scientific reports

Check for updates

OPEN

Spatial allocation of anthropogenic carbon dioxide emission statistics data fusing multi-source data based on Bayesian network

Jianbin Tao[1] & XiangBing Kong[2]✉

A gridded social-economic data is essential for geoscience analysis and multidisciplinary application. Spatial allocation of carbon dioxide statistics data is an important issue in the context of global climate change, which involves the carbon emissions accounting and decomposition of responsibility for carbon emission reductions. In this research a new spatial allocation method for non-point source anthropogenic carbon dioxide emissions (ACDE) fusing multi-source data using Bayesian Network (BN) was introduced. In addition to common-used DMSP (Defense Meteorological Satellite Program), PD (population density) and GDP (Gross Domestic Production) data, the land cover and vegetation data was imported into the model as prior knowledge to optimize the model fitting. The prior knowledge here was based on the understanding that ACDE was dominated by human activities and has strong correlations with land cover and vegetation conditions. A 1 km gridded ACDE map integrated emissions form point-source and non-point source was generated and validated. The model predicts ACDE with high accuracies and great improvement can be observed when fusing land cover and vegetation as prior knowledge. The model can achieve successful statistics data downscaling on national scale provided adequate sample data are available, offering a novel method for ACDE accounting in China.

The increasing greenhouse gas concentrations in the atmosphere and its resulting global warming has become the most important environmental issues[1] that affect the world's sustainable development[2] and is increasingly becoming the focus of global change research. Global warming is mainly due to the greenhouse effect caused by carbon dioxide, methane, etc., largely as a result of human activities over the past 50 years[3]. Carbon dioxide (CO_2) is the main source of anthropogenic greenhouse gas and the increasing of greenhouse effect caused by CO_2 now accounts for two-thirds of the total increasing[3]. The latest atmospheric carbon dioxide concentration released by Mauna Loa Observatory (Hawaii, USA) was high as 410 parts per million[4]. Anthropogenic Carbon Dioxide Emission (ACDE), or CO_2 emissions from fossil fuel consumption, is the result of human activities and an important indicator for decomposition of responsibility for emission reduction.

Since 2000 the average annual growth rate of ACDE in China was around 10%[5] , accounting for about 29% of the total global ACDE, and now is the world's leading carbon-emitting country[6]. China faces growing pressures when addressing climate change negotiations and greenhouse gas emission issues[7,8].

ACDE data are usually released officially as statistics data, which involves different statistical criteria and spatial inconsistency[9]. It's essential to decompose ACDE into pixel units and understand exactly the spatial distribution of ACDE. A unified finer spatial resolution ACDE map is proved to be crucial to climate change research and interdisciplinary research, also is of great value for China's carbon reduction strategies on a regional or national scale[10,11].

Many works have been done to develop ACDE maps on global, national and regional scales[12–14]. The existing researches depend heavily on nighttime lights data and population data, however there are several limitations in these spatial proxies. Ghost thought that the correlation of nighttime lights and ACDE was complicated and it's not possible to make independent estimates of ACDE with it[13]. Raupach also thought that the correlation between nighttime lights and human activity was significant only in developed countries[15]. Oda argued that population

[1]Key Laboratory for Geographical Process Analysis & Simulation of Hubei Province/School of Urban and Environmental Sciences, Central China Normal University, NO. 152 Luoyu Road, Wuhan 430079, China. [2]Yellow River Institute of Hydraulic Research, Zhengzhou 450003, Henan, China. ✉email: kongxb_whu@foxmaill.com

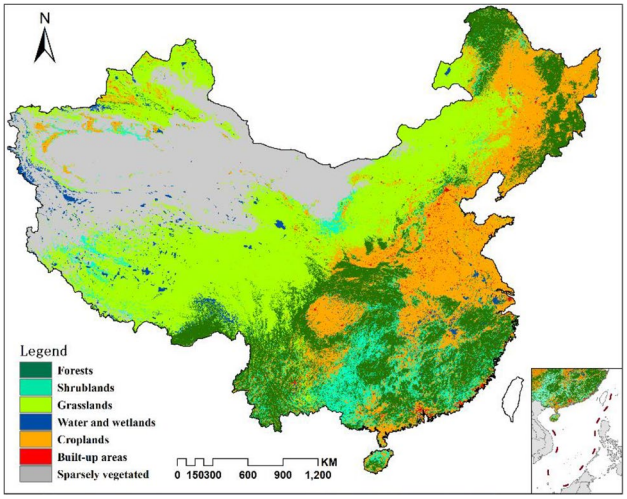


Figure 1. The research area and the major land-cover types in 2010 from MODIS land cover product (MCD12Q1 UMD). The fig was generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

statistics data did not explain the spatial pattern of nighttime lights well when spatial resolutions were finer than the country and state levels[12]. The limitation of nighttime lights was also reported by[16–18].

ACDE is closely related to human activities such as all kinds of fossil fuel consumptions. Land cover is the source of ACDE, reflecting the intensity of human activities[19] which has strong relationship with built-up areas. It has been proved that vegetation coverage has negatively correlation with impervious surfaces, the key component of built-up areas[20,21]. We can understand that high level ACDE can be observed in built-up areas, but we can't expect high level ACDE in high-vegetation-covered areas. In other words, there are strong relationships between ACDE and land-cover types (including vegetation conditions). Therefore, my idea is that land cover and vegetation condition should be considered in the spatial allocation model.

The objective of this study is to develop a spatial allocation model fusing multi-source data using Bayesian Network (BN). The specific objectives include: (1) Building a spatial allocation model for non-point source ACDE statistical data fusing multi-source data, including DMSP (Defense Meteorological Satellite Program), PD (Population Density) and GDP (Gross Domestic Production) data; (2) Introducing prior knowledge, including the probability distribution information about ACDE form land cover and vegetation data, into the model to improve the performance of the model; (3) Spatial allocation of ACDE, providing a basis for carbon emission accounting and carbon emission reduction. The novelty of the proposed model is that, the land-cover types and vegetation conditions were fused into the model as prior knowledge to mitigate the limitations in commonly used spatial proxies. Accuracies analysis were conducted by comparing the model result with the statistical data and ACDE products data at provincial, city and pixel level respectively. The advantages of the model in adding land cover and vegetation data as prior knowledge are demonstrated compared with conventional methods.

Research area and data

In this study, mainland China is selected as the research area. The major land-cover types include forests, shrublands, grasslands, water and wetlands, croplands, built-up areas and sparsely vegetated (Fig. 1). Some original IGBP (International Geosphere-Biosphere Programme) land-cover types were aggregated to obtain more general types.

Several data sources for 2005 and 2010 were used in this work (Fig. 2), including:

(1) DMSP/OLS nighttime lights data, from the US National Oceanic and Atmospheric Administration. The data has been preprocessed to eliminate the cloud and fire etc. The data values range from 1 to 63 and has a spatial resolution of 0.008333 degrees[22].
(2) PD and GDP data. The PD and GDP data were 1 km resolution raster data, which were obtained through spatial allocation based on PD and GDP statistics data (The data was source from National Earth System Science Data Sharing Infrastructure of China (http://www.geodata.cn)). The PD and GDP statistics data were also extracted from the statistical yearbook which was used for modeling their relationship with ACDE data at provincial level.
(3) Land-cover data. MCD12Q1, the global 1 km land-cover MODIS products, was selected as land-cover data. MCD12Q1 contains five land cover classification systems, within which IGBP system[23] are commonly used.
(4) Vegetation data. The vegetation data was EVI (Enhanced Vegetation Index) from the MOD13A2 v006 data, which was sourced from EOS/Terra Satellite. The products were composites synthesized over 16 days[24] and covered years 2005 and 2010. The Savitzky-Golay algorithm was used to filter and reconstruct the time-series EVI data[25] by referencing the pixel reliability data layer.

