Article

Urban Analytics and City Science

EPB: Urban Analytics and City Science
2021, Vol. 48(7) 1876–1894
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2399808320951580
journals.sagepub.com/home/epb
SAGE

Mapping China's regional economic activity by integrating points-of-interest and remote sensing data with random forest

Qian Chen and Tingting Ye
Ocean College, Zhejiang University, China

Naizhuo Zhao
McGill University, Canada

Mingjun Ding
Jiangxi Normal University, China

Zutao Ouyang
Stanford University, USA

Peng Jia
University of Twente, the Netherlands; International Initiative on Spatial Lifecourse Epidemiology (ISLE), the Netherlands

Wenze Yue
Zhejiang University, China

Xuchao Yang
Ocean College, Zhejiang University, China

Abstract

Nighttime light imageries are widely used for mapping the gross domestic product (GDP) over large areas. However, nighttime light imagery is inappropriate to disaggregate agricultural GDP and inadequate to differentiate the GDP from the secondary and tertiary sectors. Points-of-

Corresponding author:
Xuchao Yang, Ocean College, Zhejiang University, Zhoushan Campus, Haike Building 357, 1 Zheda Road, Zhoushan 316021, China.
Email: yangxuchao@zju.edu.cn

interest, a kind of geospatial big data with geographic locations and textual descriptions of the category, can effectively distinguish industrial and commercial areas, and therefore have the potential to improve the precise GDP mapping from secondary and tertiary sectors. In this study, a machine learning method, random forest, was used to disaggregate the 2010 county-level census GDP data of mainland China to 1 km × 1 km grids. Six Random Forest models were constructed for different economic sectors to explore the non-linear relationships between various geographic predictors and GDP from different sectors. By fusing points-of-interest of varying categories, the spatial distribution of economic activities from the secondary and tertiary sectors was effectively distinguished. Compared to previous studies, the strategy of developing specific Random Forest models for different sectors generated a more reasonable distribution of GDP. Our results highlight the feasibility of using point-of-interest data in disaggregating non-agricultural GDP by exploiting the complementary features of the different data sources.

Keywords

GDP, random forest, points-of-interest, remote sensing, China

Introduction

Gross domestic product (GDP) is the most widely used measure of economic activity (Geiger, 2018; Henderson et al., 2012). GDP is usually collected at the administrative unit level, such as province/state and county. Therefore, it is difficult to match GDP data with other fine-scale environmental data for conducting cross-disciplinary research. Different schemes of data collection have also been recognized as one of the major obstacles to the widespread integration of the social and natural sciences (Zandbergen and Ignizio, 2010). It is crucial that GDP data can be spatialized into a fine-scale so that data from different disciplines can be easily integrated (Jia et al., 2015). The gridded GDP data offer the flexibility to perform analysis at various spatial units, such as spatially explicit exposure assessment (Dasgupta et al., 2011; Geiger et al., 2018; Paprotny et al., 2018), and may enable policy planning for the reduction of economic inequality (Wang et al., 2019a), in a more convenient way.

In many developing countries, there is usually a paucity of reliable GDP data. The satellite-derived nighttime lights (NTL) data from the Defense Meteorological Satellite Program's Operational Linescan System (DMSP/OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS) have been proven to correlate well with GDP and are a good proxy for economic activities at all examined scales (Bennett and Smith, 2017; Ma et al., 2012; Wu et al., 2013). Elvidge et al. (1997) first reported a strong correlation between satellite observed area lit and economic activity for 21 countries. Economic researchers have also demonstrated that NTL data are a good supplementary metric to the conventional measures of GDP (Chen and Nordhaus, 2011; Henderson et al., 2012). Therefore, the NTL data have been increasingly used to estimate the spatiotemporal dynamics of GDP at different spatial scales. Using the DMSP/OLS NTL dataset, Doll et al. (2000) produced the first-ever map of Purchasing Power Parity-GDP of the world on a 1° × 1° grid and Sutton and Costanza (2002) generated the first global high-resolution (1 km) map of estimated GDP. A study by Doll et al. (2006) generated economic activity maps at 5 km resolution for the United States and European Union countries.

China has been experiencing rapid economic development since the late 1970s and became the world's second largest economy in 2010. The DMSP/OLS NTL imagery has been utilized to spatially disaggregate GDP (Zhao et al., 2017b, 2011) and estimate the poverty level across China (Wang et al., 2012). The VIIRS data showed a stronger capacity to model the regional economy in China than the DMSP/OLS NTL data, especially at levels such as prefecture or county (Dai et al., 2017; Li et al., 2013; Shi et al., 2014; Zhao et al., 2017a). Also, both DMSP/OLS and VIIRS data have been integrated to model long-term GDP dynamics in China (Zhu et al., 2017).

Although the use of NTL data as a proxy of GDP distribution has recently received growing attention across the world, this method remains problematic in some aspects. Firstly, agricultural activities usually occur in areas that emit marginal or no NTL. Therefore, NTL data are not a good proxy for disaggregating GDP from the agricultural sector, which comprises a large proportion of the national economy in many developing countries (Ghosh et al., 2010; Keola et al., 2015). Secondly, disaggregating GDP based solely on NTL data often suffers from the coarse spatial resolution, the blooming effect, and signal saturation in urban centers, especially for the DMSP/OLS data. For example, the lit area from DMSP/OLS is much larger than actual urban area due to the blooming effect (Liu et al., 2016; Small et al., 2005). Only 2.3% of the lit area is built-up land in China while cropland, grassland, and forests are the major land cover types in the lit area (Liu et al., 2016). Finally, it is impossible to use only NTL as ancillary data to accurately disaggregate GDP from secondary and tertiary sectors separately. Since medium-resolution NTL images are not directly indicative of economic activities, it is difficult to extract detailed socioeconomic features from NTL images, especially in complex urban areas (Liu et al., 2015). Highly urbanized regions typically contain a mix of commercial, business, industrial, and residential areas and infrastructure, affording different types of socio-economic activities. Such complex urban environments present a great challenge to interpreting economic activities on the basis of only NTL data.

Recently, emerging geospatial big data, such as points-of-interest (POI), provide a wealth of useful information about human activities and offer unprecedented opportunities to reveal the distribution of economic activities from a more refined spatial perspective (Gao et al., 2017; Zhao et al., 2018). POIs record the spatial and attribute information of geographic entities, such as geographic coordinates, short textual descriptions, and functional categories (McKenzie et al., 2015; Yoshida et al., 2010). POI data provide inspiring insights about the spatial distribution of population and economic activities, and therefore, they have been increasingly used to extract information in relation to urban functional regions (Gao et al., 2017; Yang et al., 2015), urban land use mapping (Hu et al., 2016; Jiang et al., 2015; Liu et al., 2017; Yao et al., 2017), and population mapping (Bakillah et al., 2014; Yang et al., 2019; Ye et al., 2019). Since certain types of entities may be associated with certain economic activities, POIs can be useful for improving GDP estimates if relevant correlations can be established and are particularly promising for disaggregating non-agricultural GDP. POIs with different categories that are highly related to various human economic activities could effectively differentiate industrial districts from service sector activity. Recently, Yang et al. (2019) and Ye et al. (2019) proved that the integration of commercial POI data and multisource remote sensing data can further refine population estimates across large geographic scales. Therefore, there is strong potential to integrate POI with remote sensing data to gain better insights into the distribution of GDP over large areas.

