAD) level across England [33]. The data contains electricity consumption amounts of domestic consumers, commercial and industrial consumers (non-domestic consumers), and all consumers. The realistic electricity consumption data is collected from household electricity meters and electricity sales in gigawatt hours (GWh). Electricity data is divided between domestic and non-domestic categories according to the meter's profile type [33]. The number of electricity sales (unit: GWh) is used to represent the levels of electricity energy consumption. We used the LAD-level electricity energy consumption data, including annual domestic and non-domestic electricity energy consumption Figure 1 maps the LAD-level annual domestic electricity energy consumption per household across England in 2016.

Population data: The Office for National Statistics (ONS) offers mid-year population estimates at the LAD level across England [34]. The mid-year population after 2011 is projected from the 2011 census data of UK. Figure 2 maps LAD-level population across England in 2016.

Nighttime light imagery data: There are two popular sources of the nighttime light satellite imageries: the DMSP/OLS and the SNPP-VIIRS. Compared to the DMSP/OLS data, the SNPP-VIIRS data have a higher spatial resolution, less saturated pixels, a wider radiance range, and a higher data quality owing to onboard calibrations [35-37]. Empirical studies also demonstrate that the SNPP-VIIRS data perform better than the DMSP/OLS data for a variety of applications, such as the GDP estimations [16], the $CO_2$ emission estimations [38], protected area detections [39], breast cancer incidence

estimations [40], and so forth. Although the SNPP-VIIRS data are superior to the DMSP/OLS data, the monthly SNPP-VIIRS data are only available from 2012 onwards and the annual SNPP-VIIRS data are only available from 2015 onwards [41]. The SNPP-VIIRS nightlight imageries can be downloaded from the webpage of the U.S. National Oceanic and Atmospheric Administration (NOAA) [41]. Specifically, the annual nighttime light data of SNPP-VIIRS were selected in this study [41]. The SNPP-VIIRS nighttime light imageries have a high spatial resolution (15 arc-second × 15 arc-second) than the DMSP/OLS imageries. Accordingly, the average area of the imagery grid is approximately 0.1-0.2 $km^2$ across England. The annual average light intensity of the SNPP-VIIRS data is represented by the annual average radiance (unit: nW $cm^{-2}$ $sr^{-1}$). Figure 3 maps the annual nighttime light intensity across England in 2016.

Twitter data: The utilized Twitter data is an open dataset [42]. The dataset is composed of geotagged tweets collected over a continuous week in April, 2016. The dataset contains 135,199 geotagged tweets located within England. Figure 4 maps the LAD-level one-week tweet count across England in 2016.

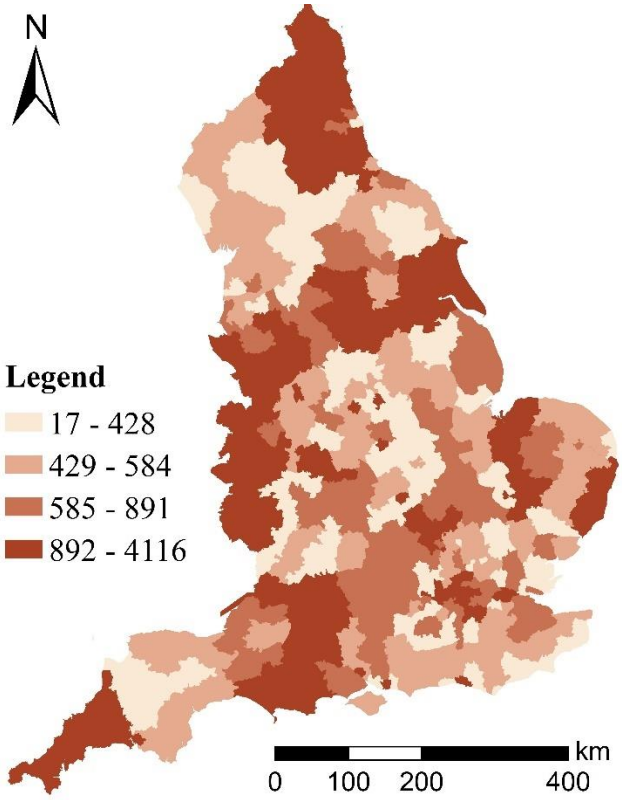



Figure 1: local authority district (LAD)-level domestic electricity energy consumption per household (unit: GWh) across England in 2016

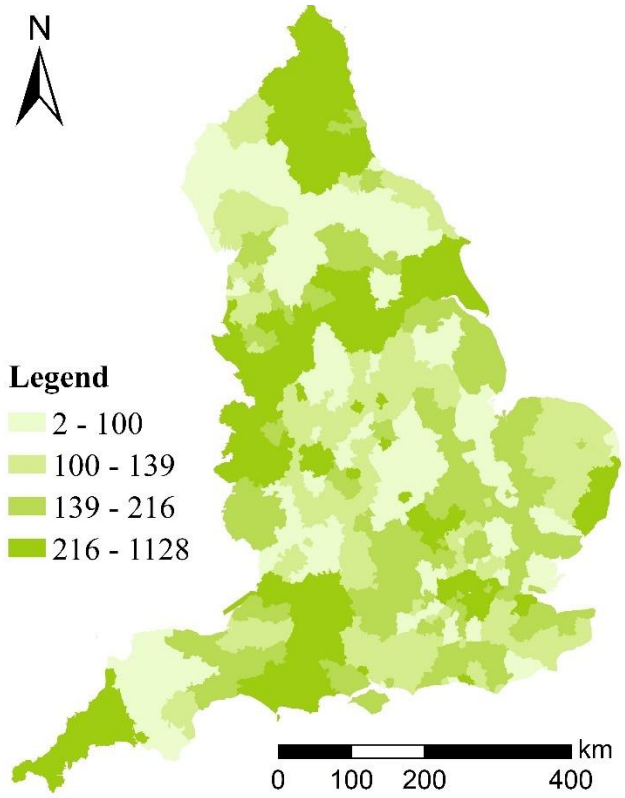



Figure 2: local authority district (LAD)-level population (unit: 1,000 persons) across England in 2016

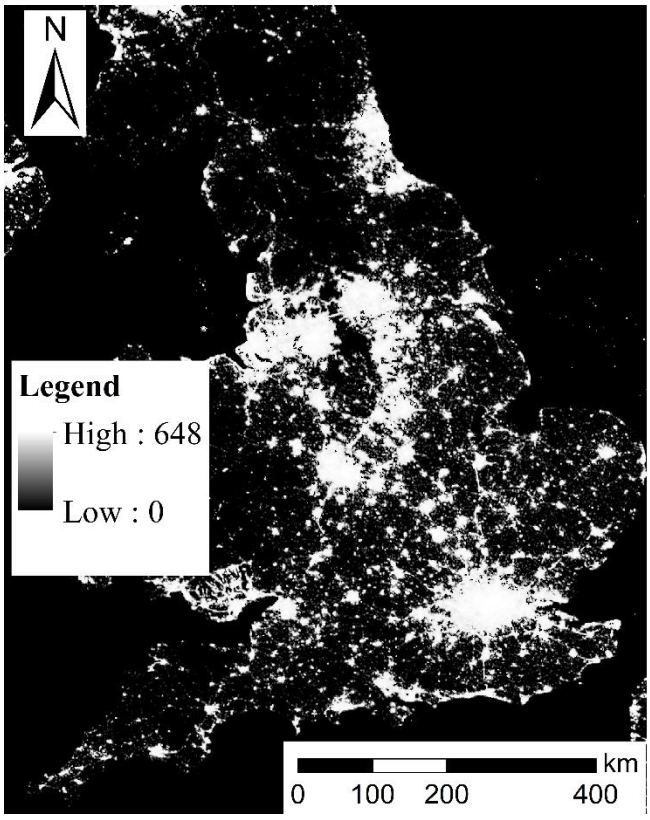



Figure 3: Annual average nighttime light intensity (unit: nW $cm^{-2}$ $sr^{-1}$) across England, 2016