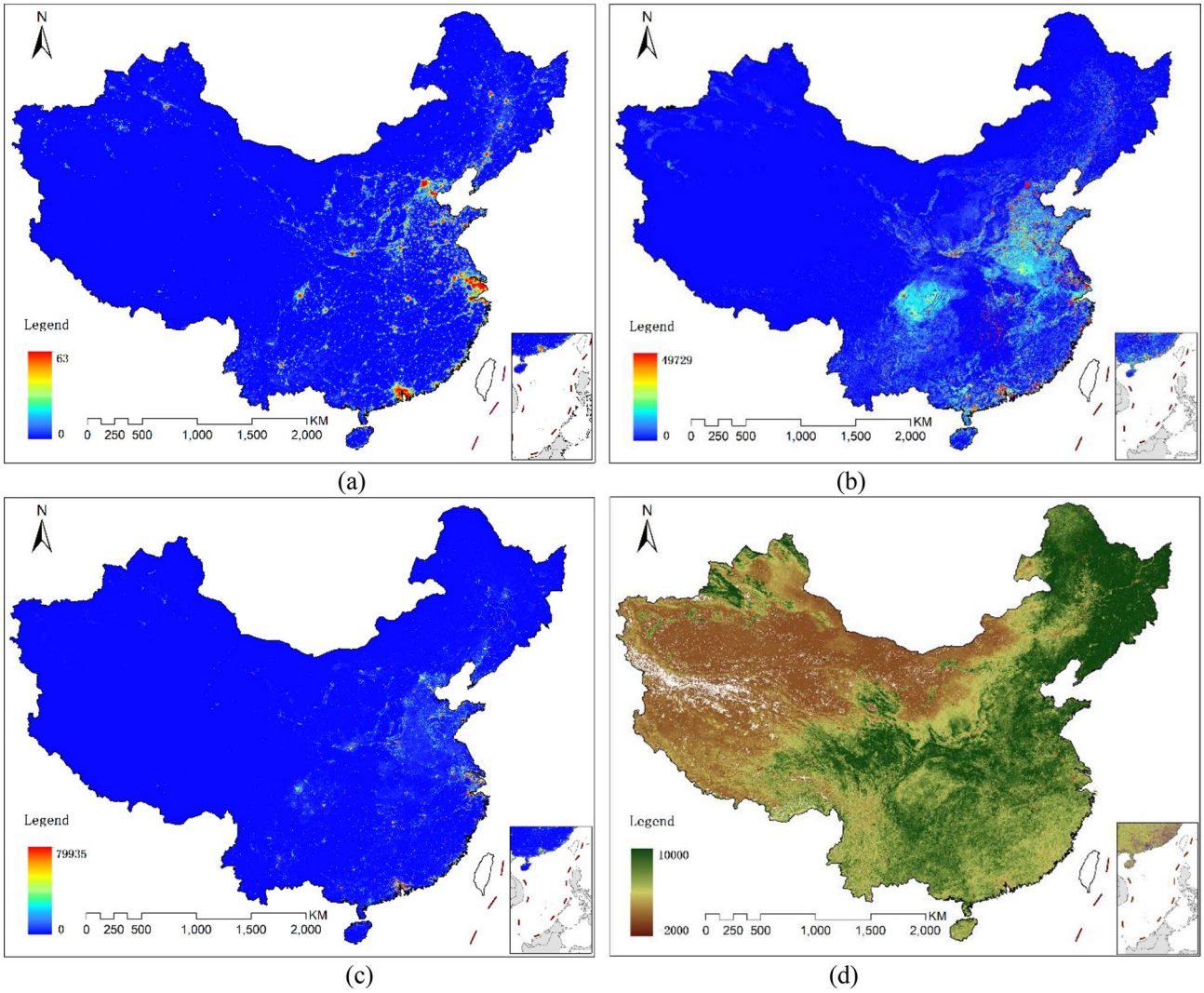


Figure 2. The datasets in 2010, (**a**) DMSP, (**b**) PD, (**c**) GDP, (**d**) EVI. The figs were generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

(5) ACDE statistical data. The energy consumption statistical data of year 2005 and 2010 covering 30 provinces or municipalities (excluding Tibet, Taiwan, Hong Kong and Macau) were sourced from the China Energy Statistical Yearbook. The data used to calculate ACDE is from the primary energy consumption of end energy consumption in the balance sheet. The 2006 IPCC Guideline[26] was used to calculate carbon-dioxide emissions from energy sources. Nine types of fossil energies, including coal, coke, crude oil, gasoline, kerosene, diesel oil, fuel oil, natural gas and electricity, were selected to calculate carbon emissions form energy consumption. The formula is as follows:

$$A_C = \frac{44}{12} \times \sum_{i=1}^{9} K_i E_i \tag{1}$$

A_c is carbon emissions form energy consumption. E_i is carbon emissions for energy i with unit 10^4 t. K_i is carbon emission coefficient for energy i with unit (10^4 t carbon)/(10^4 t coal equivalent). i denotes the type of energy. K_i are default carbon emission coefficients from IPCC Guideline and are given in Table 1. Provincial ACDE statistical data are given in Table 2.

Provincial ACDE statistical data are presented in Table 2.

(6) Carbon dioxide emissions from point sources. The carbon dioxide emissions from point sources are calculated separately by using the power-plant database CARMA (Carbon Monitoring and Action, http://carma.org). All power plants that use fossil fuels were selected from the database and their emissions were recorded. All power plants located in mainland China were reviewed using Google Earth images. Those power plants that cannot match with the locations will be included in non-point source.

Energy types	Convertion to coal equivalent (t coal equivalent) (t^{-1})	Carbon emission coefficients (10^4 t coal) (10^4 t coal equivalent)$^{-1}$
Coal	0.7143	0.7559
Coke	0.9714	0.855
Crude oil	1.4286	0.5857
Gasoline	1.4714	0.5538
Kerosene	1.4714	0.5714
Diesel oil	1.4571	0.5921
Fuel oil	1.4286	0.6185
Natural gas	1.33	0.4483
Electricity	0.345	0.272

Table 1. Carbon emission coefficients for different energy types. *Conversion coefficient for natural gas is t/ten thousand m^3. Conversion coefficient for electricity is t/ten thousand Wh.

Provinces	2005	2010	Provinces	2005	2010
Beijing	3211.23	3488.87	Henan	11940.63	17032.00
Tianjin	3460.45	5077.85	Hubei	6798.00	10142.17
Heibei	16377.65	22951.56	Hunan	6094.49	8269.43
Shanxi	16121.15	18925.19	Guangdong	10841.27	15923.43
Inner Mongolia	8838.09	16948.17	Guangxi	2804.41	4876.04
Liaoning	13708.68	18921.23	Hainan	344.00	1351.58
Jilin	4897.36	6996.86	Chongqing	2846.83	4160.74
Heilongjiang	6903.82	9521.93	Sichuan	6036.17	8853.56
Shanghai	6372.66	7357.40	Guizhou	4679.64	6581.40
Jiangsu	13544.28	18681.41	Yunnan	5005.58	6801.31
Zhejiang	8443.77	11974.65	Shaanxi	4968.40	9370.20
Anhui	5575.45	8856.60	Gansu	3604.74	4792.75
Fujian	3730.64	6268.16	Qinghai	703.74	1043.99
Jiangxi	3325.59	4889.76	Ningxia	2055.82	3613.54
Shandong	20118.87	30514.86	Xinjiang	4014.35	7362.58

Table 2. Provincial ACDE statistical data (ten thousand tons). *Data for Tibet, Taiwan, Hongkong and Macau are missing.

All the raster datasets were re-projected into the same coordinate system (WGS 1984), and re-sampled to 1 km grid using nearest neighbor method. Zonal statistics function in ArcGIS was used to get the mean and sum value of all variables, which were used as case values for modeling and validation in provincial and city level.

Results

The experimental analyses were conducted from two aspects: model calibration and accuracy validation. The performance of two models, the model fusing multi-source data (Model 1), and the model adding land cover and vegetation data as prior knowledge (Model 2), was compared.

Model calibration. To test the prediction success rate of the model, the logarithmic loss, quadratic loss and spherical payoff measures[27] were calculated to evaluate the performance of the models by using the sampling ACDE values as reference values. Their equations are[28]:

$$\text{Logarithmic loss} = M(-\ln S) \tag{2}$$

$$\text{Quadratic loss} = M\left(1 - 2S + \sum_{j=1}^{n} P_j^2\right) \tag{3}$$

$$\text{Spherical payoff} = M\left(\frac{S}{\sum_{j=1}^{n} P_j^2}\right) \tag{4}$$

	Relative error (%)	Logarithmic loss	Quadratic loss	Spherical payoff
Model 1	18.75	2.16	0.31	0.82
Model 2	9.68	0.17	0.12	0.93

Table 3. Overall accuracy measures of the models.