Our study aimed to adopt a machine learning method, random forest (RF), that can fuse the information from multi-sensor remote sensing data and POIs to disaggregate GDP from different sectors (using GDP_1, GDP_2, and GDP_3 for the production of the primary sector, the secondary sector, and the tertiary sector, respectively) in China from the county level to 1 × 1 km grids. To our knowledge, this is the first study to disaggregate GDP using a machine learning method on the basis of remote sensing and geospatial big data in China.

Data collection and pre-processing

Census data

This study takes mainland China as its case study area. Taiwan, Hong Kong, and Macao were excluded due to their distinct political and economic status from mainland China. Census data of GDP_1 (including farming, forestry, livestock, and fishery), GDP_2, and GDP_3 at the county level in China (with 1200, 2495, and 2495 units, respectively) were collected from the China Statistical Yearbook for Regional Economy published in 2011. All census GDP data for 2010 from different sectors were spatially joined to the corresponding GIS-based administrative boundaries.

Remote sensing datasets

1. *NTL data*. The radiance calibrated NTL images from the DMSP/OLS were downloaded from the National Geophysical Data Center (https://ngdc.noaa.gov/eog/dmsp/download_radcal.html). This global NTL dataset has eliminated the "saturation" problem in the bright cores of urban areas, which happens in the ordinary DMSP/OLS NTL product due to its very limited dynamic range (Hsu et al., 2015).
2. *Vegetation Index data*. The normalized difference vegetation index (NDVI) products based on Satellite Pour I'Observation de la Terre Vegetation were downloaded from the Vlaamse Instelling Voor Technologisch Onderzoek at a spatial resolution of 1 km (http://www.vito-eodata.be/). An annual mean image ($NDVI_{mean}$) for 2010 was calculated based on the original 10-day composite NDVI data.
3. *Land cover data*. Land cover data in China for 2010 at a spatial resolution of 30 m were obtained from the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (Liu et al., 2014). The proportions of four agriculture-related land cover categories within each 1 km × 1 km grid (i.e. farm-rate, forest-rate, grass-rate, and water-rate) were calculated.
4. *Net Primary Production (NPP) data*. The version-55 Terra/MODIS NPP products (MOD17A3) at a spatial resolution of 1 km, produced by the Numerical Terra dynamic Simulation Group/University of Montana were download from National Aeronautics and Space Administration (https://ladsweb.modaps.eosdis.nasa.gov). The MODIS NPP is an annual value and provides information on spatial patterns of productivity in the biosphere (kg C/m^2).
5. *Digital elevation model (DEM) data*. The ASTER GDEM (Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model) Version 2 dataset was downloaded from the website of the Earth Remote Sensing Data Analysis Center of Japan (http://www.gdem.aster.ersdac.or.jp/search.jsp). The original DEM data, with a spatial resolution of 30 m, were re-projected into the Albers Conical Equal Area projection and resampled to a new image at a spatial resolution of 1 km. Elevation and slope layers were generated by the new image in ArcGIS 10.4.1.
6. *Temperature data*. Land surface temperature (LST) data in this study were taken from the MODIS MOD11A2 product (https://ladsweb.modaps.eosdis.nasa.gov/search/), which is

an eight-day composite dataset averaging the daily MOD11A1 LST product. The average daytime and nighttime measurements are stored in separate files with a spatial resolution of 1 km. We choose the LST images observed from June to August and calculated the mean daytime and mean nighttime LSTs (LST-day and LST-night) for the summer of 2010. The MODIS re-projection tool was used for data mosaicking and re-projection into the Albers Conical Equal Area projection at a resolution of 1 km.

Road network data

The road network data across China, including China National Highways, railways, provincial-level, county-level, and township-level roads and city roads, were obtained from the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences. The road network data were used to generate the corresponding road density (road length per unit of area) raster layers (Railway-density, City-Rd-density, and County-Rd-density) at a spatial resolution of 1 km in ArcGIS 10.4.1.

POI data

The POI dataset in this study was derived from the Baidu Map (http://map.baidu.com), which is the largest desktop and mobile map service provider in China (Yao et al., 2017). The application programming interface provided by the Baidu Map Service was used to extract 5,152,850 Baidu POI records for 2010 within 20 categories, *e.g.*, factories, commercial buildings, retail, restaurants, companies, banks, educational facilities, governmental agencies, and residential communities. Each POI record provides not only its location information but also thematic information in the form of a Chinese phrase.

Methodology

RF model fitting and dasymetric GDP mapping

RF is a classic ensemble machine learning approach proposed by Breiman (2001). Differing from traditional linear regression models, RF is a non-parametric method against overfitting that can deal with non-linear relationships, utilize continuous and categorical variables, and model complex interactions (Hastie et al., 2009). Variable importance rankings can be measured by calculating the increased mean squared error (%IncMSE; Liaw and Wiener, 2002). With this quantified measure of variable importance in the RF model, the variables were then further analyzed to reveal their respective effectiveness and the potential reasons. RF model fitting and prediction of GDP density used the "randomForest" package in R 3.5.0. In this study, six RF models were constructed to effectively identify complex associations between various independent variables and GDP from different sectors.

Generally, it is difficult to estimate GDP_1 from the NTL data (Ghosh et al., 2010; Keola et al., 2015). Remote sensing data potentially related to agricultural productivity, such as agriculture-related land cover, NDVI, NPP, LST, slope, elevation, and distance to the coastline, were used as independent variables in RF models for GDP_1 mapping from each agricultural subsector. The proportions of agriculture-related land cover types within each 1 km × 1 km grid were used to restrict the redistribution of GDP from farming, forestry, and fishery to corresponding land cover grid cells. For livestock GDP estimation, built-up land and water bodies were masked out. In doing so, built-up areas were given a GDP_1 value of 0.

Various types of POIs that provide physical infrastructure for different economic activities can be considered as good indicators of non-agricultural GDP distribution. For instance, factory POIs are geographical points that mark the place of various industrial facilities such as iron and steel plants, power plants, cement and lime plants, papermaking plants, and chemical plants. A region with more factory POIs or close to factory POIs usually produce a larger GDP_2. Similarly, service-industry-related POIs, such as banks, retail, restaurants, accommodation services and companies, represent the occurrence of economic activities which mainly contributed to GDP_3. In addition, thermal anomalies can be used to detect industrial heat emission (Liu et al., 2018; Zhang and Zhu, 2019) and thus LST data were used in estimating non-agricultural GDP. Road network density is also significantly related to the growth of GDP per capita (Fan and Chan-Kang, 2008). Therefore, factory POIs, railways and county-level road network density, NTL, and LST were used in the RF model for GDP_2 mapping. Service-related POIs, NTL, county and city road network density, and LST were employed in the RF model for GDP_3 mapping.