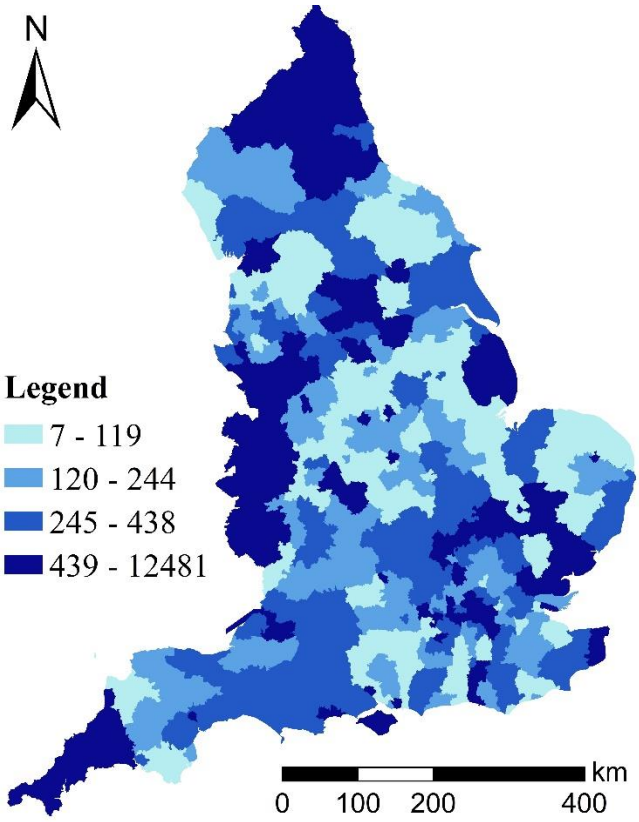



Figure 4: local authority district (LAD)-level one-week tweet count across England in 2016

*2.2. Exploratory spatial data analyses*

Before the electricity energy estimation, the exploratory spatial data analyses were first performed to explore spatial patterns of the LAD-level electricity energy consumption, with a focus on the spatial heterogeneity and the spatial association. The positive or negative spatial dependence (spatial autocorrelation) indicates that observations at spatially close locations tend to have similar or dissimilar values. The Moran's *I* is widely used to quantify the level of spatial autocorrelation between adjacent locations [43, 44]. Specifically, the Moran's *I* test statistic is defined as [43]

$$I = \frac{n}{W}\frac{\sum w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum w_{ij}(x_i - \bar{x})^2} \quad (1)$$

where *n* is the number of locations, $\boldsymbol{x_i}$ is the value of the variable *x* at the location *i*; $\boldsymbol{w_{ij}}$ is the spatial weight of the location *i* and *j*; and *W* is the sum of all $\boldsymbol{w_{ij}}$.

Specifically, a significantly positive Moran's *I* value indicates a high (low) value neighboured by high (low) values; whilst a significantly negative Moran's *I* value

indicates a high (low) value neighboured by low (high) values. Conventionally, the spatial matrix used to determine spatial relationships of the observations was computed based on the contiguity of the target areas (polygons). Consequently, the global and local forms of Moran's *I* statistic were used to measure the spatial dependence (spatial autocorrelation) globally and locally.

### *2.3. Estimates of models*

#### 2.3.1. Explanatory variables and responses

Table 1 lists the explanatory variables and responses used in this study. The two responses are the annual domestic electricity energy consumption (ADEEC) and the annual non-domestic electricity energy consumption (ANEEC) in 2016; whilst the explanatory variables are the population (POP), the annual NTL intensity (ANTLI), and the one-week tweet count (OWTC) in 2016. Table 1 also shows the statistical description for the variables used in this study. All the variables used are measured at the LAD level. As the NTL intensity is observed at the grid level, the aggregation of the NTL intensity from the grids to the LADs is needed. Specifically, the area-weighted average NTL intensity in each LAD is computed to represent the LAD-level NTL intensity. Supposing *i* is a LAD, its NTL intensity (NTLI) is calculated as:

$$LAD\_NTLI(i) = \sum_{j \in S(i)} NTLI(j) * \frac{area(i,j)}{LAD\_area(i)} \quad (2)$$

where *NTLI*(*j*) represents the NTL intensity of the grid *j*. $Area(i,j)$ represents the area of the overlapping part of grid *j* and LAD *i;* $S(i)$ is the set of grids overlapping with LAD *i*.

Besides, Table 2 shows the Pearson correlation coefficients of the explanatory variables and responses. Owing to the positive Pearson correlation coefficients, there are positive relationships between the responses (i.e., ADEEC and ANEEC) and the explanatory variables (i.e., POP, ANTLI, and OWTC). This indicates that the explanatory variables were properly selected for the estimation.

**Table 1.** Explanatory variables and responses and their statistical description

| Type | Variables | Full Names | Mean | SD |
|---|---|---|---|---|
| Responses | ADEEC | Annual domestic electricity energy consumption in 2016 (unit: GWh) | 286.95 | 179.26 |
| | ANEEC | Annual non-domestic electricity energy consumption in 2016 (unit: GWh) | 450.11 | 371.27 |
| Explanatory variables | POP | Population in 2016 (unit: 1,000 persons) | 172.83 | 117.01 |
| | ANTLI | Annual nighttime light intensity in 2016 (unit: nW $cm^{-2}$ $sr^{-1}$) | 8.99 | 13.92 |
| | OWTC | One-week tweet count in 2016 | 430.25 | 831.43 |

**Table 2.** Pearson correlation coefficients of explanatory variables and responses

| **Pearson coefficients** | POP | ANTLI | OWTC |
|---|---|---|---|
| ADEEC | 0.969 | 0.199 | 0.372 |
| ANEEC | 0.713 | 0.550 | 0.690 |

2.3.2. Model estimation

Conventional spatial regression models: The classical forms of the spatial regression models are modified based on the generalized linear models (OLS models). Specifically, the linear regression models are initially modified into two basic forms of the spatial regression models: the spatial lag (SL) model (also called spatial autoregressive model) and the spatial error (SE) model (also called spatial moving average model) by integrating spatially endogenous interactions and spatial interactions in the errors respectively. Moreover, as another basic form of the spatial regression models, the spatial autoregressive combined (SAC) model, also known as the spatially autoregressive moving average (SARMA) model, is a combination of the SL and SE models by incorporating spatially endogenous and spatial interactions in the error simultaneously into the generalised linear regression models. Furthermore, the three basic forms can be extended into the advanced forms by incorporating further the exogeneous interactions. Accordingly, the three advanced forms are the spatial Durbin (SD) model, the spatial Durbin error (SDE) model, and the spatial autoregressive

combined Durbin (SACD) model. Additionally, the Lagrange Multiplier (LM) diagnostics can recommend the appropriate form from the three types of classical spatial regression models (i.e., SL, SE and SAC).