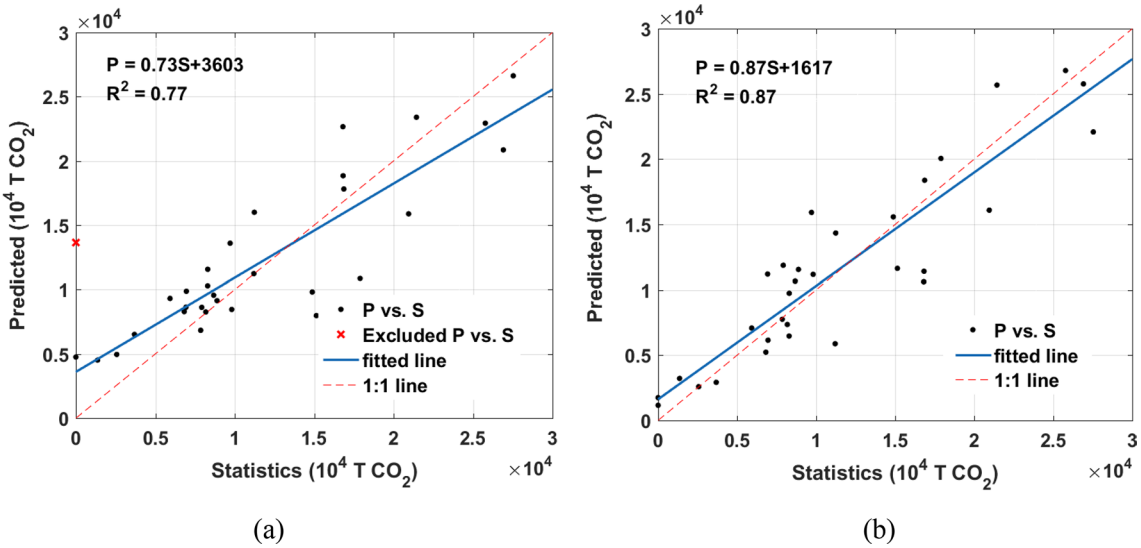


Figure 3. Trend line of the predicted ('P' for short) and the statistics ('S' for short) data at provincial level, (**a**) Model 1, (**b**) Model 2. The figs were generated using the MATLAB software package (The MathWorks, Version 2013, https://www.mathworks.com). The predicted data was summarized to provincial level using zonal statistics tool in ArcGIS. Two outliers were excluded because of the missing value in statistic data.

where M represents the mean probability value of a given state, S is the probability of the correct state for a given class, P_j is the probability for class j and n is the number of states.

I took 2010 data as training data and 2005 data as validation data. The root mean square error (RMSE) and the relative error rate were used to test the accuracy of the model outputs. The relative error rates were 18.75% and 9.68% for the two models respectively (Table 3).

Accuracy validation. A quantitative accuracy comparison with statistical data and the products from FFDAS, EDGAR (the Emission Database for Global Atmospheric Research) and ODIAC (Open-Data Inventory for Anthropogenic Carbon dioxide) were performed. EDGAR is a global-covered 10 km × 10 km resolution emission inventory data developed by EC-JRC/PBL[29] by using point source and road network data in addition to population data. FFDAS is a global fossil fuel CO_2 emissions inventory developed by Rayner et al.[30] by assimilating nightlights data together with population data. ODIAC is a global high-resolution emission data product for carbon dioxide emissions, originally developed under the Greenhouse gas Observing SATellite (GOSAT) project at the National Institute for Environmental Studies (NIES), Japan (http://odiac.org/).

The estimated ACDE were analyzed and compared at provincial, city and pixel level respectively. On provincial scale, the ACDE statistical data in 2005 were used as reference. At city level, FFDAS ACDE product data in 2005 was used as reference to compare the precision of the estimated ACDE. Furthermore, the spatial distributions and qualitative evaluations of the estimated ACDE on local scale comparing with ACDE products data were carried out.

Provincial level. The scatter diagrams of the result predicted by the BN model with or without prior knowledge versus statistics data at provincial level are plotted in Fig. 3.

The proposed BN model predicted ACDE with higher accuracies at provincial level. Furthermore, the Model 2 predicted ACDE better than Model 1 with R^2 of 0.87 because of its full consideration of land cover and vegetation coverage as prior knowledge.

City level. Although built-up areas cover a small percentage of earth surface, they influence vast areas due to the massive energy demands[31], and are main source of ACDE[32]. So we can believe that the accuracy at city level can explain most variations of the BN modeled result.

At city level, the statistic units are not administrative region but are image blocks generated through segmentation of nighttime lights data. Multiresolution segmentation was conducted on DMSP data to generate homogeneous areas using Definiens Developer 7.0. The scale was set to 120, shape *vs* color was set to 0.7:0.3,

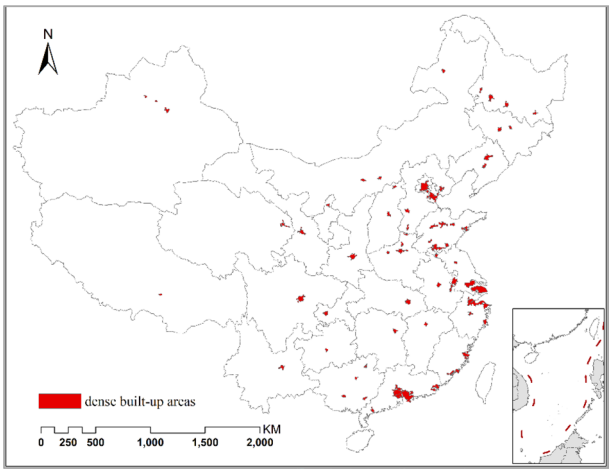


Figure 4. The distribution of the selected homogeneous areas covering dense built-up areas. The fig was generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

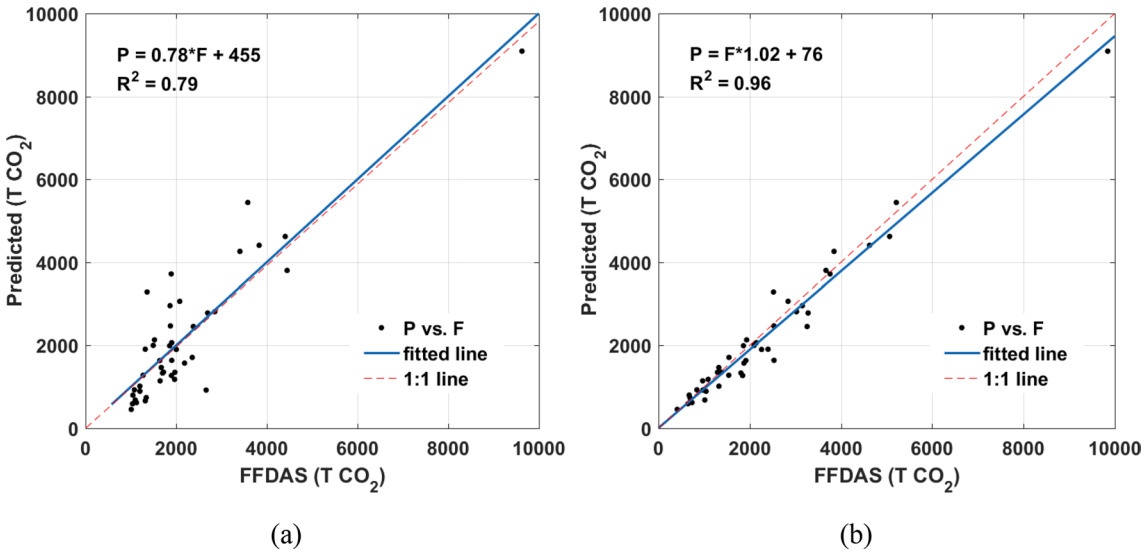


Figure 5. Trend line of the predicted ('P' for short) and FFDAS ('F' for short) data at selected city level, (**a**) Model 1, (**b**) Model 2. The figs were generated using the MATLAB software package (The MathWorks, Version 2013, https://www.mathworks.com).

and compactness *vs* smoothness was set to 0.9:0.1. Those homogeneous areas with mean value greater 10 and sum value greater than 10,000 were selected and merged. These areas cover most of the built-up areas, and also represent most prefectures-above-level cities (Fig. 4). In this sense, this level is a selected city level.

The precisions of Model 1 and Model 2 are evaluated by comparing with FFDAS data at city level (Fig. 5). Since it's hard to collect ACDE statistics data at city level, FFDAS data was used as reference and summarized to city level (considering the borders I defined through image segmentation), although it also has substantial error.

The R^2 of Model 2 reaches 0.96 at confidence level of 95%. Great improvement can be observed from the Model 2 at city level when fusing the land cover and vegetation as prior knowledge.

Pixel level. The map of the modeled ACDE was shown in Fig. 6. The map clearly shows the high ACDE emission areas in China, covering the major metropolitan areas including Beijing-Tianjin-Tangshan, Yangtze River Delta, Pearl River Delta, North China Plain, Central Urban Cluster and Sichuan Basin.

On local scale, abundant details can be observed in ACDE map derived by Model 2, especially great variations in high-density built-up areas (Fig. 7, take Wuhan urban areas as an example).

The modeled ACDE was compared with ACDE products, FFDAS, EDGAR and ODIAC at pixel level (Fig. 8). The BN modeled ACDE maps in major metropolitan areas showed good spatial agreement with ACDE products, at the same time with higher spatial resolution.