The raster layers of predictors at 1 km resolution were aggregated by county and then linked with the logarithm of the census GDP density from different sectors to fit the specific RF models. Once all the RF models were fitted, the corresponding raster layers of the input variables were utilized to predict GDP-distribution weight layers for different sectors at a 1 km resolution. Census GDP data for the individual sectors were then disaggregated using a dasymetric mapping approach based on these weighting layers (Stevens et al., 2015; Ye et al., 2019). Figure 1 illustrates the workflow of the data processing, RF model fitting, and dasymetric GDP mapping. The variables used to fit the RF models are summarized in Supplementary Table 1.

POI data processing

Relationship between non-agricultural GDP and POIs. According to the thematic characteristics of POI data, we first calculated the Spearman's correlation coefficient between the non-agricultural GDP in China and the total number of corresponding POIs at the county level. Table 1 shows that the total number of the factory POIs correlates with the census GDP_2 with an R of 0.64. Service-related POIs are also highly correlated with GDP_3, with R values ranging from 0.58 to 0.93. The categories of commercial buildings ($R = 0.93$) and restaurant and entertainment POIs ($R = 0.91$) are the most strongly correlated with GDP_3. This analysis demonstrates that POI data have the potential to refine non-agricultural GDP mapping.

Secondary sector-related POI data. Factory POIs were used to create two raster layers, i.e., factory POI density (POIs-density-Sec) and the distance to the nearest factory POI (DtN-POIs-Sec), as predictor variables in the RF model for GDP_2 mapping. To generate the POIs-density-Sec layer, the kernel density estimation (KDE) tool in ArcGIS 10.4.1 was employed to convert discrete factory POIs to a smoothly continuous surface. According to the method by Ye et al. (2019), we tested different KDE bandwidths from 500 m to 6000 m with an interval of 100 m and finally determined an optimal bandwidth with 4700 m. The raster layer of DtN-POIs-Sec was generated with a 1 km × 1 km cell size and each grid cell was assigned the Euclidean distance from the center of the grid to the nearest factory POI in ArcGIS 10.4.1.

Tertiary sector-related POI data. The service-industry-related POIs were incorporated into the RF model to estimate GDP_3 density. Using the same method mentioned in the previous

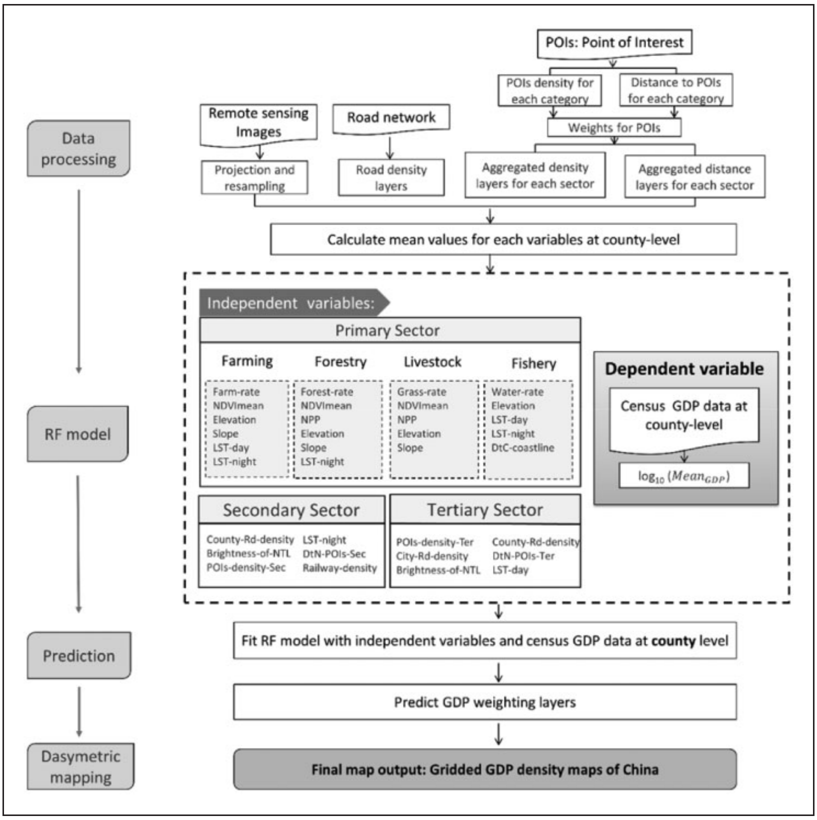

Figure 1. The flow diagram for producing the GDP maps of China.

Table 1. Correlation between selected categories of POIs and GDP_2 and GDP_3 at county level ($n = 2495$).

POI category	GDP_2	GDP_3
Factory	0.64	
Gasoline station		0.58
Bank		0.89
Commercial building		0.93
Retail		0.83
Accommodation services		0.85
Restaurant and entertainment		0.91
Company		0.71
Others (post office, etc.)		0.88

GDP: gross domestic product; POI: point-of-interest.

subsection, eight selected service-related POI categories were used to produce raster layers of POI density and distance to the nearest service-sector POI, respectively. Following Ye et al. (2019), the eight service-sector POI-density layers were combined into one layer (POIs-density-Ter) by utilizing %IncMSEs as weights. Similarly, the eight layers of distance to the

Table 2. Cross-validated accuracies of the Random Forest model for the logarithm of agricultural ($n = 1200$) and non-agricultural GDP estimation ($n = 2495$).

	R^2	RMSE	MAE
Farming	0.87	0.30	0.20
Forestry	0.70	0.47	0.34
Livestock	0.81	0.33	0.24
Fishery	0.82	0.54	0.39
GDP_2	0.90	0.31	0.24
GDP_3	0.95	0.20	0.15

MAE: mean absolute error; RMSE: root mean squared error.

nearest service-sector POI were combined into one raster layer (DtN-POIs-Ter). The % IncMSEs and corresponding weights of different POI categories used to generate the final POIs-density-Ter and DtN-POIs-Ter layers for tertiary industry are shown in Supplementary Table 2.

Results

GDP density maps

Combining multi-source remote sensing data (*e.g.*, NPP, temperature, land use data, vegetation index, and DEM data), four primary sector GDP density maps of China for farming, forestry, livestock and fishery with a spatial resolution of 1 km were produced for the year 2010 (Figure 2). Areas of high farming GDP density are located in the North China Plain, the Yangtze River Delta, and the Guanzhong Plain, which are low relief areas with vast tracts of arable land. The Northeast China Plain and Sichuan basin also have high economic production from farming (Figure 2(a)). For forestry, relatively higher economic production occurs in the mountainous area of eastern China (Figure 2(b)). Most of the livestock-sector GDP contributions are from Sichuan, Xinjiang, Hubei, and Hunan Provinces, the Northeast China Plain, and the North China Plain (Figure 2(c)). Coastal areas and some inland areas with numerous lakes or rivers in east China support a large fishery GDP density (Figure 2 (d)). A composite GDP_1 density map of mainland China, based on the above four maps, is shown in Figure 3(a). Strong and heterogeneous GDP_1 density was mainly distributed in the plains of eastern China and the Sichuan basin. Some areas in Liaoning, Hubei, Hunan, and Guangdong Provinces also have high GDP_1 density.