Eigenvector spatial filtering model: as a new type of the spatial regression model, the spatial regression model with eigenvector spatial filtering has been also applied to process geospatial data [45, 46]. The Moran's eigenvector-based spatial regression approach is called the "eigenvector spatial filtering (ESF)" [45] in regional science, and the ESF with a small number of eigenvectors can greatly reduce model misspecification errors and increases the model accuracy [47]. Compared to the conventional spatial regression models estimated based on parametric methods (e.g., maximum likelihood estimation or Bayesian estimation), the eigenvector spatial filtering is computationally intensive since it is a nonparametric statistical method which is distribution free without sacrificing too much information within a sample [46]. Although the eigenvector spatial filtering (ESF) models are computationally demanding, they are likely to outperform the conventional spatial regression models for applications related to urban and regional studies, ecological studies, and so on [47]. Furthermore, the 'random effects eigenvector spatial filtering' (REESF) model had been developed because of its usefulness for spatial dependence analysis considering the spatial confounding [48]. The REESF models are found to outperform the conventional ESF models [47, 48]. Additionally, the MESS-SAR model produces $R^2$ values which are direct measures of the explanation capacity of the model; whilst the conventional spatial regression models do not.

### 2.3.3. Model validation

All the models were applied to an additional dataset (i.e., test dataset) for the estimations to further evaluate the performance. Apart from the regression models, a popular machine learning model (i.e., the Random Forest regression model) and a popular deep learning model (i.e., the Long Short-Term Memory regression model) were used to predict the test dataset for a broader comparison. The Random Forest

models have been proven to outperform other machine learning models (e.g., support vector regression) in previous studies [24]. The Long Short-Term Memory models perform well in energy prediction as reported by some studies [25, 54].

### *2.4. Implementation of analysis*

In this study, model estimates can be implemented in *R*. Specifically, a package named 'spatialreg' is designed for the conventional spatial regression models [49]; whilst another package named 'spmoran' is developed to implement estimates of the eigenvector spatial filtering models [50]. The package 'randomForest' supports the random forests for the classification and regression [51]; while the package 'keras' allows users to build a LSTM network [52]. Besides, the package 'spdep' supports the Lagrange Multiplier (LM) diagnostics [53]. Both global and local forms of Moran's *I* statistic can be implemented in the open software GeoDA [54].

## 3. Results and discussion

In this section, empirical results of the exploratory spatial data analysis and electricity energy consumption estimates are presented.

### *3.1. Exploratory spatial data analysis*

First of all, we explored the spatial patterns of the LAD-level electricity energy consumption intensity by examining spatial autocorrelation (dependence) in the 'annual domestic electricity energy consumption per household' (ADEECPH). As a population-based measure, the ADEECPH is used to represent electricity energy consumption intensity. In this study, the global and local forms of the Moran's *I* statistic were computed and simulated according to the 317 observations (317 LADs). First, we explored the global spatial autocorrelation in ADEECPH across England. Figure 5 shows the global Moran scatterplot of the England-wide LAD-level ADEECPH. A global Moran's *I* statistic value of 0.326 and a *p*-value of less than 0.001 indicate statistically significant presence of the global spatial autocorrelation in ADEECPH across England (see Figure 5). And, a significantly positive Moran's *I* statistic value (0.326) indicates that the observations at spatially close locations tend to have similar

values. Specifically, some LADs with high-value ADEECPH are neighboured by the LADs with the high-valued ADEECPH; and some LADs with the low-valued ADEECPH are neighboured by the LADs with low-valued ADEECPH across England.

Moreover, we explored the local spatial autocorrelation in ADEECPH across England. Figure 6 shows the clusters and outliers of LAD-level ADEECPH across England. Specifically, the clusters ('High - High' or 'Low - Low') indicate the LADs with high-valued (low-valued) ADEECPH are neighboured by the LADs with high-valued (low-valued) ADEECPH; whilst outliers ('Low - High' or 'High - Low') indicate LADs with low-valued (high-valued) ADEECPH are neighboured by LADs with high-value (low-valued) ADEECPH (see Figure 6). In this study, we focused more on the highly energy-intensive clusters ('High - High') and the lowly energy-intensive clusters ('Low - Low') since the outliers ('Low - High' or 'High - Low') are few. Particularly, three large concentrations of the highly energy-intensive clusters presented in southern England; whilst the two large concentrations of lowly energy-intensive clusters presented in northern England.

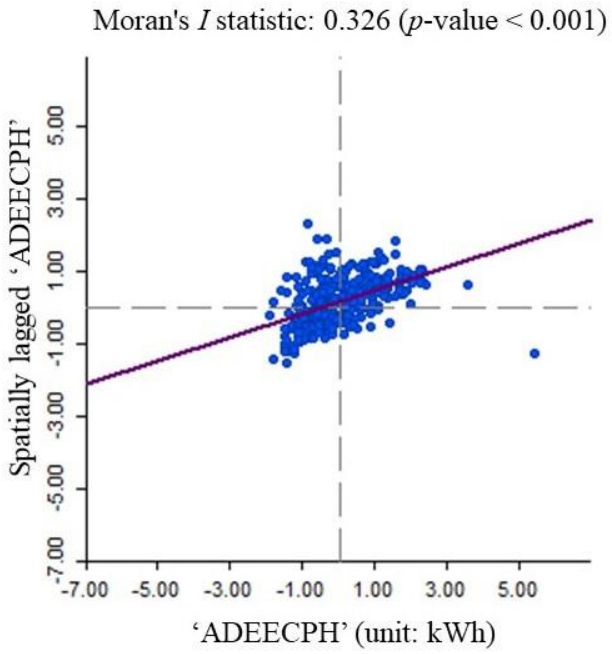



**Figure 5.** Global Moran scatterplot of England-wide 'annual domestic electricity energy consumption per household' (ADEECPH) in 2016