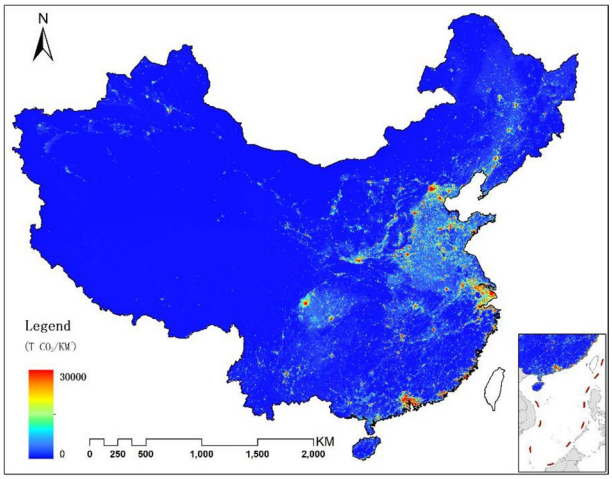


Figure 6. BN modeled ACDE map of mainland China in 2010. The fig was generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

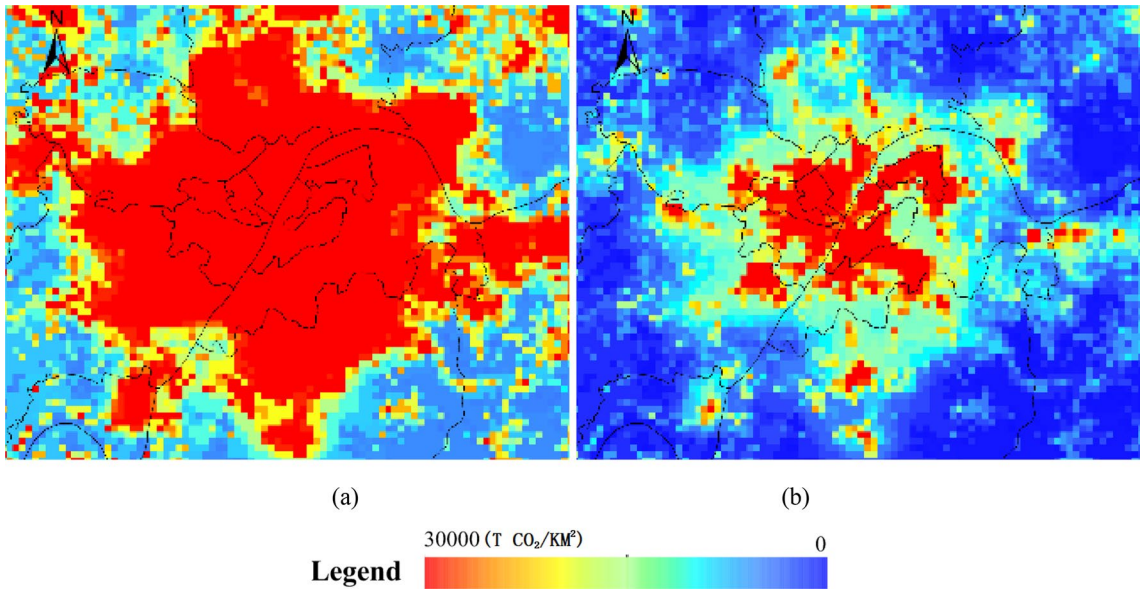


Figure 7. ACDE maps of major metropolitan areas in 2010, (**a**) Model 1, (**b**) Model 2. The figs were generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

Discussions

When conducting land-surface parameter estimation, simply learning a model from a few independents without considering prior knowledge is ill-posed. The existing researches usually utilize regression method to model the relationships between spatial proxies and ACDE, depending heavily on nighttime lights data and population data[12–14,34], without considering model's bias caused by spatial proxies such as DMSP data. The prior knowledge that ACDE was dominated by human activities and has strong correlation with land cover and vegetation condition was usually ignored in conventional methods. The proposed model mitigated the limitations in these spatial proxies to some extent by introducing geo knowledge into earth observation, providing a new way in land-surface parameters inversion.

Predicting ACDE from multi-source data involve the ability of model to combine evidence from remote sensing observations with prior knowledge. Different from conventional statistical methods such as regression, BN can model not only conditional dependence, but also causation relationship between variables. The causal relationships between ACDE and land cover and vegetation were added as prior knowledge to the original BN model. The model structure was very simple and can be explained because the only adaption to a conventional BN model was adding a node into the model to represent the prior knowledge. The model was then locally trained and at the same time globally optimal[33].

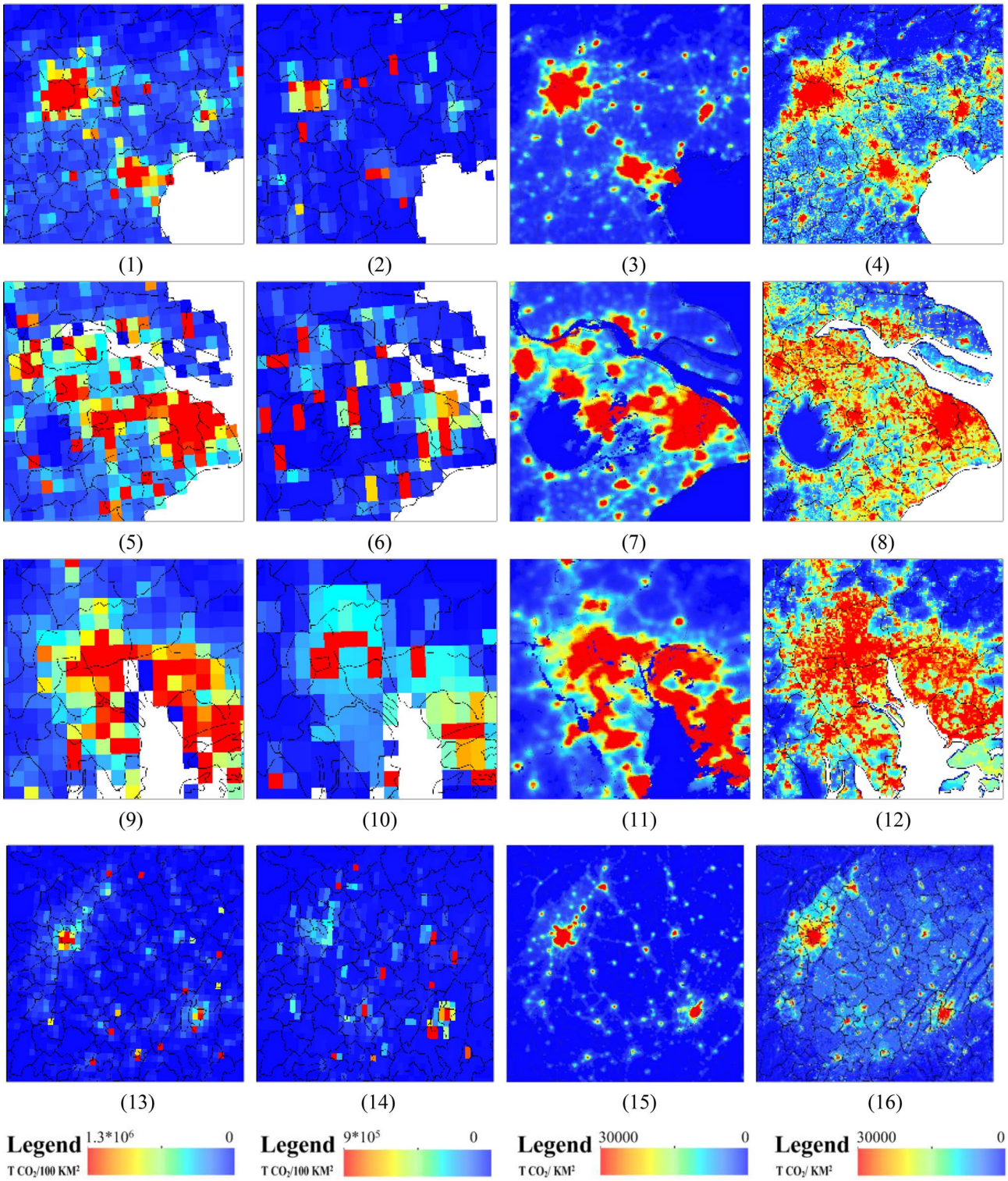


Figure 8. BN modeled ACDE maps comparing with ACDE products in major metropolitan areas of mainland China in 2010, (1) Beijing-Tianjin-Tangshan FFDAS, (2) Beijing-Tianjin-Tangshan EDGAR, (3) Beijing-Tianjin-Tangshan ODIAC, (4) Beijing-Tianjin-Tangshan BN modeled, (5) Yangtze River delta FFDAS, (6) Yangtze River delta EDGAR, (7) Yangtze River delta ODIAC, (8) Yangtze River delta BN modeled, (9) Pearl River delta FFDAS, (10) Pearl River delta EDGAR, (11) Pearl River delta ODIAC, (12) Pearl River delta BN modeled, (13) Sichuan Basin FFDAS, (14) Sichuan Basin EDGAR, (15) Sichuan Basin ODIAC, (16) Sichuan Basin BN modeled. The figs were generated using the ArcGIS Desktop (ESRI, Inc, Version 10.2, https://desktop.arcgis.com/zh-cn/).