The density of non-agricultural GDP across China in 2010 was spatialized separately with the joint use of NTL satellite images and POI data (Figure 3(b) and (c)). Generally, the urban areas exhibit high non-agricultural GDP densities. According to the spatial distribution of the final disaggregated total GDP density across mainland China for the year 2010 (Figure 3(d)), most of the GDP is distributed in plains and basin regions with a high urbanization level. The highest GDP densities are located in the metropolitan area of Beijing–Tianjin–Hebei, the Yangtze River Delta, the Pearl River Delta, and some provincial capitals.

Figure 4 shows the predicted GDP_1, GDP_2, and GDP_3 density maps in four metropolitan areas of Beijing–Tianjin, the Yangtze River Delta, the Pearl River Delta, and Chengdu–Chongqing. There are no GDP_1 in urban areas in the four regions. Areas with higher GDP_2 density are mainly scattered at the urban fringe or some cities with relatively low

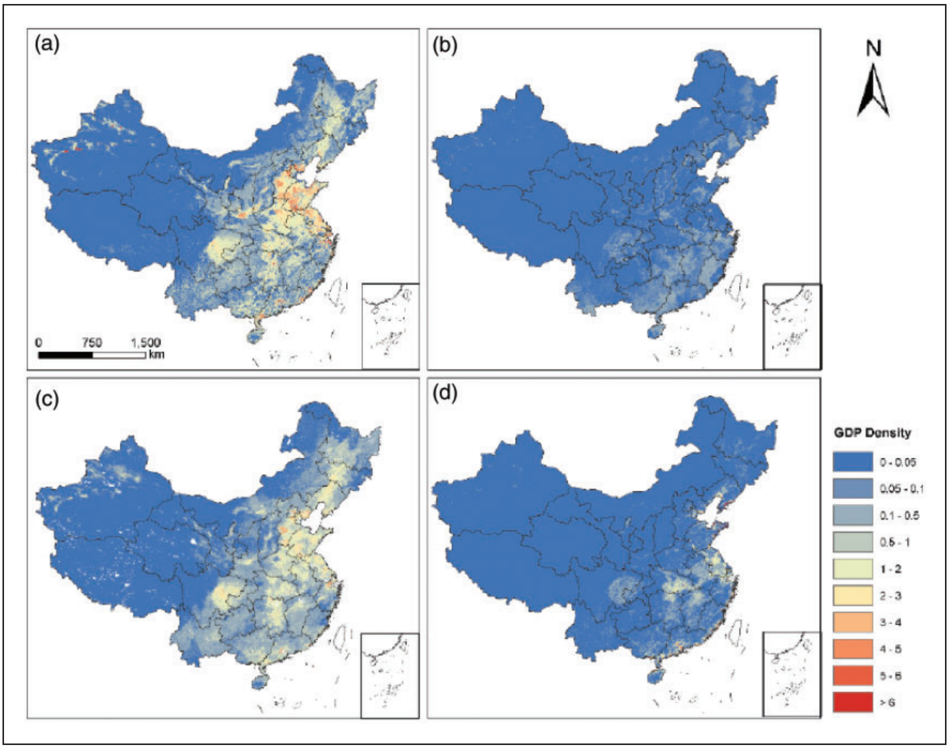

Figure 2. GDP density maps for the primary sectors of (a) farming, (b) forestry, (c) livestock, and (d) fishery in 2010 (unit: million CNY/km^2).

urbanization level, probably due to low land prices and less stringent environmental standards in this region (Hu et al., 2016). By contrast, very high GDP$_3$ density occurs in the highly urbanized metropolizes, particularly in downtown areas, presenting the character of agglomeration (Figure 4). Our methodology captured the main spatial characteristics of the different economic sectors and generated a more reasonable spatial distribution of GDP density.

Accuracy assessment

While an ideal measure to validate the gridded GDP dataset would involve a cell-by-cell comparison, the collection of such data is difficult and costly. Another common method uses a summed gridded GDP value compared to a finer-level original census GDP. Unfortunately, finer-level (such as township) census GDP data are also not available in China. Here, we use 10-fold cross-validation to assess the predictive accuracy of each RF model by calculating cross-validated R^2, root mean squared error (RMSE), and mean absolute error (MAE; Table 2). Overall, the RF models developed for the non-agricultural sector (with R^2 from 0.90 to 0.95) performed better than those for the agricultural sectors (with R^2 ranging from 0.70 to 0.87). Agreement between the logarithm of census GDP and cross-validated predictions was good, with RMSE varying from 0.20 to 0.54 and MAE from 0.15 to 0.39, which means that the proposed GDP modeling method has good predictive ability.

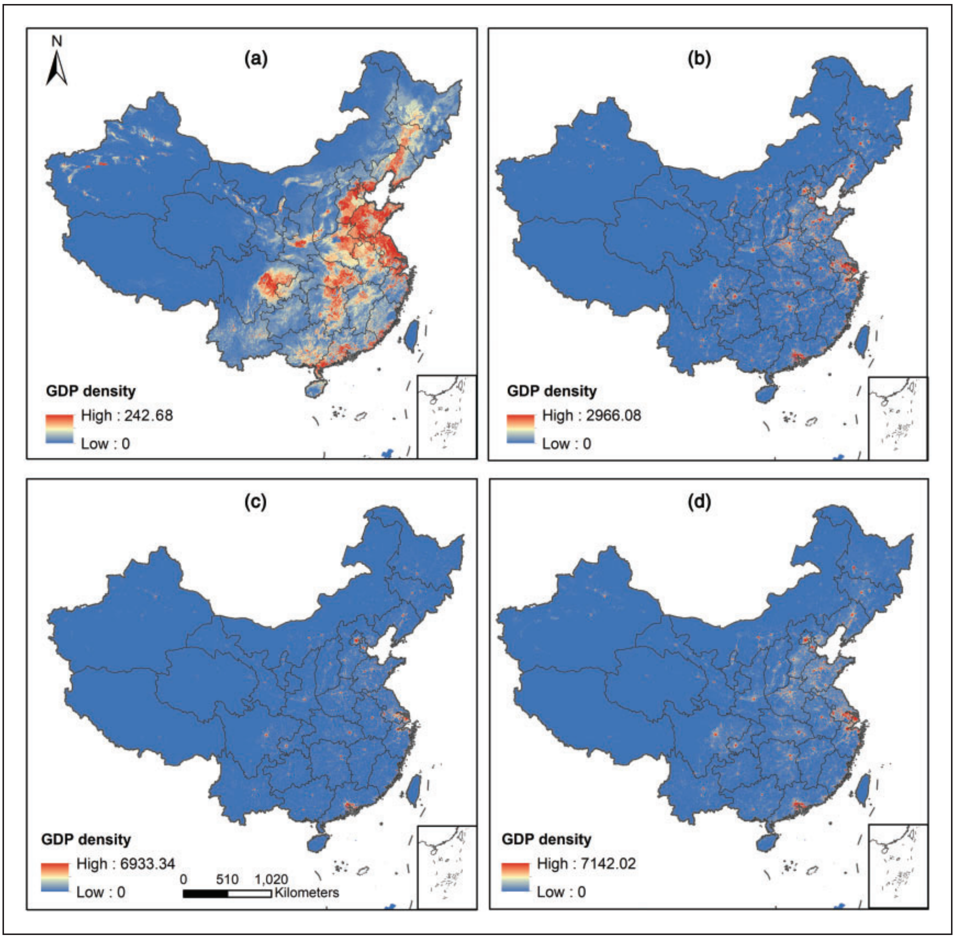

Figure 3. GDP density maps for (a) the primary sectors, (b) the secondary sectors, (c) the tertiary sectors, and (d) the summed GDP for all sectors in 2010 (unit: million CNY/km^2).