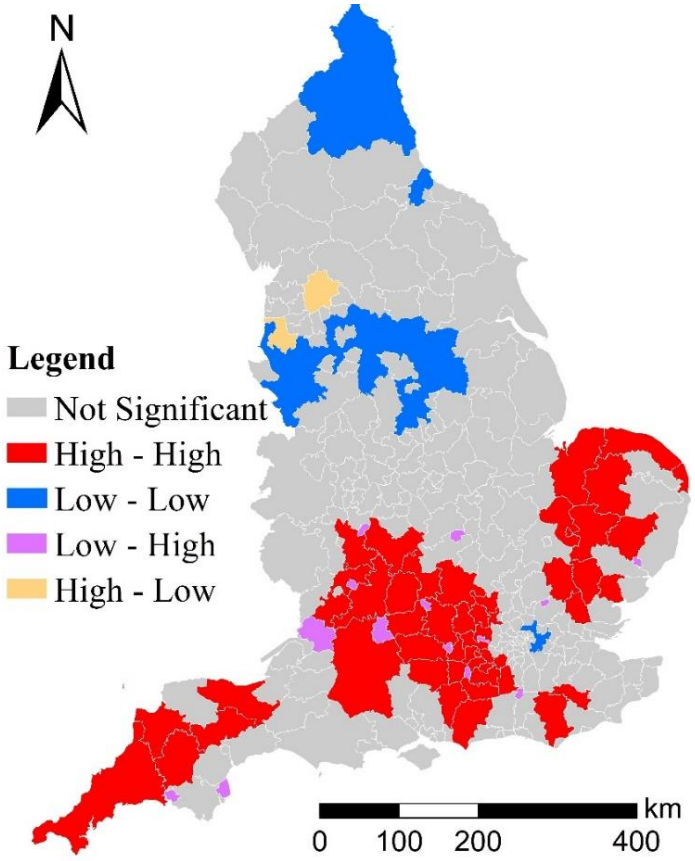



**Figure 6.** Clusters and outliers of 'annual domestic electricity energy consumption per household' (unit: kWh) across England in 2016

*3.2 Model estimation: electricity energy consumption estimates*

There are 3 explanatory variable combinations: "POP + ANTLI", "POP + OWTC", and "POP + ANTLI + OWTC". Accordingly, six (2 responses × 3 explanatory variable combinations) models were estimated for each regression model type.

3.2.1 Estimates of explanatory models

1) Estimates of OLS models

Initially, six non-spatial regression models (OLS models) were estimated based on population together with ANTLI, TC, or both respectively. As a result, six OLS models were estimated according to 317 observations (317 LADs).

2) Estimates of convetional spatial regression models (SL/SD/SE/SDE/SAC/SACD)

The Lagrange Multiplier (LM) diagnostics were implemented to conduct testing for

spatial dependence in the non-spatial regression model (i.e., OLS model). According to the LM diagnostics results, spatially simultaneous lag and error dependence is statistically significantly present in all the 6 OLS models estimated ($p$-values < 0.01). This verified the selection of SAC models or SACD models in this study. Accordingly, we chose the SACD models instead of the SAC models since the SACD model is an advanced form of SAC model. Eventually, 6 SACD models were estimated based on 317 observations (317 LADs).

3) Estimates of REESF models

Likewise, six REESF models were estimated based on population together with ANTLI, TC, or both respectively. As a result, six OLS models are estimated based on 317 observations (317 LADs).

3.2.2 Comparison of explanatory models estimated

Table 3 shows estimation results for all the explanatory models estimated. The Akaike information criterion (AIC) is used to measure the goodness-of-fit level of model. Unlike the OLS models, the SACD models were estimated through searching for maximum log likelihood. Therefore, the SACD models do not have adjusted $R^2$. In addition, the variance inflation factor (VIF) for all the explanatory variables (predictors) are below seven indicate that there is no serious multicollinearity exists in the models estimated. This means all the explanatory variables (predictors) are not highly correlated to each other. In all the models estimated, all the explanatory variables have statistically significant impact on the two responses. The exceptions are the coefficients of OWTC in the OLS and SACD models estimated for the ADEEC according to POP and OWTC. Expectedly, POP and OWTC are positively associated with both ADEEC and ANEEC. Interestingly, although ANTLI is positively correlated with ADEEC (see Table 2), it is negatively associated with ADEEC after other explanatory variables (i.e., POP and OWTC) are incorporated.

The comparison of the models estimated indicates some findings. First, for each explanatory variable combination or each response, the REESF model is likely to

outperform the SACD and OLS models due to higher $R^2$ or lower AIC values. Second, owing to higher $R^2$ or lower AIC values, the models accounting for both ANTLI and OWTC appear to outperform their counterparts accounting only for either ANTLI or OWTC alone. Likewise, owing to higher $R^2$ or lower AIC values, the models accounting for OWTC alone appear to outperform their counterparts accounting only for ANTLI alone. Third, for each explanatory variable combination or each regression model type, the ADEEC-aimed model outperforms its ANEEC-aimed counterpart.

Additionally, the Moran's *I* tests were used to check whether spatial correlation is present in the residuals of the models estimated. In Table 3, the Moran's *I* tests for regression residuals are statistically significant in all the OLS models but are not in all the SACD and REESF models. Two exceptions are in the ANEEC-aimed REESF models based on the two explanatory variable combinations: "POP + ANTLI" and "POP + OWTC". The Moran's *I* tests for regression residuals indicate that replacing non-spatial regression models with spatial regression models are likely to reduce the presence of residual spatial dependence.

**Table 3.** Estimation results for the explanatory models estimated

a) Response: ADEEC

| **Variables** | **POP + ANTLI** | | | **POP + OWTC** | | | **POP + ANTLI + OWTC** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **OLS** | **SACD** | **REESF** | **OLS** | **SACD** | **REESF** | **OLS** | **SACD** | **REESF** |
| Intercept | 34.583*** | 34.434*** | 32.905*** | 31.219*** | 39.25*** | 28.318*** | 36.239*** | 32.784 | 34.778*** |
| POP | 1.504*** | 1.504*** | 1.513*** | 1.455*** | 1.465*** | 1.465*** | 1.476*** | 1.473*** | 1.486*** |
| ANTLI | -1.104*** | -1.105*** | -1.081** | | | | -1.547*** | -1.578*** | -1.533*** |
| OWTC | | | | 0.005 | 0.005 | 0.008** | 0.017*** | 0.017*** | 0.016*** |
| Adjusted $R^2$ | 0.945 | | 0.951 | 0.938 | | 0.954 | 0.949 | | 0.954 |
| AIC | 3,276 | 3,278 | 3,263 | 3,311 | 3,311 | 3,287 | 3,253 | 3,255 | 3,250 |
| Residual Moran's *I* | 0.164*** | 0.398 | 0.078 | 0.192*** | 0.004 | -0.004 | 0.102*** | 0.003 | 0.017 |

Note: '.', '*', '**', and '***' mean the *p*-values are below 0.1, 0.05, 0.01, and 0.001 respectively.

b) Response: ANEEC

| Variables | POP + ANTLI | | | POP + OWTC | | | POP + ANTLI + OWTC | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **OLS** | **SACD** | **REESF** | **OLS** | **SACD** | **REESF** | **OLS** | **SACD** | **REESF** |
| Intercept | 28.732 | 53.733 | 10.931 | 63.656** | 23.66*** | 66.102*** | 46.437* | 63.542*** | 30.902** |
| POP | 1.899*** | 1.921*** | 1.839*** | 1.671*** | 1.621*** | 1.646*** | 1.598*** | 1.615*** | 1.535*** |
| ANTLI | 10.051*** | 10.982*** | 13.192*** | | | | 5.307*** | 5.962*** | 8.406*** |
| OWTC | | | | 0.221*** | 0.212*** | 0.225*** | 0.18*** | 0.179*** | 0.176*** |
| Adjusted $R^2$ | 0.637 | | 0.679 | 0.72 | | 0.725 | 0.748 | | 0.788 |
| AIC | 4,335 | 4,336 | 4,318 | 4,253 | 4,248 | 4,256 | 4,220 | 4,221 | 4,202 |
| Residual Moran's *I* | 0.116*** | -0.008 | 0.03* | 0.095** | 0.004 | 0.069* | 0.104** | -0.005 | -0.014 |

Note: '.', '*', '**', and '***' mean the *p*-values are below 0.1, 0.05, 0.01, and 0.001 respectively.