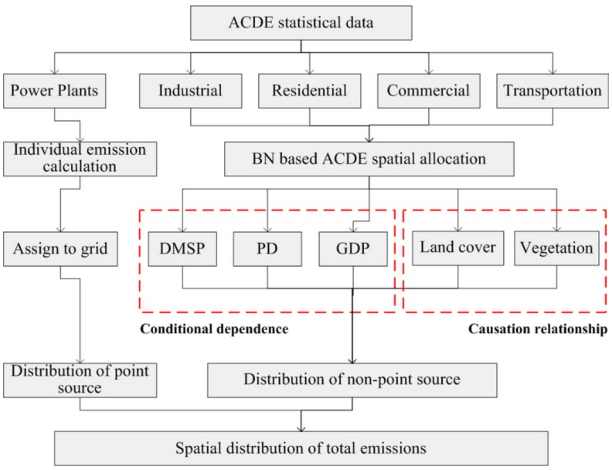


Figure 9. Flow chart of spatial allocation of ACDE statistical data.

The proposed BN model can be applied to high accurate mapping of ACDE on national scale successfully. Furthermore, the model can be applied to any scale land-surface parameters mapping, provided the relationships between dependent variable and independent variables can be modeled by CPT and prior information about the dependent variable can be obtained.

Information loss sourced from data preprocessing method of BN is one limitation of the study. The model has relatively high accuracies in urban and built-up areas, however poor accuracy in rural areas. This is partly because the ACDE value was overestimated in rural areas, and the summation of massive rural pixels pushed up the total ACDE value. The overestimation was partly due to the discrete of variables in Netica, which cause the BN to lose statistics accuracy to some extent[35]. If the information loss caused by Netica can be avoided, the performance could be further improved.

Conclusions

Information on the extent and spatial details of ACDE is a key requirement for understanding how human activities affect climate change. This paper reports my work on spatial allocation of ACDE statistical data in mainland China using Bayesian Network. Bayesian Network combines the robustness of probability theory with the expressiveness of graphs, offering a mechanism for handling prior knowledge in estimating land-surface parameters. A spatial allocation BN model for non-point source ACDE statistics data fusing multi-source data was developed. The BN model predicted ACDE with R-square of 0.87 at provincial level and R-square of 0.96 at city level and great improvement can be observed from the proposed model (Model 2) when fusing the land cover and vegetation as prior knowledge. The study is of great value to the development of spatial allocation method for ACDE statistics data on national scale.

The contribution of this research is that it provides a novel way of combining DMSP, PD, GDP data and land cover and vegetation data and can be applied to large-scale ACDE mapping. BN model can model not only conditional dependence between ACDE and proxies, but also causation relationship between ACDE and land covers and vegetation well. The method has the superiority of fusing land cover and vegetation data as prior knowledge into BN model, mitigating the limitations in conventional spatial proxies greatly. Conventional approaches for downscaling statistical data into grid level based on nightlight and population data have certain limitations. BN model can therefore provide a good solution to spatial allocation of ACDE statistical data.

Methods

ACDE gridded data were estimated by combining emissions from point source and non-point source. ACDE from point source were calculated separately using the CARMA power plants database. Emissions form power plants were spatially allocated to the exact grid indicated by CARMA. Emissions from non-point source (including industrial, residential, commercial and transportation sectors) were obtained approximately by subtracting the emissions from point source, and then allocated to grids using Bayesian Network by fusing multi-source data as proxies. Finally, emissions form these two sources were integrated by summarizing at 1 km grid level.

By utilizing the advantages of BN, a spatial allocation model for ACDE statistics data (from non-point source) fusing multi-source data was made. Here I present the method of fusing multi-source data using BN (modeling conditional dependence between variables), and adding prior knowledge to the model (modeling causation relationship between variables). The flow chart of spatial allocation was presented in Fig. 9. The software packages used for modeling and the subsequent calibration and validation were Netica 5.02 and Matlab.

Bayesian network. Bayesian Network is a powerful mathematical tool combing the advantages of probability theory and graph theory to reason about uncertainty in land-surface parameters estimation. A Bayesian Network (BN) is a DAG (directed acyclic graph) combined with a CPT (conditional probability table), in which

each node represents a random variable and the arcs linking the nodes represent relationships between variables. In a BN the joint probability distribution of all random variables can be decomposed as the product of a series of probabilities each representing a subset of the variables. Therefore the joint probability distribution of a BN with n nodes can be calculated as[36]:

$$p(X) = \sum_{i=1}^{n} p\left(x_i | pa_i\right), \tag{5}$$

where x_i is a random variable, pa_i denotes the parent node sets of x_i, and $X = \{x_1, \ldots, x_n\}$.

Naïve Bayesian Network (NB for short) is the simplest type of Bayesian Network in which class node is the parent node of all other feature nodes and the features are assumed to be independent of each other. Therefore on the basis of conditional independence hypothesis the joint probability of all nodes is[37]:

$$P(X_1, X_2, \ldots, X_n, C) = P(C) \prod_{i=1}^{n} P(X_i | C), \tag{6}$$

where X_1, X_2, …, X_n represent the features and C represents the class variable.

The Bayesian Network model has the advantage of having bidirectional inference capability which enables probability to be propagated forward or backward through nodes[38].

Fusing multi-source data. The spatial distributions of ACDE are highly correlated with human activities and land-cover conditions. As spatial proxies of ACDE, the independent variables include: (1) DMSP nighttime lights data; (2) socioeconomic data such as PD and GDP data; (3) land cover and vegetation data.

A Bayesian Network model was constructed to model the conditional dependence between ACDE (dependent variable) and multi-source data (independent variables). In this research, I used term variable to indicate any features. Since we know the positive relationships between DMSP, PD, GDP and ACDE, the link was added manually to construct the DAG. The initial Bayesian network was a NB which included three child nodes (Fig. 10a).

The continuous variables need to be converted to discrete quantities before any probabilistic inference because all of probabilistic inference in Netica is done with discrete tables[39]. Jenks Natural Breaks[40] was used to discrete all the variables. After a series of tests (initializing 3–15 states) to figure out how many states can improve the overall accuracy, ten states were finally selected. The CPT of the model was estimated using the EM (expectation-maximization) algorithm. The maximum number of iterations was set to 500 to ensure convergence.

When the CPTs of each node have been determined, the network is 'solved'[40], as shown in Fig. 10a. The observations of the independent can then easily be interpreted by individual case. The effect of entering an evidence on one node can be examined by the response of other nodes through propagation, as illustrated in Fig. 10b, c. The rapid information propagation through the nodes make us quickly observe how conditions at one node will affect the whole network[40]. The BN is complete and can then be applied to validation and prediction of the samples after evaluation tests.

Since the relationship between independents and ACDE remain constant when change the scale from district polygon to pixel grid, the BN model can allocate ACDE statistics data to pixel level spatially. Because I can only get a continuous and complete official data at provincial level, the model was built at this level (forward), and then backward to pixel level.

In the inference stage the probability values with maximum posterior probabilities were selected as the predicted ACDE values. The Bayesian Network model can then be run on every pixel in the research area.

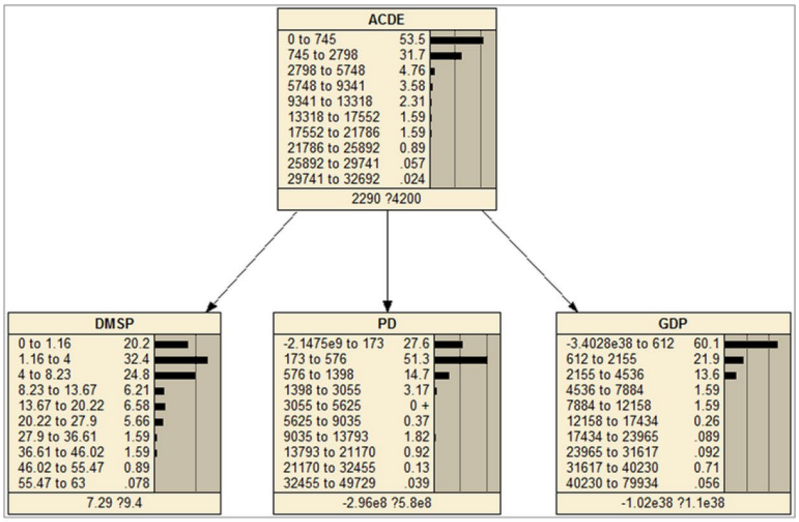
Adding land cover and vegetation data as prior knowledge. Here the prior knowledge refers to the knowledge about the distribution of ACDE in addition to the probability distribution function from DMSP, GDP and PD data through training. The basis is that simply determining a model from training samples without prior knowledge is ill-posed. According to Bayes' theorem, if X and Y are two variables, with their probabilities being $P(X)$ and $P(Y)$ respectively, then