Random Forest variable importance

Variable importance for the agricultural GDP estimation. According to the statistical outputs of RF models for the primary industry, the variance explained in predicting GDP$_1$ of farming, forestry, livestock and fishery at the county level was 84.24%, 67.90%, 80.49%, and 78.52%, respectively. For farming and fishery (Figure 5(a) and (d)), the proportion of corresponding land use in a county contributes largely in prediction of GDP density because land use is the fundamental requirement for these agricultural activities. Due to the paucity of information about forest land use rather than forest land cover, the proportion of forest land had less importance (Figure 5(b)). For example, many forest lands were classified as nature reserves for the requirements of ecological protection (Huang et al., 2019), and forests around cities were usually developed as parks, rather than used for forestry production. This may also explain the relatively poor performance of the RF model for forestry GDP estimation compared to other agricultural sectors. NDVI and NPP, which reflect the growth conditions of vegetation and are widely used to monitor agricultural and forest production (Baisden, 2006; Huang et al., 2014; Mkhabela et al., 2011; Son et al., 2014;

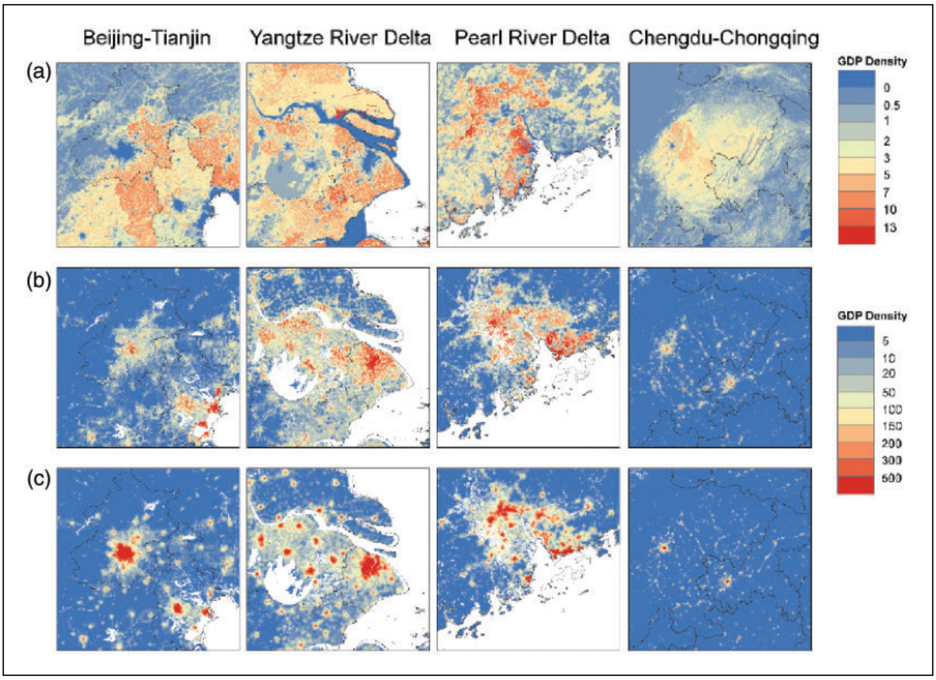

Figure 4. Comparison of the GDP density maps from (a) the primary sector, (b) the secondary sector, and (c) the tertiary sector for four metropolitan agglomerations in China (unit: million CNY/km^2).

Tao et al., 2005), are top-ranked variables in the farming and forestry GDP modeling (Figure 5(a) and (b)). These two vegetation-related variables also linked to the potential output of livestock systems (Figure 5(c); Yu et al., 2010). The two topography-related variables, elevation, and slope, were also important for the prediction of GDP density in the four agricultural subsectors (Figure 5). Higher GDP$_1$ was generally concentrated in low-elevation, flat areas across China. Temperature is also crucial for agricultural production with more importance of nighttime LST than daytime LST. Fishery activities include both marine fishery and freshwater fishery. Distance to the ocean was an important predictor in the RF model for fishery GDP estimation (Figure 5(d)).

Variable importance for the non-agricultural GDP estimation. For non-agricultural GDP estimation, six predictors explained 89.91% and 95.26% of the variance of GDP$_2$ and GDP$_3$ in China, respectively. At night, ceaseless industrial processes release large amounts of light and waste heat and make industrial area brighter and warmer compared to other areas (Liu et al., 2018; Zhang and Zhu, 2019; Zhao et al., 2013). Therefore, the brightness of NTL and nighttime LST are the most important variables in predicting GDP$_2$ (Figure 6(a)). County road infrastructure also has a high variable importance. The coordinates of factory POIs show the precise geographic location of industrial enterprises (Huang et al., 2018), which can refine the GDP$_2$ mapping. The DtN-POIs-Sec and railway density are identified as the least important predictor variables in GDP$_2$ estimation.

Most service industries are located in built-up areas. Daytime LST, which was positively correlated with built-up areas (Xu, 2008), is the most important predictor in GDP$_3$ estimation. The two POI-based variables had large contributions to GDP$_3$ estimation and the %