## 3.3 Validation of models estimated

In this study, the data for 2015 were used as the test data while the data for 2016 were used as the train data by which the models were estimated. Specifically, the POP and the ANTLI are represented by population in 2015 (unit: 1,000 persons) and the annual nighttime light intensity in 2015 (unit: nW $cm^{-2}$ $sr^{-1}$); whilst OWTC is still represented by one-week tweet count in 2016 owing to the absence of the Twitter data for 2015. The two responses are the annual domestic electricity energy consumption (ADEEC) in 2015 (unit: GWh) and the annual non-domestic electricity energy consumption (ANEEC) in 2015 (unit: GWh). Apart from the SACD and REESF models, Random Forest (RF) and Long Short-Term Memory (LSTM) models were also applied to the test data for a broader comparison. Specifically, the key parameters of the RF models were set as the previous studies suggested [24]: 1 as 'number of variables randomly sampled as candidates at each split' and 5 as 'minimum size of terminal nodes'; and the key parameters of the LSTM models were set as the previous studies suggested [55-57]:

2 as ‘number of hidden layers’, 1,000 as ‘number of epochs’, *Adam* as ‘optimizer category’, and 0.001 as ‘learning rate’.

The Normalized Mean Absolute Error (NMAE) and the Normalized Root Mean Square Error (NRMSE) are both used to measure the difference of the predicted and actual values after adjusting for scales. The NMAE is the average of mean error normalized over the average of all the actual values; while the NRMSE is the square root of the mean of the squares of the deviations normalized over the average of all the actual values. Table 4 shows that the NMAE and the NRMSE values for the predictions of the annual electricity energy consumption in 2015 by different models.

Owing to the lower NMAE or NRMSE values, the models combining ‘POP, ANTLI and OWTC’ outperform those combining ‘POP and ANTLI’ or ‘POP and OWTC’ (see the 6$^{th}$ and 7$^{th}$ columns of Table 4). This consistently indicates that incorporating both nighttime light intensity and tweet volume can improve the prediction of energy consumption. Owing to the lower NMAE or NRMSE values, the REESF models consistently outperform the SACD models (see the 3$^{rd}$ and 4$^{th}$ rows of Table 4). The performance of the RF models is similar to that of the REESF models in the prediction of ADEEC, as the RF models have lower NMAE values while the REESF models have lower NRMSE values (see the 4$^{th}$ and 5$^{th}$ rows of Table 4a). In addition, the performance of the RF models is higher than that of the REESF models in the prediction of ANEEC, as the RF models have lower NMAE and NRMSE values than the REESF models (see the 4$^{th}$ and 5$^{th}$ rows of Table 4b). Owing to the lower NMAE or NRMSE values, the LSTM models outperform the other three models in the prediction of both ADEEC and ANEEC (see the last rows of Table 4). Besides, we plotted the actual and predicted ADEEC based on POP, ANTLI and OWTC as the prediction accuracies of ADEEC are higher than those of ANEEC (see Figure 7). We also plotted the distribution of prediction errors. Specifically, we plotted relative absolute error (RAE) of predictions based on POP, ANTLI and OWTC (see Figure 8). Both Figure 7 and 8 are consistent with Table 4.

**Table 4.** Prediction accuracies of the regression models estimated

a) Response: ADEEC

| **Variables** | **POP + ANTLI** | | **POP + OWTC** | | **POP + ANTLI + OWTC** | |
|---|---|---|---|---|---|---|
| **Model** | ***NMAE*** | ***NRMSE*** | ***NMAE*** | ***NRMSE*** | ***NMAE*** | ***NRMSE*** |
| SACD | 0.086 | 0.155 | 0.092 | 0.152 | 0.086 | 0.152 |
| REESF | 0.077 | 0.137 | 0.082 | 0.133 | 0.075 | 0.132 |
| RF | 0.069 | 0.149 | 0.066 | 0.129 | 0.069 | 0.146 |
| LSTM | 0.018 | 0.025 | 0.018 | 0.025 | 0.018 | 0.025 |

b) Response: ANEEC

| **Variables** | **POP + ANTLI** | | **POP + OWTC** | | **POP + ANTLI + OWTC** | |
|---|---|---|---|---|---|---|
| **Model** | ***NMAE*** | ***NRMSE*** | ***NMAE*** | ***NRMSE*** | ***NMAE*** | ***NRMSE*** |
| SACD | 0.395 | 0.61 | 0.258 | 0.447 | 0.325 | 0.501 |
| REESF | 0.275 | 0.461 | 0.239 | 0.426 | 0.238 | 0.375 |
| RF | 0.191 | 0.338 | 0.178 | 0.322 | 0.164 | 0.283 |
| LSTM | 0.030 | 0.046 | 0.031 | 0.047 | 0.030 | 0.047 |