$$P(X|Y) = P(Y|X)P(X)/P(Y) \tag{7}$$

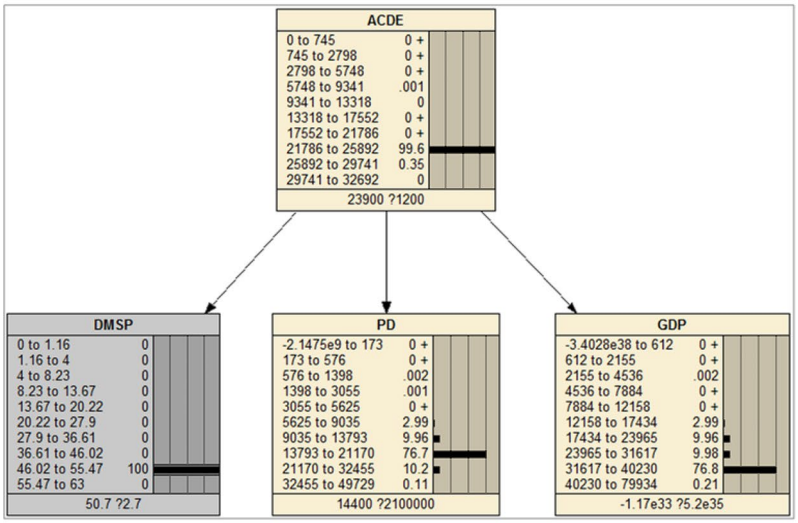
where $P(X|Y)$ is posterior probability (the probability of X given Y), and $P(Y|X)$ is prior probability (the probability of Y given X). The prior probability is useful because it gives us important information about the predicted variable which cannot obtain from the training data.

Land covers document the places where human activities take place and are source of ACDE, reflecting the intensity of human activities. General speaking, ACDE usually take place in built-up areas, especially in dense urban areas. At the same time, we cannot expect significant ACDE at high-level vegetation-covered areas. To see if BN can model this causation relationship, a statistics analysis of total ACDE from FFDAS data and total built-up areas from MCD12Q1 land-cover data was conducted to explore the relationship between ACDE and land-cover types. Figure 11 shows the proportions of ACDE caused by every land-cover types. We can observe a varied proportion between them, within which built-up areas ranks top one. This percentage can be interpreted as probability of ACDE level of a pixel, given that we know the land-cover type of that pixel.

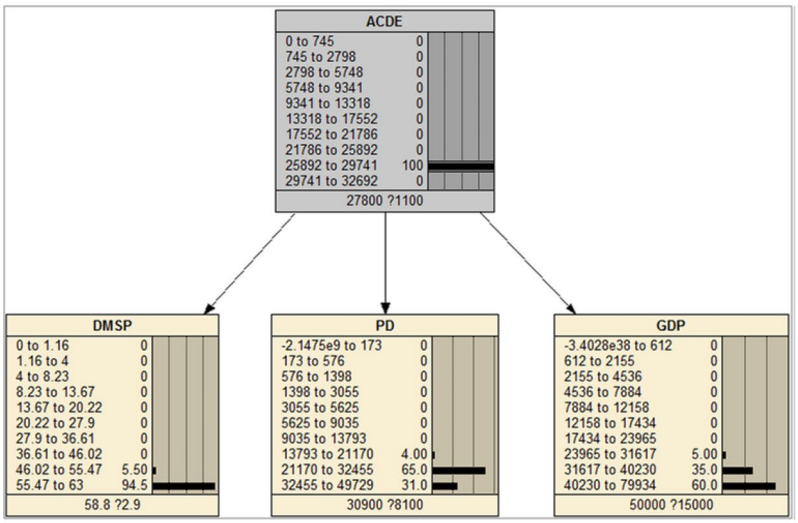
The land cover and vegetation data were fused into the initial model (named as Model 1) as prior knowledge to generate the final model (Model 2). The original MODIS EVI data was reclassified into five classes using Jenks Natural Breaks method which represent different levels of vegetation coverage (Table 4). Unlike the training data such as nighttime lights having continuous measurement, the land cover and vegetation data is a kind of data with nominal measurement, which can't be modeled as a natural node in BN and can't learning CPT through



(a)



(b)



(c)

Figure 10. The Naïve Bayesian Networks and the CPTs of the nodes. (**a**) CPTs when learning the network, (**b**) CPTs when giving an evidence of DMSP, (**c**) CPTs when giving an evidence of ACDE. The figs were generated using the Netica software (Norsys Software Corp, Version 5.02, https://www.norsys.com/).

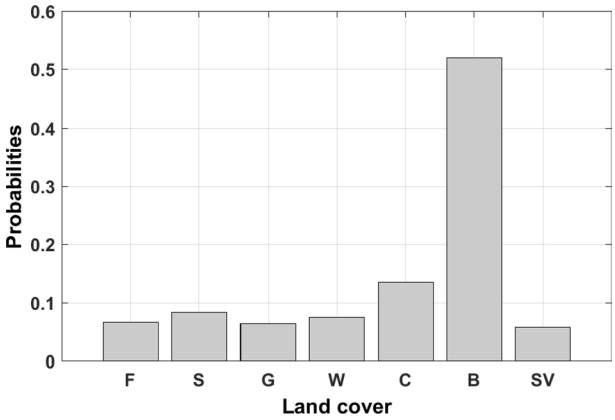


Figure 11. The statistics analysis of the FFDAS data compared with land-cover data. The horizontal axis indicates land-cover types: F-forests, S-shrublands, G-grasslands, W-water and wetlands, C-croplands, B-built-up areas and SV-sparsely vegetations.

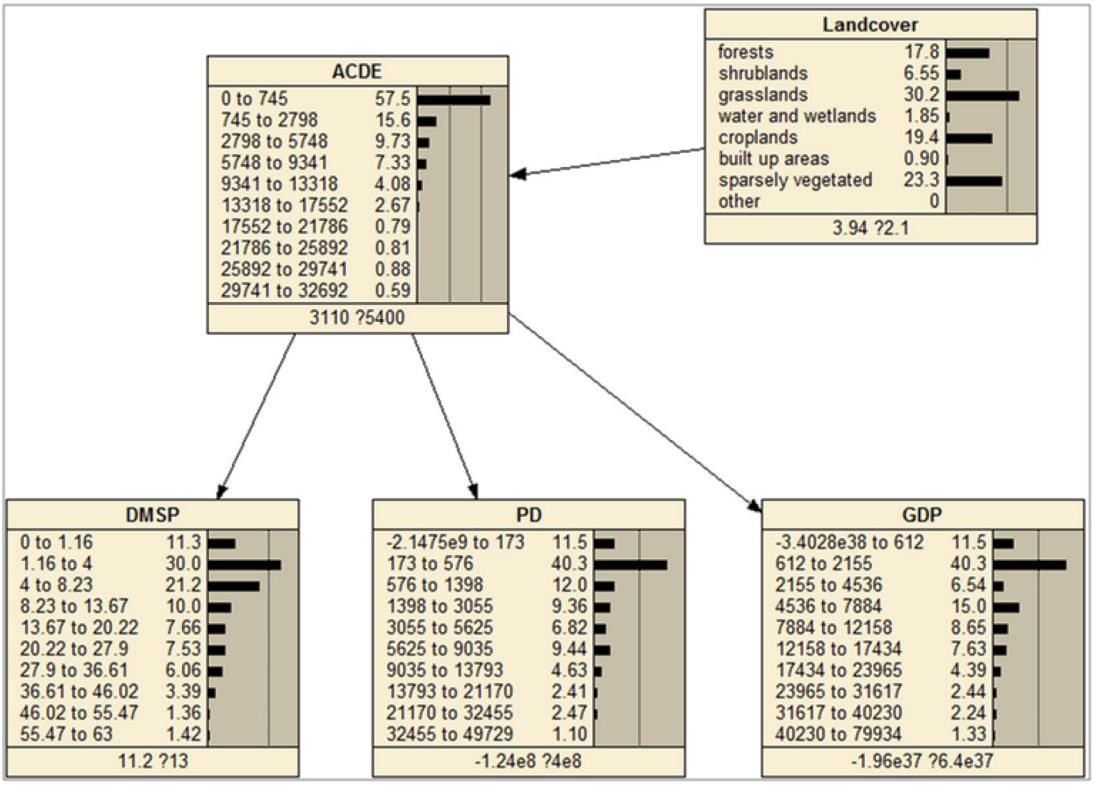
EVI value	Class label	Summary class
0–1058	1	Very low
1059~2431	2	Low
2432–3632	3	Medium
3633–4961	4	High
4962–10,000	5	Very high

Table 4. Summary classes of vegetation.