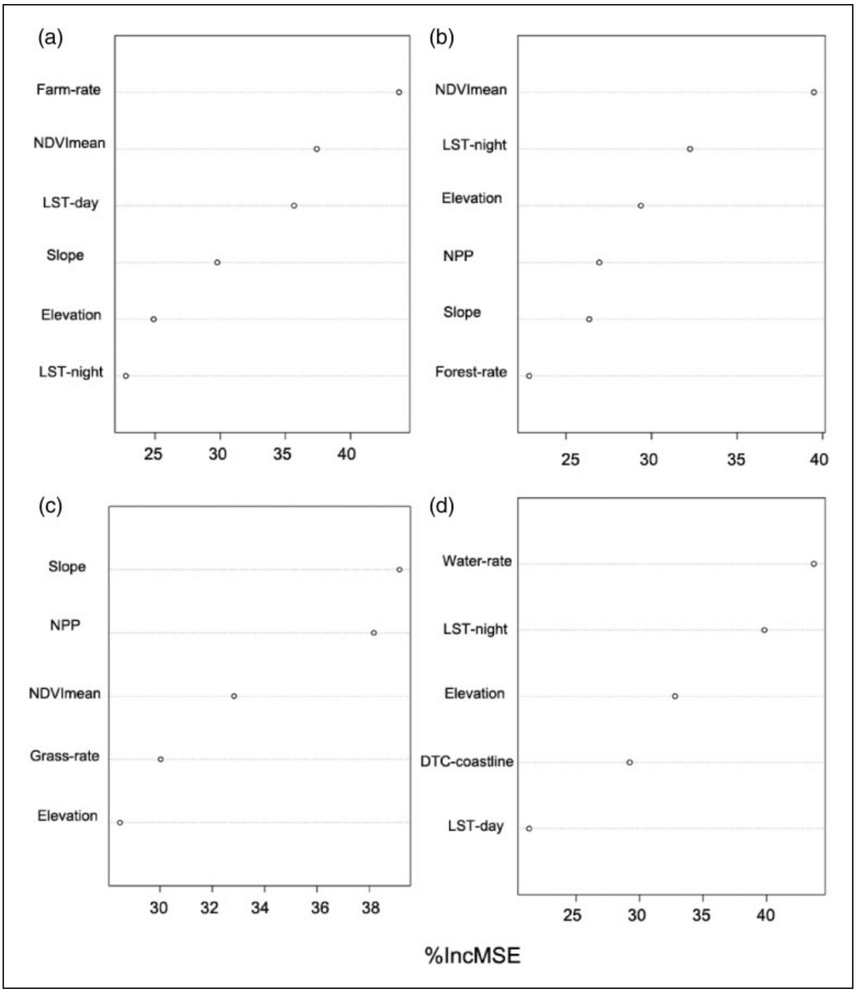

Figure 5. Percent increased mean square error (%IncMSE) that indicates the variable importance in RF regression for the (a) farming, (b) forestry, (c) livestock, and (d) fishery GDP estimation of the primary sector.

IncMSE value of POI-density-Ter was much larger than that of the brightness of NTL (Figure 6(b)). High density service-related POIs can represent an area with high GDP_3 density and exclude industrial regions.

Discussion

NTL data offer a unique view of the Earth's surface and can be used to estimate and map economic activities across a range of spatial scales. The main limitations of previous studies are their overdependence on NTL data (Keola et al., 2015). Economic activities from agriculture are generally distributed in the darker areas at night. Therefore, the overall GDP contributed by the primary sector cannot be accurately reflected by only NTL data, which represents a higher value-added economic activity (Wang et al., 2019b). This is particularly

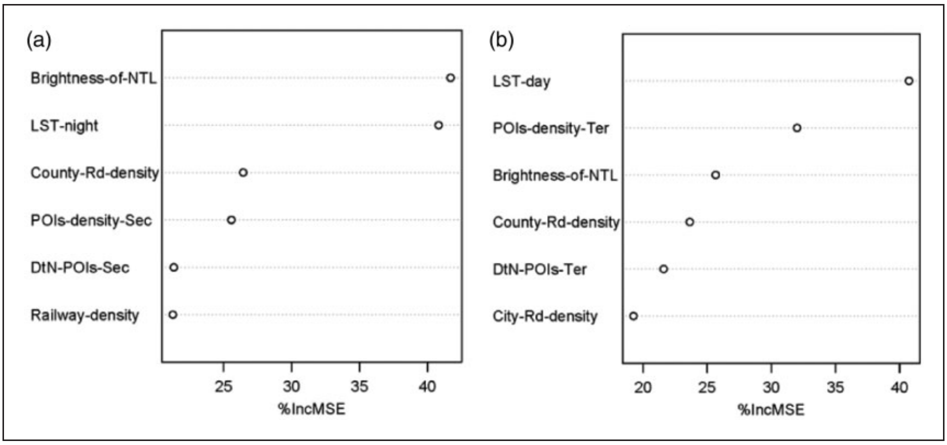

Figure 6. Same as in Figure 5 but for (a) GDP_2 and (b) GDP_3.

troublesome in developing countries that often have a large share of total GDP contributed from agriculture or forestry (Keola et al., 2015; Wang et al., 2019b). Recently, land use/cover data have been used to estimate primary sector production (Cao et al., 2016; Chen et al., 2016; Han et al., 2012; Keola et al., 2015; Yue et al., 2014). In this study, we further integrated NDVI, NPP, elevation, slope, and LST which potentially influence agricultural production in the RF models. Statistical outputs of the RF models indicated that the abovementioned variables can play an important role in estimating GDP_1.

Another limitation of using NTL images for GDP mapping is that the GDP_2 and GDP_3 cannot be distinguished effectively. Compared with NTL data, POIs have an inherent advantage in revealing the actual distribution of different economic activities, especially in urbanized areas. Based on their taxonomy and names, the POIs can be classified into different types, such as commerce, business offices, catering, recreation, and factories. The density of a certain type of POI shows the spatial distribution pattern of a certain group of related economic activities. Different functional distribution patterns within cities can be identified based on the POI data (Figure S1), with a high probability that the specific distributions of GDP_2 and GDP_3 can be differentiated effectively. However, in urban fringe areas or rural areas, the POI density is significantly lower than in urban areas and thus have limited capability to estimate GDP_1. The NTL data and POIs have specific disadvantages regarding the description of economic activities. Given that GDP distribution should be recognized using both natural/physical and social attributes, disaggregating GDP based on the sole use of either NTL or geospatial big data has unavoidable limitations. Therefore, integrating the NTL data, land use data, POIs, and other remote sensing data can effectively capture the distribution of economic activities from the different sectors and significantly improve the accuracy of GDP mapping. The spatialized GDP maps also provide insights into the pattern of the value-added commodity chain involving different economic sectors in China. Low value-added agricultural production activities are generally founded in rural areas. The value of raw material builds up across the process of manufacturing and assembly (medium value-added activities), usually locating at the urban fringe or in less-urbanized cities while high value-added activities such as design and marketing are often located in highly urbanized areas.

Currently, the most widely used POI is generally extracted from online platforms or based on volunteered geographic information (VGI). The coverage and accuracy of these types of geospatial big data depend heavily on the completeness of the online sources and the veracity of user-contributed content (Jiang et al., 2015). VGI-based POIs were generally recorded in urban areas and areas with more service entities or infrastructure (*e.g.*, the Central Business District or tourist attractions), while there were insufficient POIs in rural areas as well as residential areas (Ma et al., 2015). The POI data used in this study were based on commercial navigation data. Commercial POI data are collected by trained persons and undergo rigorous inspections and corrections. As a result, the positional and thematic accuracy of Baidu POI data is reasonably reliable. The superiority of commercial POI data as good indicators of various economic activities without substantial user biases make them more prominent in spatially disaggregating census GDP from different non-agricultural activities to a very fine geographic scale.

However, an inevitable limitation of this study was that our current approach focuses on the quantity rather than the quality of the POIs. One example is that a large department store and a small convenience store, which generate distinct economic outputs, are treated equally. Possible further incorporation of other forms of big data would provide supplementary information to identify the intensity of economic activities in specific fine-scale areas (Zhao et al., 2018). Moreover, the accuracy of GDP estimation is potentially limited by the uncertainty of the RF models. The main limitation of the RF regression model is that it cannot predict the values beyond the range in the training dataset (Heenkenda et al., 2015; Houborg and McCabe, 2018).