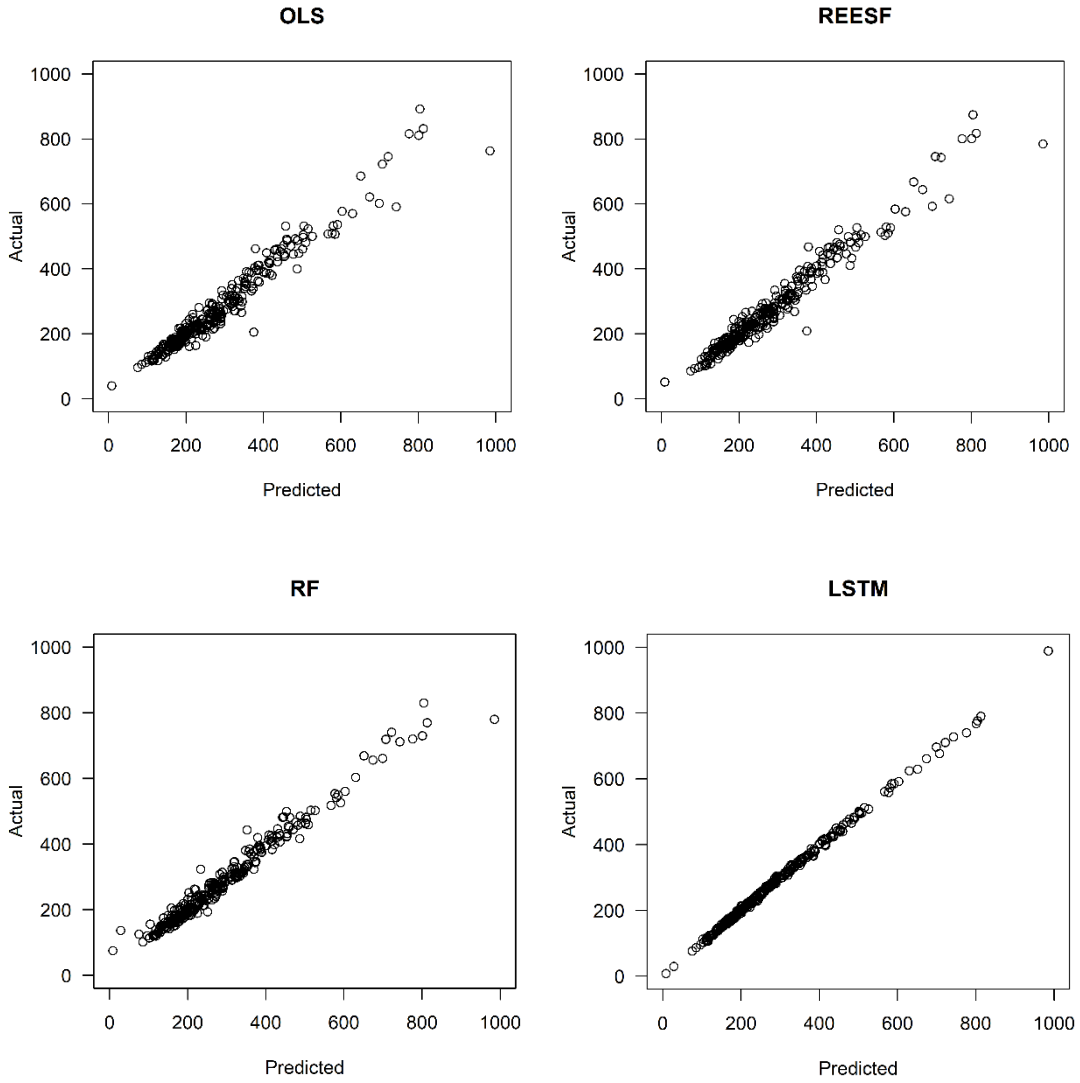
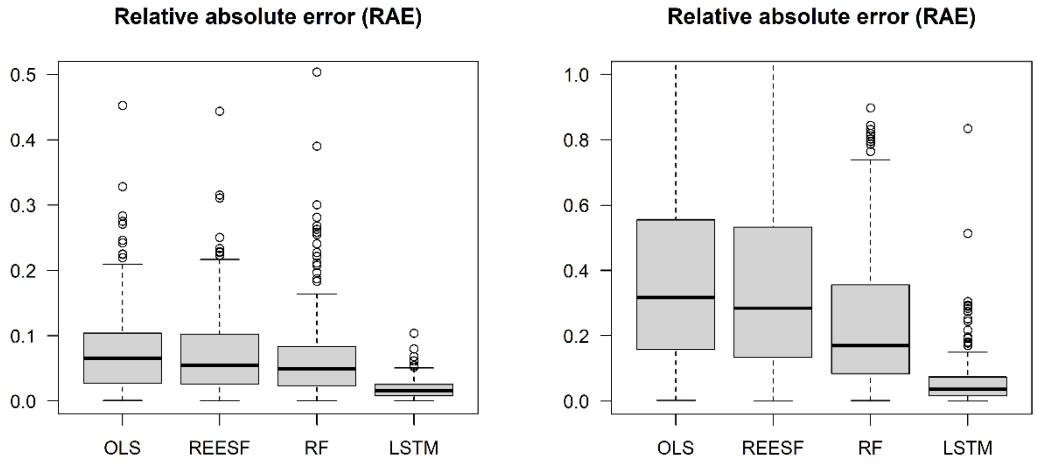



**Figure 7.** Actual and predicted ADEEC based on POP, ANTLI and OWTC




a) ADEEC b) ANEEC

**Figure 8.** Relative absolute error (RAE) of predictions based on POP, ANTLI and OWTC

*3.4. Discussion*

Additionally, we estimated the electricity consumption at a finer scale. In the UK demographic system, the output area (OA) is the basic unit of the census area or census tract. The OA is further aggregated into the lower layer super output area (LSOA) and the middle layer super output areas (MSOA). The MSOA is a higher geography level than the LAD. On average, each English LAD consists of approximately 21 MSOAs. We additionally estimated the MSOA-level electricity consumption across England. Due to a lower *R*-squared value and a higher AIC value, the England-wide MSOA-level electricity consumption is ambiguously explained (determined) as the LAD-level consumption. This exhibits the influence of geography level (areal unit) on the estimated results. The modifiable areal unit problem (MAUP) widely exists in statistical analysis of the aggregate geospatial data [58-61]. Several studies have confirmed that the statistical results vary based on scale and aggregation, attracting attentions from many individuals in the field of geospatial analysis. [60]. When areal units are aggregated into fewer, larger units for statistical analysis, the values associated with the variation of the data decreases, which will affect any associated statistical analysis [60]. In this study, as imagery grids were aggregated to the district units (LADs) and the neighbourhood units (MSOAs), modelling results tend to differ from districts to neighbourhood levels. Specifically, compared with the aggregation of grids into

neighbourhood units, the aggregation of grids into district units tends to lose more information associated with the variation of data, leading to a decrease in statistical analysis results (e.g., modelling results). Therefore, the MAUP issue might theoretically explain part of the decrease in the *R*-squared values of models estimated from the LAD level (district level) to the MSOA level (neighbourhood level).

One of the challenges for the extension of the models is that the coefficients estimated might differ from one country to another. Therefore, it is important to estimate the coefficients in different countries to explore whether and how the coefficients vary over space and time. The first step of extending the established model in this study is to apply these models to other European countries. The reason is that owing to similar economic development levels, the energy-related policies, the climate and cultural characteristics, the associations of electricity energy consumption, the nighttime light intensity, and the tweet count might be similar. This could be empirically checked through applying this approach to electricity energy consumption in other European countries where the governments also offer the local-scale reference data based on household meters. More specifically, we may use the models estimated to predict the Europe-wide local-scale electricity energy consumption based on the remote sensing and the social sensing data if the coefficients estimated are unlikely to differ from one country to another. Another challenge for the extension of the models is the data availability issue. Annual SNPP-VIIRS data (nighttime light imagery data) are only available for 2015 and 2016, and we have only Twitter data collected for 2016. Due to the limited data availability, we are not able to extend the spatial models to the spatiotemporal models.

Governments dedicated to reducing energy consumption and carbon emission reduction are likely to propose or modify energy-related policies based on some new research findings [62, 63]. It is vital to assess the implementations of policies targeting energy consumption and carbon emission reduction. This study demonstrates that the remote sensing data (e.g., the SNPP-VIIRS imageries) or/and the social sensing data (e.g., geotagged tweets) can provide a new approach to rapidly and easily monitor the energy consumption, enhancing policy implementation assessment and reducing operation cost. Unlike England, many countries do not offer the energy consumption data at local scales. The remote sensing data and the social sensing data would potentially play an important role in energy-related policy assessment in the countries where the governments do not offer local-scale energy consumption statistics.