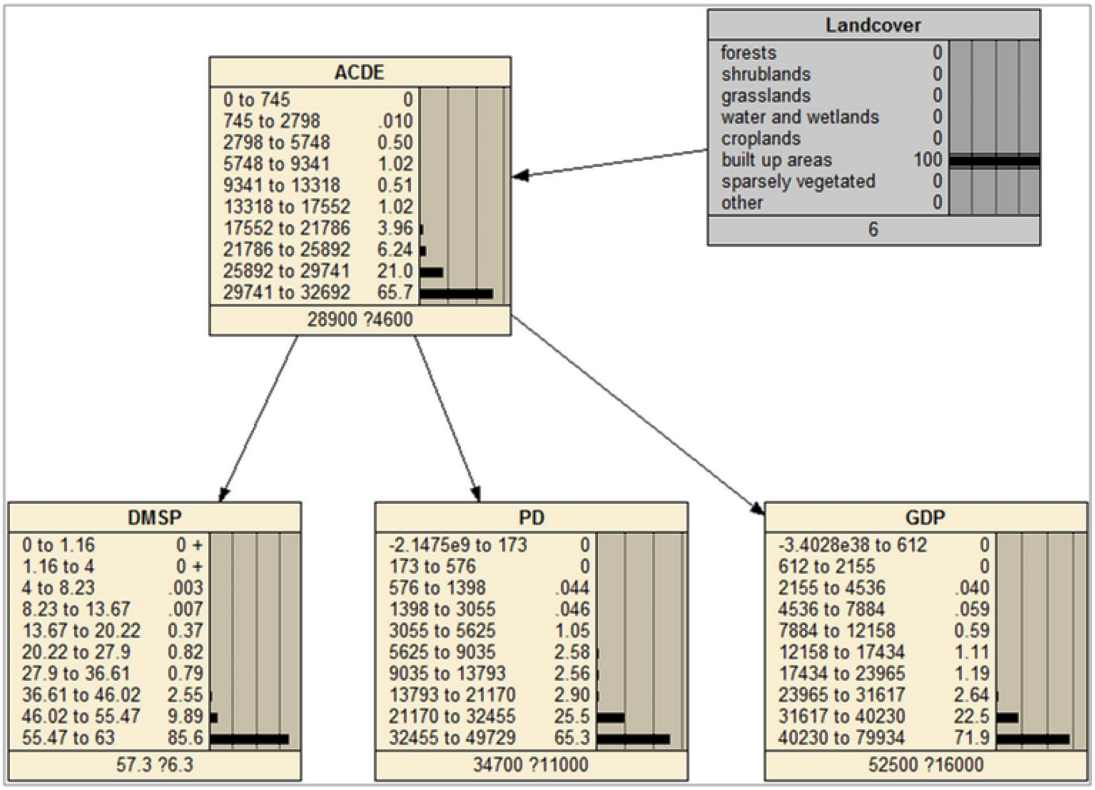
training. The CPTs of land cover and vegetation nodes were input manually according to above statistics analysis. Figures 12 and 13 give the DAGs and CPTs of Model 2.

The CPT of land-cover node came from the statistics of FFDAS data. After adding this node, all the CPTs in other nodes update automatically. Most places have relative low level ACDE value, as well as DMSP, POP and GDP values, which can be observed from the CPTs. This is consistent with the fact that human activities and ACDE exist in relative low percent of the total land covers (Fig. 12a). If give an evidence of land cover type as built-up at 100% probability, a high level ACDE can be observed, and all the other nodes update the CPT through propagation of probabilities. If land cover tells it is built-up, it is very likely (more than 65% probability) it has high level ACDE, even if other three independents observe low values (Fig. 12b). These facts demonstrated that BN can model the causation relationship of ACDE and land covers well.

The CPT of vegetation node came from the statistics of EVI data. After adding this node, all the CPTs in other nodes update automatically (Fig. 13a). If give an evidence of vegetation condition as "very high" at 100% probability, a low level ACDE can be observed, although giving an evidence of DMSP as high at 100% probability (Fig. 13b). This seemingly contradictory can be explained by the saturation problem of DMSP in dense built-up areas. These facts demonstrated that BN can model the causation relationship of ACDE and vegetation well.

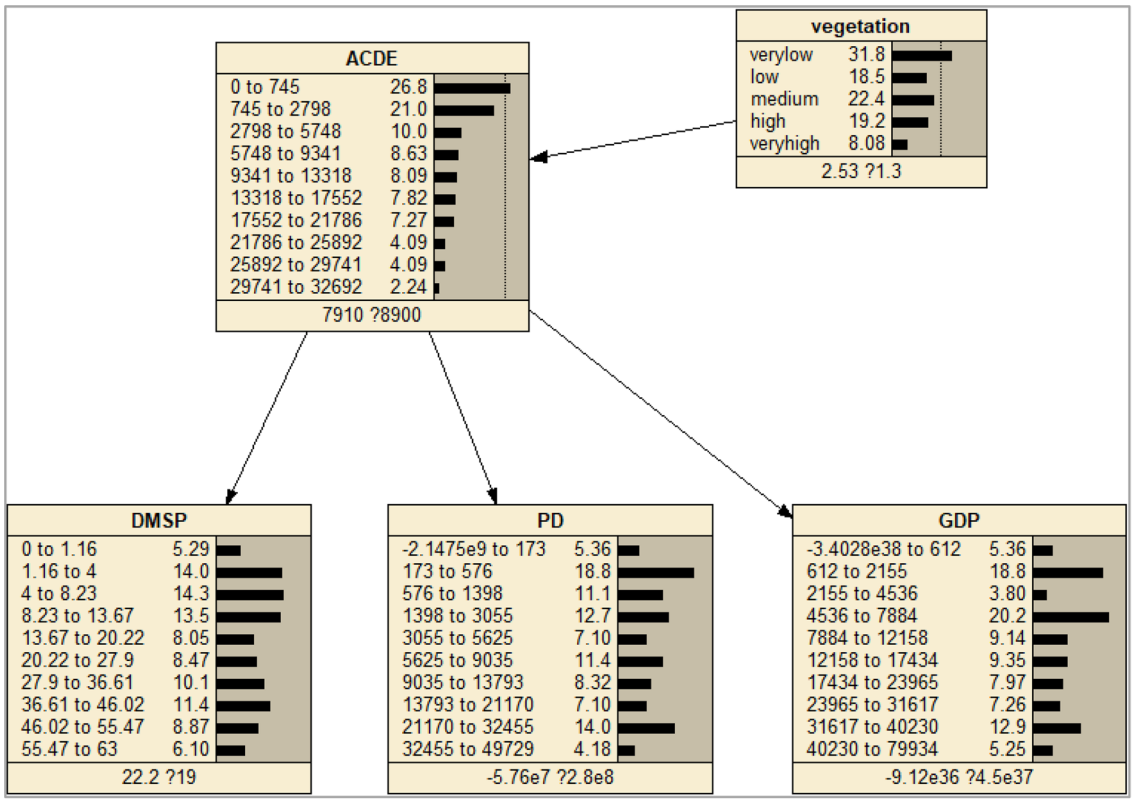


(a)

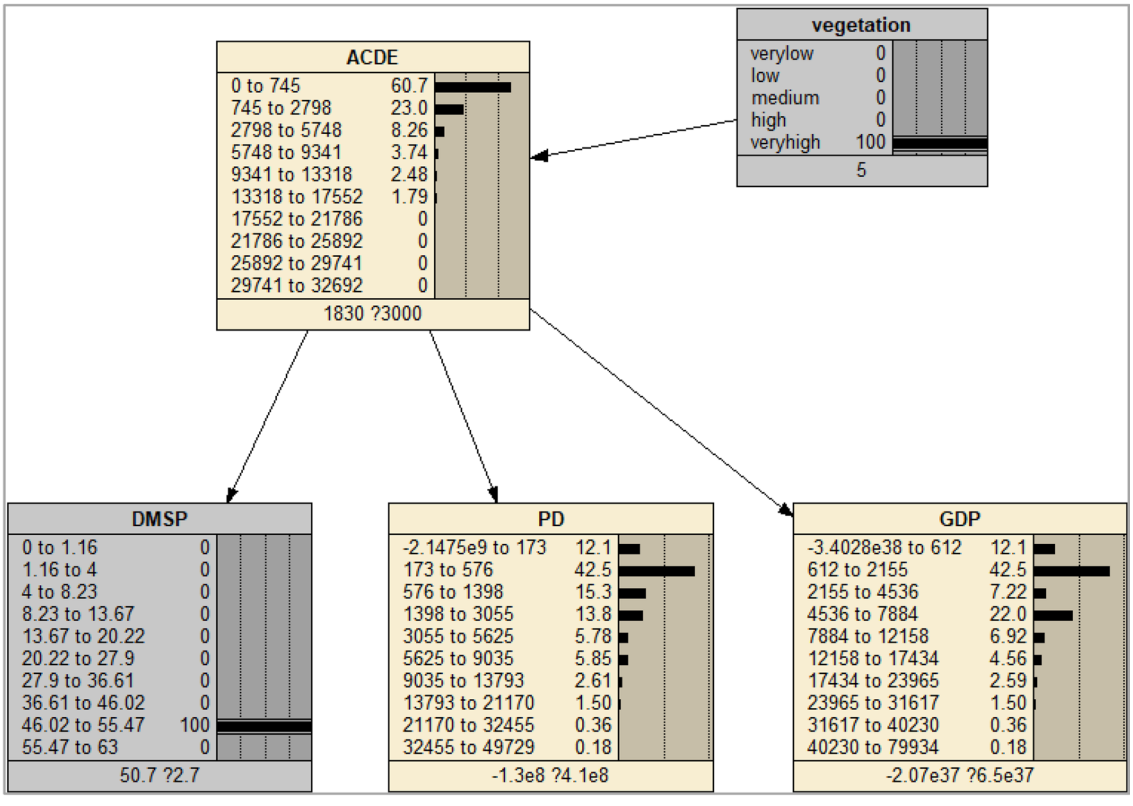


(b)

Figure 12. The Naïve Bayesian Network fusing land-cover data as prior knowledge. (**a**) CPTs when adding a land-cover node as the prior knowledge, (**b**) CPTs when give an evidence of land cover. The figs were generated using the Netica software (Norsys Software Corp, Version 5.02, https://www.norsys.com/).