Conclusions

This study proposes a novel solution for GDP mapping at the pixel level through RF modeling using both remote sensing and POI data. Differing from previous studies, we constructed individual RF models for each sector in order to capture the non-linear relationships between the predictors and census GDP from different sectors. Benefiting from the combination of the advantages of NTL and POI, we explored the potential for spatializing GDP_2 and GDP_3 separately. The proposed methodology of GDP modeling shows good predictive ability. On the countrywide scale, our case study of mainland China suggests that the integration of multi-source data helps to generate a reasonable estimation of GDP density, which offers an intuitive way of visualizing GDP distribution. The gridded GDP maps can be used as an important scientific reference for governments and organizations when formulating developmental strategies. Besides revealing the effectiveness of POIs in non-agricultural GDP mapping, this study suggests the potential of POIs to improve the disaggregation of other socioeconomic parameters in the future.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Second Tibetan Plateau Scientific Expedition and Research program (STEP) (No. 2019QZKK0603), the National Natural Science

Foundation of China (No. 41671035), the Fundamental Research Funds for the Central Universities (No. 2019QNA4050).

ORCID iD

Xuchao Yang https://orcid.org/0000-0002-8130-7447

References

Baisden WT (2006) Agricultural and forest productivity for modelling policy scenarios: Evaluating approaches for New Zealand greenhouse gas mitigation. *Journal of the Royal Society of New Zealand* 36(1): 1–15.

Bakillah M, Liang S, Mobasheri A, et al. (2014) Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science* 28(9): 1940–1963.

Bennett MM and Smith LC (2017) Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment* 192: 176–197.

Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.

Cao Z, Wu Z, Kuang Y, et al. (2016) Coupling an intercalibration of radiance-calibrated nighttime light images and land use/cover data for modeling and analyzing the distribution of GDP in Guangdong, China. *Sustainability* 8(2): 108.

Chen Q, Hou X, Zhang X, et al. (2016) Improved GDP spatialization approach by combining land-use data and night-time light data: A case study in China's continental coastal area. *International Journal of Remote Sensing* 37(19): 4610–4622.

Chen X and Nordhaus W (2011) Using luminosity data as proxy for economic statistics. *Proceedings of the National Academy of Sciences of the United States of America* 108(21): 8589–8594.

Dai Z, Hu Y and Zhao G (2017) The suitability of different nighttime light data for GDP estimation at different spatial scales and regional levels. *Sustainability* 9(2): 305.

Dasgupta S, Laplante B, Murray S, et al. (2011) Exposure of developing countries to sea-level rise and storm surges. *Climatic Change* 106(4): 567–579.

Doll CNH, Muller JP and Elvidge CD (2000) Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *AMBIO: A Journal of the Human Environment* 29(3): 157–162.

Doll CNH, Muller JP and Morley JG (2006) Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics* 57(1): 75–92.

Elvidge CD, Baugh KE, Kihn EA, et al. (1997) Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing* 18(6): 1373–1379.

Fan S and Chan-Kang C (2008) Regional road development, rural and urban poverty: Evidence from China. *Transport Policy* 15(5): 305–314.

Gao S, Janowicz K and Couclelis H (2017) Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21(3): 446–467.

Geiger T (2018) Continuous national gross domestic product (GDP) time series for 195 countries: Past observations (1850–2005) harmonized with future projections according to the shared socioeconomic pathways (2006–2100). *Earth System Science Data* 10(2): 847–856.

Geiger T, Frieler K and Bresch DN (2018) A global historical data set of tropical cyclone exposure (TCE-DAT). *Earth System Science Data* 10(1): 185–194.

Ghosh T, Powell R, Elvidge CD, et al. (2010) Shedding light on the global distribution of economic activity. *The Open Geography Journal* 3: 148–161.

Han X, Zhou Y, Wang S, et al. (2012) GDP spatialization in China based on DMSP/OLS data and land use data. *Remote Sensing Technology and Application* 27(3): 396–405.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer Science+Business Media, LLC.

Heenkenda MK, Joyce KE, Maier SW, et al. (2015) Quantifying mangrove chlorophyll from high spatial resolution imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 108: 234–244.

Henderson JV, Storeygard A and Weil DN (2012) Measuring economic growth from outer space. *The American Economic Review* 102(2): 994–1028.

Houborg R and McCabe MF (2018) A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 135: 173–188.

Hsu F-C, Baugh K, Ghosh T, et al. (2015) DMSP-OLS radiance calibrated nighttime lights time series with intercalibration. *Remote Sensing* 7(2): 1855–1876.

Hu T, Yang J, Li X, et al. (2016) Mapping urban land use by using landsat images and open social data. *Remote Sensing* 8(2): 151.

Huang J, Wang H, Dai Q, et al. (2014) Analysis of NDVI data for crop identification and yield estimation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7(11): 4374–4384.

Huang L, Wu Y, Zheng Q, et al. (2018) Quantifying the spatiotemporal dynamics of industrial land uses through mining free access social datasets in the mega Hangzhou Bay region, China. *Sustainability* 10(10): 3463.

Huang Y, Fu J, Wang W, et al. (2019) Development of China's nature reserves over the past 60 years: An overview. *Land Use Policy* 80: 224–232.

Jia P, Sankoh O and Tatem AJ (2015) Mapping the environmental and socioeconomic coverage of the INDEPTH international health and demographic surveillance system network. *Health & Place* 36: 88–96.

Jiang S, Alves A, Rodrigues F, et al. (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems* 53: 36–46.

Keola S, Andersson M and Hall O (2015) Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. *World Development* 66: 322–334.

Li X, Xu H, Chen X, et al. (2013) Potential of NPP-VIIRS nighttime light imagery for modeling the regional economy of China. *Remote Sensing* 5(6): 3057–3081.

Liaw A and Wiener M (2002) Classification and regression by. *RandomForest*. *R News* 3: 18–22.

Liu J, Kuang W, Zhang Z, et al. (2014) Spatiotemporal characteristics, patterns, and causes of land-use changes in China since the late 1980s. *Journal of Geographical Sciences* 24(2): 195–210.

Liu X, He J, Yao Y, et al. (2017) Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science* 31(8): 1675–1696.

Liu Y, Delahunty T, Zhao N, et al. (2016) These lit areas are undeveloped: Delimiting China's urban extents from thresholded nighttime light imagery. *International Journal of Applied Earth Observation and Geoinformation* 50: 39–50.

Liu Y, Hu C, Zhan W, et al. (2018) Identifying industrial heat sources using time-series of the VIIRS nightfire product with an object-oriented approach. *Remote Sensing of Environment* 204: 347–365.

Liu Y, Liu X, Gao S, et al. (2015) Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers* 105(3): 512–530.

Ma D, Sandberg M and Jiang B (2015) Characterizing the heterogeneity of the OpenStreetMap data and community. *ISPRS International Journal of Geo-Information* 4(2): 535.

Ma T, Zhou C, Pei T, et al. (2012) Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities. *Remote Sensing of Environment* 124: 99–107.

McKenzie G, Janowicz K, Gao S, et al. (2015) POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization* 50(2): 71–85.