## 4. Conclusion

To implement a new mixed approach for the estimation of the electricity energy consumption, this study modeled the England-wide local-scale electricity energy consumption by combining the demographic data, the nighttime light satellite imageries, and the geotagged tweets. Specifically, both the annual electricity energy consumption volumes of the domestic and non-domestic usages were estimated based on the population along with the nighttime light intensity or/and the tweet volume. Moreover, compared with the conventional spatial regression models, a newly developed spatial regression model (i.e., the random effects eigenvector spatial filtering model) was used. The empirical results uncover that 1) the electricity energy consumption can be best explained by the population along with both the nighttime light intensity and tweet volume; 2) the domestic electricity energy consumption can be better explained than its non-domestic counterpart; 3) The REESF models appear to outperform conventional spatial regression models; and 4) the performance of REESF models is similar to that of random forest models and is lower than that of the long short-term memory models. The governments and policymakers could use the SNPP-VIIRS data and the Twitter data to estimate local-scale electricity energy consumption annually. This approach is low-cost, and thus has high potential in monitoring the spatiotemporal variations in domestic electricity energy consumption.

As the Twitter data used was collected during a one-week period, we will attempt to collect the Twitter data during the whole 2016 to reduce the spatially sampling bias of the geotagged tweets likely caused by the occurrence of seasonal events. In the future, to check whether the approach can be extensively applicable to other countries, we will apply the approach to other European countries if local-scale electricity consumption statistics based on household meters are publicly available. We may also attempt to apply the approach to estimates of the Europe-wide local-scale electricity energy consumption in the near future. Furthermore, since there are some emerging spatiotemporal models, e.g., chain-structure echo state network (CESN) and CNN-LSTM models, used in other applications (e.g., solar irradiance prediction) [64, 65], we will attempt to apply those emerging models to the spatiotemporal energy consumption forecasting once monthly electricity energy consumption data were available in the future.

**References:**


1. Liu Z., 2015. Global Energy Development: The Reality and Challenges. *Global Energy Interconnection*, 1, pp.1-64.

2. Yu, H., 2012. The influential factors of China's regional energy intensity and its spatial linkages: 1988–2007. *Energy Policy*, *45*, pp.583-593.

3. Hao, Y., Liu, Y., Weng, J.H. and Gao, Y., 2016. Does the Environmental Kuznets Curve for coal consumption in China exist? New evidence from spatial econometric analysis. *Energy*, *114*, pp.1214-1223.

4. Hao, Y. and Peng, H., 2017. On the convergence in China's provincial per capita energy consumption: new evidence from a spatial econometric analysis. *Energy Economics*, *68*, pp.31-43.

5. Ding, Y., Qu, W., Niu, S., Liang, M., Qiang, W. and Hong, Z., 2016. Factors influencing the spatial difference in household energy consumption in China. *Sustainability*, *8*(12), p.1285.

6. Huang, J., Du, D. and Hao, Y., 2017. The driving forces of the change in China's energy intensity: an empirical research using DEA-Malmquist and spatial panel estimations. *Economic Modelling*, *65*, pp.41-50.

7. Liu, Y., Xiao, H., Lv, Y. and Zhang, N., 2017. The effect of new-type urbanization on energy consumption in China: a spatial econometric analysis. *Journal of Cleaner Production*, *163*, pp.S299-S305.

8. Xin-gang, Z., Yuan-feng, Z. and Yan-bin, L., 2019. The spillovers of foreign direct investment and the convergence of energy intensity. *Journal of cleaner production*, *206*, pp.611-621.

9. de Assis Cabral, J., Legey, L.F.L. and de Freitas Cabral, M.V., 2017. Electricity consumption forecasting in Brazil: A spatial econometrics approach. *Energy*, *126*, pp.124-131.

10. Gomez, L.M.B., Filippini, M. and Heimsch, F., 2013. Regional impact of changes in disposable income on Spanish electricity demand: A spatial econometric analysis. *Energy economics*, *40*, pp.S58-S66.

11. Townsend, A.C. and Bruce, D.A., 2010. The use of night-time lights satellite imagery as a measure of Australia's regional electricity consumption and population distribution. *International Journal of Remote Sensing*, *31*(16), pp.4459-4480.

12. Letu, H., Hara, M., Yagi, H., Naoki, K., Tana, G., Nishio, F. and Shuhei, O., 2010.

Estimating energy consumption from night-time DMPS/OLS imagery after correcting for saturation effects. *International Journal of Remote Sensing*, *31*(16), pp.4443-4458.

13. Zhao, N., Ghosh, T. and Samson, E.L., 2012. Mapping spatio-temporal changes of Chinese electric power consumption using night-time imagery. *International journal of remote sensing*, *33*(20), pp.6304-6320.

14. Cao, X., Wang, J., Chen, J. and Shi, F., 2014. Spatialization of electricity consumption of China using saturation-corrected DMSP-OLS data. *International Journal of Applied Earth Observation and Geoinformation*, *28*, pp.193-200.

15. Xie, Y. and Weng, Q., 2016. World energy consumption pattern as revealed by DMSP-OLS nighttime light imagery. *GIScience & Remote Sensing*, *53*(2), pp.265-282.

16. Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., Chen, L. and Wu, J., 2014. Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing*, 6(2), pp.1705-1724.

17. Shi, K., Chen, Y., Yu, B., Xu, T., Yang, C., Li, L., Huang, C., Chen Z., Liu, R., & Wu, J. 2016. Detecting spatiotemporal dynamics of global electric power consumption using DMSP-OLS nighttime stable light data. *Applied energy*, *184*, pp.450-463.

18. Xiao, H., Ma, Z., Mi, Z., Kelsey, J., Zheng, J., Yin, W., & Yan, M., 2018. Spatio-temporal simulation of energy consumption in China's provinces based on satellite night-time light data. *Applied energy*, *231*, pp.1070-1078.

19. He, C., Ma, Q., Li, T., Yang, Y. and Liu, Z., 2012. Spatiotemporal dynamics of electric power consumption in Chinese Mainland from 1995 to 2008 modeled using DMSP/OLS stable nighttime lights data. *Journal of Geographical Sciences*, *22*(1), pp.125-136.

20. Xie, Y., and Weng, Q., 2016. Detecting urban-scale dynamics of electricity consumption at Chinese cities using time-series DMSP-OLS (Defense Meteorological Satellite Program-Operational Linescan System) nighttime light imageries. *Energy*, *100*, pp.177-189.

21. Steiger, E., Resch, B., de Albuquerque, J.P., and Zipf, A., 2016. Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. *Transportation Research Part C: Emerging Technologies*, 73, pp. 91-104.

22. Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., & Tatem, A. J. 2017. Improving large area population mapping using geotweet densities. *Transactions in GIS*, *21*(2), pp. 317-331.

23. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E., 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5), pp. 4366.

24. Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S. and Ahrentzen, S., 2018. Random Forest based hourly building energy prediction. *Energy and Buildings*, 171, pp.11-25.

25. Wang, H., Lei, Z., Zhang, X., Zhou, B. and Peng, J., 2019. A review of deep learning for renewable energy forecasting. *Energy Conversion and Management*, 198, p.111799.

26. Kim, T.Y. and Cho, S.B., 2019. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, 182, pp.72-81.

27. Ullah, F.U.M., Ullah, A., Haq, I.U., Rho, S. and Baik, S.W., 2019. Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks. *IEEE Access*, 8, pp.123369-123380.

28. Yan, K., Li, W., Ji, Z., Qi, M. and Du, Y., 2019. A hybrid LSTM neural network for energy consumption forecasting of individual households. *IEEE Access*, 7, pp.157633-157642.

29. Huang, J. and Gurney, K.R., 2016. The variation of climate change impact on building energy consumption to building type and spatiotemporal scale. *Energy*, 111, pp.137-153.

30. Camargo, L.R. and Stoeglehner, G., 2018. Spatiotemporal modelling for integrated spatial and energy planning. *Energy, Sustainability and Society*, 8(1), p.32.

31. NHS, 2020. NHS Data Model and Dictionary. https://datadictionary.nhs.uk/nhs_business_definitions/local_authority_district.html

32. Office for National Statistics, 2020. Local Authority Districts (April 2019) Boundaries UK BFE. http://geoportal1-ons.opendata.arcgis.com/datasets/b6d2e15801de45328b760a4f55d74318_0

33. GOV.UK, 2019. Sub-national electricity consumption data. https://www.gov.uk/government/collections/sub-national-electricity-consumption-data

34. Office for National Statistics, 2019. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland

35. Elvidge, C. D., Baugh, K. E., Zhizhin, M., & Hsu, F. C., 2013. Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network*, *35*(0), pp.62.

36. Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C. and Ghosh, T., 2017. VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21), pp.5860-5879.

37. Li, X., Li, D., Xu H., & Wu C., 2017. Intercalibration between DMSP/OLS and VIIRS night-time light images to evaluate city light dynamics of Syria's major human settlement during Syrian Civil War. *International Journal of Remote Sensing*, 38(21), pp.5934-5951.

38. Ou, J., Liu, X., Li, X., Li, M., & Li, W., 2015. Evaluation of NPP-VIIRS nighttime light data for mapping global fossil fuel combustion CO2 emissions: a comparison with DMSP-OLS nighttime light data. *PloS one*, *10*(9), e0138310.

39. Xu, P., Wang, Q., Jin, J., & Jin, P., 2019. An increase in nighttime light detected for protected areas in mainland China based on VIIRS DNB data. *Ecological Indicators*, *107*, pp.105615.

40. Rybnikova, N. A., & Portnov, B. A., 2017. Outdoor light and breast cancer incidence: a comparative analysis of DMSP and VIIRS-DNB satellite data. *International Journal of Remote Sensing*, 38(21), pp.5952-5961.

41. NOAA, 2019. Version 1 VIIRS Day/Night Band Nighttime Lights. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

42. Followthehashtag, 2020. 170,000 UK geolocated tweets. Free Twitter Dataset. http://www.followthehashtag.com/datasets/170000-uk-geolocated-tweets-free-twitter-dataset/

43. Moran PA, 1950. Notes on continuous stochastic phenomena. Biometrika, 37(1/2): pp. 17-23.

44. Getis A, Ord JK, 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24, pp. 189-206.

45. Griffith, D. A. 2003. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer Science & Business Media.

46. Tiefelsdorf, M., and Griffith, D. A. 2007. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A*, 39 (5), pp. 1193-1221.

47. Murakami, D. and Griffith, D.A. 2019. Eigenvector spatial filtering for large data sets: fixed and random effects approaches. *Geographical Analysis*, 51 (1), pp.23-49.

48. Murakami, D. and Griffith, D.A., 2015. Random effects specifications in eigenvector spatial filtering: a simulation study. *Journal of Geographical Systems*, *17*(4), pp.311-331.

49. Bivand R. et al., 2019. spatialreg: Spatial Regression Analysis. https://cran.r-project.org/web/packages/spatialreg/index.html

50. Murakami, D., 2020. spmoran: Moran Eigenvector-Based Scalable Spatial Additive Mixed Modeling. https://cran.r-project.org/web/packages/spmoran/index.html

51. Liaw, A., and Wiener, M., 2018. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. https://cran.r-project.org/web/packages/randomForest/

52. Keydana, S. et al., 2020. keras: R Interface to 'Keras'. https://cran.r-project.org/web/packages/rnn/

53. Bivand, R. et al., 2020. spdep: Spatial Dependence: Weighting Schemes, Statistics. https://cran.r-project.org/web/packages/spdep/index.html

54. Anselin, L., 2013. GeoDa: An Introduction to Spatial Data Analysis. http://geodacenter.github.io/index.html

55. Xie, Y., Yang, H., Yuan, X., He, Q., Zhang, R., Zhu, Q., Chu, Z., Yang, C., Qin, P. and Yan, C., 2020. Stroke prediction from electrocardiograms by deep neural network. Multimedia Tools and Applications, pp.1-7.

56. Han, Y., Fan, C., Xu, M., Geng, Z. and Zhong, Y., 2019. Production capacity analysis and energy saving of complex chemical processes using LSTM based on attention mechanism. Applied Thermal Engineering, 160, p.114072.

57. Laubscher, R., 2019. Time-series forecasting of coal-fired power plant reheater metal temperatures using encoder-decoder recurrent neural networks. *Energy*, 189, p.116187.

58. Nakaya, T., 2000. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A*, *32*(1), pp. 91-109.

59. Zhang, M. and Kukadia, N., 2005. Metrics of urban form and the modifiable areal unit problem. *Transportation Research Record*, *1902*(1), pp. 71-79.

60. Dark, S.J. and Bram, D., 2007. The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, *31*(5), pp. 471-479.

61. Wong, D., 2009. The modifiable areal unit problem (MAUP). *The SAGE handbook of spatial analysis*, *105*(23), 2.

62. Sepehri, A., and Sarrafzadeh, M. H., 2018. Effect of nitrifiers community on fouling mitigation and nitrification efficiency in a membrane bioreactor. *Chemical Engineering and Processing-Process Intensification*, *128*, pp.10-18.

63. Sepehri, A., and Sarrafzadeh, M.H., 2019. Activity enhancement of ammonia-oxidizing bacteria and nitrite-oxidizing bacteria in activated sludge process: metabolite

reduction and CO2 mitigation intensification process. *Applied Water Science*, *9*:131.

64. Li, Q., Wu, Z. and Zhang, H., 2020. Spatio-temporal modeling with enhanced flexibility and robustness of solar irradiance prediction: A chain-structure echo state network approach. *Journal of Cleaner Production*, p.121151.

65. Zang, H., Liu, L., Sun, L., Cheng, L., Wei, Z. and Sun, G., 2020. Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. Renewable Energy, 160, pp.26-41.