(a)



(b)

Figure 13. Naïve Bayesian Network fusing vegetation data as prior knowledge. (**a**) CPTs when adding a vegetation node as the prior knowledge, (**b**) CPTs when give an evidence of vegetation. The figs were generated using the Netica software (Norsys Software Corp, Version 5.02, https://www.norsys.com/).

Received: 9 April 2020; Accepted: 22 February 2021
Published online: 13 September 2021

References

1. Church, J. A., Clark, P. U., Cazenave, A., Gregory, J. M., Jevrejeva, S., Levermann, A., Merrifield, M. A., Milne, G. A., Nerem, R. S. & Nunn, P. D. Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Sea Level Change, 1137-1216 (2013).
2. UNEP, G. Global Environment Outlook 6. Report GEO-6. (UNEP, 2019) (2019).
3. Stocker, T. *Climate change 2013: The physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2014).
4. McGee, M. Mauna Loa Observatory, Hawaii (Scripps) Preliminary data See Supplemental Material at https://www.co2.earth/.released Nov 6, 2019. (2019).
5. Wang, S., Huang, Y. & Zhou, Y. Spatial spillover effect and driving forces of carbon emission intensity at the city level in China. *J. Geogr. Sci.* **29**(02), 231–252 (2019).
6. Olivier, J., Janssens-Maenhout, G., Muntean, M. & Peters, J. Trends in global CO2 emissions: 2018 Report. PBL Netherlands Environmental Assessment Agency, The Hague; European Commission, Joint Research Centre (JRC). Institute for Environment and Sustainability (IES), JRC Report (2018).
7. Liu, L., Chen, C., Zhao, Y. & Zhao, E. China's carbon-emissions trading: Overview, challenges and future. *Renew. Sustain. Energy Rev.* **49**, 254–266 (2015).
8. Green, F. & Stern, N. China's changing economy: Implications for its carbon dioxide emissions. *Clim. Policy* **17**(4), 423–442 (2017).
9. Wu, J., Wang, X., Wang, C., He, X. & Ye, M. The status and development trend of disaggregation of socio-economic data. *J. Geo Inform. Sci* **20**, 1252–1262 (2018).
10. Goodchild, M. F., Anselin, L. & Deichmann, U. A framework for the areal interpolation of socioeconomic data. *Environ. Plan. A* **25**(3), 383–397 (1993).
11. Malone, P., McBratney, A. B., Minasny, B. & Wheeler, I. A general method for downscaling earth resource information. *Comput. Geosci.* **41**, 119–125 (2012).
12. Oda, T. & Maksyutov, S. A very high-resolution (1 km× 1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmos. Chem. Phys.* **11**(2), 543–556 (2011).
13. Ghosh, T. *et al.* Creating a global grid of distributed fossil fuel CO_2 emissions from nighttime satellite imagery. *Energies* **3**(12), 1895–1913 (2010).
14. Wang, J. *et al.* High resolution carbon dioxide emission gridded data for China derived from point sources. *Environ. Sci. Technol.* **48**(12), 7085–7093 (2014).
15. Raupach, M., Rayner, P. & Paget, M. Regional variations in spatial structure of nightlights, population density and fossil-fuel CO_2 emissions. *Energy Policy* **38**(9), 4756–4764 (2010).
16. Ou, J., Liu, X., Li, X., Li, M. & Li, W. Evaluation of NPP-VIIRS nighttime light data for mapping global fossil fuel combustion CO_2 emissions: A comparison with DMSP-OLS nighttime light data. *PloS One* **10**(9), e0138310 (2015).
17. Hogue, S., Marland, E., Andres, R. J., Marland, G. & Woodard, D. Uncertainty in gridded CO2 emissions estimates. *Earth's Fut.* **4**(5), 225–239 (2016).
18. Zhang, X., Wu, J., Peng, J. & Cao, Q. The uncertainty of nighttime light data in estimating carbon dioxide emissions in China: A comparison between DMSP-OLS and NPP-VIIRS. *Remote Sens.* **9**(8), 797 (2017).
19. Xu, Y., Xu, X. & Tang, Q. Human activity intensity of land surface: Concept, method and application in China. *J. Geogr. Sci.* **9**, 1068–1079 (2016).
20. Pozzi, F. & Small, C. Analysis of urban land cover and population density in the United States. *Photogram. Eng. Remote Sens.* **71**(6), 719–726 (2005).
21. Zhang, Q., Schaaf, C. & Seto, K. C. The vegetation adjusted NTL urban index: A new approach to reduce saturation and increase variation in nighttime luminosity. *Remote Sens. Environ.* **129**, 32–41 (2013).
22. Imhoff, M. L., Lawrence, W. T., Stutzer, D. C. & Elvidge, C. D. A technique for using composite DMSP/OLS "city lights" satellite data to map urban area. *Remote Sens. Environ.* **61**(3), 361–370 (1997).
23. Friedl, M. A. *et al.* MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **114**(1), 168–182 (2010).
24. Didan, K., Munoz, A. B., Solano, R. & Huete, A. MODIS vegetation index user's guide (MOD13 Series). Vegetation Index and Phenology Lab, The University of Arizona, 1–38 (2015).
25. Jönsson, P. & Eklundh, L. TIMESAT—A program for analyzing time-series of satellite sensor data. *Comput. Geosci.* **30**(8), 833–845 (2004).
26. IPCC, R.: IPCC guidelines for national greenhouse gas inventories. Kanagawa, Japan (2006).
27. Morgan, M. G., Henrion, M. & Small, M. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis* (Cambridge University Press, 1992).
28. Dlamini, W. M. Application of a Bayesian network for land-cover classification from a Landsat 7 ETM+ image. *Int. J. Remote sens.* **32**(21), 6569–6586 (2011).
29. EC-JRC/PBL. Emission Database for Global Atmospheric Research (EDGAR), release version 4.0. http://edgar.jrc.ec.europa.eu (2009).
30. Rayner, P., Raupach, M., Paget, M., Peylin, P. & Koffi, E. A new global gridded data set of CO2 emissions from fossil fuel combustion: Methodology and evaluation. *J. Geophys. Res. Atmos.* https://doi.org/10.1029/2009JD013439 (2010).
31. Pacifici, F., Chini, M. & Emery, W. J. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sens. Environ.* **113**(6), 1276–1292 (2009).
32. Bagan, H. & Yamagata, Y. Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GISci. Remote Sens.* **52**(6), 765–780 (2015).
33. Tao, J., Wu, W. & Xu, M. Using the Bayesian network to map large-scale cropping intensity by fusing multi-source data. *Remote Sens.* **11**(2), 168 (2019).
34. Li, H., Long, M. & Li, G. Spatial-temporal dynamics of carbon dioxide emissions in China based on DMSP/OLS nighttime stable light data. *China Environ. Sci.* **38**(7), 2777–2784 (2018).
35. Friedman, N. & Goldszmidt, M. Discretizing continuous attributes while learning Bayesian networks. In: ICML 1996, 157–165 (1996)
36. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **29**(2–3), 131–163 (1997).
37. Cooper, G. F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9**(4), 309–347 (1992).
38. Kalácska, M., Sánchez-Azofeifa, G. A., Caelli, T., Rivard, B. & Boerlage, B. Estimating leaf area index from satellite imagery using Bayesian networks. *IEEE Trans. Geosci. Remote Sens.* **43**(8), 1866–1873 (2005).
39. Norsys. http//www.norsys.com (2016)

40. Pollino, C., Henderson, C. Bayesian networks: A guide for their application in natural resource management and policy. Landscape Logic, Technical Report 14 (2010).

Acknowledgements
This work was finally supported by the National Natural Science Foundation of China (Grant Nos. 41971371 and 61501200), the National Key R&D Program of China (2017YFC0504501) and the Fundamental Research Funds for the Central Universities of Central China Normal University (CCNU19TS004).

Author contributions
J.T. drafted and revised the manuscript. X.K. took part of experiment and validation work. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to X.K.

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.