Mkhabela MS, Bullock P, Raj S, et al. (2011) Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology* 151(3): 385–393.

Paprotny D, Morales-Nápoles O and Jonkman SN (2018) HANZE: A pan-European database of exposure to natural hazards and damaging historical floods since 1870. *Earth System Science Data* 10(1): 565–581.

Shi K, Yu B, Huang Y, et al. (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): 1705–1724.

Small C, Pozzi F and Elvidge C (2005) Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sensing of Environment* 96(3–4): 277–291.

Son NT, Chen CF, Chen CR, et al. (2014) A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. *Agricultural and Forest Meteorology* 197: 52–64.

Stevens FR, Gaughan AE, Linard C, et al. (2015) Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS One 10(2): e0107042.

Sutton PC and Costanza R (2002) Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics* 41: 509–527.

Tao F, Yokozawa M, Zhang Z, et al. (2005) Remote sensing of crop production in China by production efficiency models: Models comparisons, estimates and uncertainties. *Ecological Modelling* 183(4): 385–396.

Wang W, Cheng H and Zhang L (2012) Poverty assessment using DMSP/OLS night-time light satellite imagery at a provincial scale in China. *Advances in Space Research* 49(8): 1253–1264.

Wang X, Rafa M, Moyer JD, et al. (2019b) Estimation and mapping of sub-national GDP in Uganda using NPP-VIIRS imagery. *Remote Sensing* 11(2): 163.

Wang X, Sutton PC and Qi B (2019a) Global mapping of GDP at 1 km^2 using VIIRS nighttime satellite imagery. *ISPRS International Journal of Geo-Information* 8(12): 580.

Wu J, Wang Z, Li W, et al. (2013) Exploring factors affecting the relationship between light consumption and GDP based on DMSP/OLS nighttime satellite imagery. *Remote Sensing of Environment* 134(0): 111–119.

Xu H (2008) A new index for delineating built-up land features in satellite imagery. *International Journal of Remote Sensing* 29(14): 4269–4276.

Yang T, Li M and Shen Z (2015) Between morphology and function: How syntactic centers of the Beijing city are defined. *Journal of Urban Management* 4(2): 125–134.

Yang X, Ye T, Zhao N, et al. (2019) Population mapping with multisensor remote sensing images and point-of-interest data. *Remote Sensing* 11(5): 574.

Yao Y, Li X, Liu X, et al. (2017) Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science* 31(4): 825–848.

Ye T, Zhao N, Yang X, et al. (2019) Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *The Science of the Total Environment* 658: 936–946.

Yoshida D, Song X and Raghavan V (2010) Development of track log and point of interest management system using free and open source software. *Applied Geomatics* 2(3): 123–135.

Yu L, Zhou L, Liu W, et al. (2010) Using remote sensing and GIS technologies to estimate grass yield and livestock carrying capacity of Alpine grasslands in Golog prefecture. *Pedosphere* 20(3): 342–351.

Yue W, Gao J and Yang X (2014) Estimation of gross domestic product using multi-sensor remote sensing data: A case study in Zhejiang province. *Remote Sensing* 6(8): 7260.

Zandbergen PA and Ignizio DA (2010) Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science* 37(3): 199–214.

Zhang G and Zhu AX (2019) A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena. *International Journal of Geographical Information Science* 33(9): 1873–1893.

Zhao M, Cheng W, Zhou C, et al. (2017a) GDP spatialization and economic differences in South China based on NPP-VIIRS nighttime light imagery. *Remote Sensing* 9(7): 673.

Zhao N, Cao G, Zhang W, et al. (2018) Tweets or nighttime lights: Comparison for preeminence in estimating socioeconomic factors. *ISPRS Journal of Photogrammetry and Remote Sensing* 146: 1–10.

Zhao N, Currit N and Samson E (2011) Net primary production and gross domestic product in China derived from satellite imagery. *Ecological Economics* 70(5): 921–928.

Zhao N, Liu Y, Cao G, et al. (2017b) Forecasting China's GDP at the pixel level using nighttime lights time series and population images. *GIScience & Remote Sensing* 54(3): 407–425.

Zhao X, Jiang H, Wang H, et al. (2013) Remotely sensed thermal pollution and its relationship with energy consumption and industry in a rapidly urbanizing Chinese city. *Energy Policy* 57: 398–406.

Zhu X, Ma M, Yang H, et al. (2017) Modeling the spatiotemporal dynamics of gross domestic product in China using extended temporal coverage nighttime light data. *Remote Sensing* 9(6): 626.

Qian Chen is a PhD candidate at Ocean College, Zhejiang University. Her research focuses on the relationship between climate change and coastal disaster risk management and the comprehensive risk assessment in coastal areas. In particular, she aims to develop an open-data-based procedure integrating remote sensing and geospatial big data for improving spatial representations of risk components.

Tingting Ye is a Student at Ocean College, Zhejiang University as project assistant until January 2020. She is specialized in spatial analysis.

Naizhuo Zhao received the PhD degree in environmental geography from Texas State University, San Marcos, the United States, in 2014. He worked as a lecturer and postdoctoral research associate at Texas Tech University, Lubbock, the United States, during 2014 to 2018. At present, he is a professor at Northeastern University, Shenyang, China. His research interests include nighttime lights imagery, location-based social media data, machine learning, and air pollution.

Mingjun Ding is a Professor in Land use policy and leads the research group of Land use change Group at the Key Lab of Poyang Lake Wetland and Watershed Research, Ministry of Education, Jiangxi Normal University. His research mainly focuses on land use/cover change and its ecological and environmental effects basing on the observations, GIS and RS approaches. In addition, as an expert consultant, he often provides some advisory reports about land use for the local government.

Zutao Ouyang is a postdoctoral researcher in Earth System Science Department, Stanford University. He hold a PhD degree in geography from Michigan State University. Dr Ouyang is an ecosystem ecologist and geographer using data-driven approaches to study global environmental changes, including climate change, land use land cover change, and coupled natural and human (CNH) systems.

Peng Jia is a faculty member at the University of Twente. He coined the term "Spatial Lifecourse Epidemiology" and founded the International Initiative on Spatial Lifecourse Epidemiology (ISLE) to facilitate this area. He received BE degree in environmental engineering, two MS degrees in spatial science and spatial epidemiology, and PhD degree in health geography. He uses statistical, spatial, location-based, and artificial intelligence

technologies to conduct spatial lifecourse epidemiologic research. He is also expert in planning health-care resource allocation, and optimizing hierarchical health-care systems.

Wenze Yue is a Professor in the Department of Land Management, the director of the Institute of Land Science and Property Management, Zhejiang University. His research interests are urbanization, urban growth analysis, and spatial planning.

Xuchao Yang is an Associate Professor in Ocean College, Zhejiang University. His research interests include natural disaster risk management in coastal zone and urban environmental health due to the combined effect of urbanization and climate change. He received the BS degree in physical geography from the Lanzhou University in 2002, the MS degree from the Institute of Earth Environment, Chinese Academy of Sciences, in 2005, and the PhD degree in Physical Geography from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, in 2008.