https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

A long-term consistent socioeconomic dataset of Chinese cities generated by Bayesian spatiotemporal modeling with multi-source Earth observations

Zhangying Tang[1†], Xianteng Tang[1,2†], Lingfeng Liao[2], Guoqiang Yan[1], Zhenyan Wang[1], Yuju Wu[2,3,4], Mingyu Xie[5], Yumeng Zhang[2,3,4], Chengwu Wang[1], Zhoufeng Wang[1], Yangting Zeng[6], Chao Song[2,3,4*], Jay Pan[2,3,4]

[1]State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, School of Geoscience and Technology, Southwest Petroleum University, Chengdu, 610500, China
[2]HEOA–West China Health & Medical Geography Group, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, 610041, China
[3]Health Promotion and Food Nutrition & Safety Key Laboratory of Sichuan Province, Chengdu, 610041, China
[4]Institute for Healthy Cities and West China Research Center for Rural Health Development, Sichuan University, Chengdu, 610041, China
[5]Chengdu Center for Disease Prevention and Control, China Railway Chengdu Group Co., Ltd., 4 Xiyi Lane, Chengdu North Railway Station, Jinniu District, Chengdu, 610081, China
[6]Institute for Disaster Management and Reconstruction (IDMR), Sichuan University, Chengdu, 610207, China

† These authors contributed equally to this work.

Correspondence to: Chao Song (chaosong@scu.edu.cn)

Abstract. Within the Healthy Cities and Sustainable Development Goals (SDGs) agendas, socioeconomic data are fundamental for tracking regional development. China, however, lacks a complete, long-term subnational socioeconomic dataset due to severe spatiotemporal missingness in official statistical yearbooks. We compiled 35 official socioeconomic indicators for 366 Chinese cities from 2000 to 2021, incorporated remote-sensing-derived covariates as auxiliary information, and applied a Bayesian spatiotemporal interacting varying intercepts (BSTIVI) model to capture the target variables' spatial, temporal, and coupled spatiotemporal dependence. Model performance was evaluated using global Bayesian criteria and cross-validation, while local error distributions and temporal trends were visualized to examine imputation outcomes. Based on the completed dataset, we further derived a composite development index using entropy weighting and assessed spatial inequality with the Gini coefficient, coefficient of variation and hotspot analysis. The results show that BSTIVI achieved markedly better fit than traditional multiple linear regression (MLR). In cross-validation, 32 of 35 indicators achieved R^2 >= 0.95, RMSE and MAE remained low. The resulting data product showed strong imputation performance in both spatial and temporal dimensions. Analyses of the completed dataset revealed marked spatial inequality and clustering in urban socioeconomic development across China during 2000-2021. We ultimately produced the first long-term city-level socioeconomic dataset for China, comprising 35 indicators and one composite index, with Bayesian credible intervals for imputed values. This study provides both a new city-level data resource for China and a transferable framework for imputing missing subnational socioeconomic data worldwide, thereby supporting Earth system research and SDG implementation.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

1 Introduction

Socioeconomic data comprise diverse indicators that characterize regional economic and social development and are essential for policy formulation, economic monitoring, and scientific research (Zhang et al., 2024). Within the Earth system science framework, socioeconomic processes are increasingly recognized not as exogenous drivers, but as integral components of coupled human-natural systems. Population dynamics, economic activity, and urban development both shape and respond to environmental change, thereby influencing the trajectories of system evolution. High-quality socioeconomic datasets therefore provide essential information on human-system exposure, supporting the analysis of human activities and their feedback mechanisms, and providing an important empirical foundation for investigating human-environment interactions. At the same time, such data are indispensable for tracking progress toward the UN Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being), SDG 4 (Quality Education), and SDG 11 (Sustainable Cities and Communities), for which they provide essential empirical support. However, national averages often conceal substantial subnational disparities (Zhao et al., 2017; India State-Level Disease Burden Initiative Child Mortality, 2020). By contrast, fine-scale socioeconomic data more effectively characterize spatially explicit patterns of inequality (Lozano et al., 2018; Neal et al., 2019), and support refined assessments of human-environment interactions at local scales. Consequently, such fine-grained data are essential not only for improving the accuracy of SDG monitoring (Mckeen et al., 2023; Oh et al., 2024), but also for advancing the understanding of complex interactions within human-Earth systems. Nevertheless, on the global scale, long-term subnational socioeconomic datasets suffer from severe missing-value problems.

One common approach to imputing missing socioeconomic data is to increase sample size or incorporate auxiliary information. In practice, however, socioeconomic data collection often requires substantial human, material, and financial resources, and official statistics in most countries rely primarily on sample surveys to produce representative estimates more efficiently and cost-effectively (Amaral et al., 2015). Consequently, when missingness occurs, additional samples or suitable auxiliary variables are often unavailable, especially in resource-limited settings. Recent advances in satellite remote sensing, however, have created new opportunities to derive auxiliary variables relevant to socioeconomic conditions. This technology not only expands pathways for acquiring auxiliary information but has also been widely applied in population downscaling (Alegana et al., 2015; Sorichetta et al., 2015) and the estimation of health indicators across countries (Tatem, 2014; Wang et al., 2016; James et al., 2018). As auxiliary variables in these studies, remote sensing data have demonstrated substantial value and potential in both academic research and practical applications (Utazi et al., 2018; Lloyd et al., 2019).

The second strategy is to use model-based imputation when additional samples or auxiliary information are unavailable (Ferreira et al., 2020). Common methods include statistical methods such as expectation maximization (EM) (De Souto et al., 2015) and linear regression (LR) (Pati and Das, 2017), as well as machine-learning approaches including k-nearest neighbors (KNN) (De Silva and Perera, 2016), decision trees (DT) (Purwar and Singh, 2015), and random forests (RF) (Xia et al., 2017). However, these approaches often ignore or only partially capture spatial dependence in the target variables (Seu, 2022), which can reduce fit and predictive accuracy while increasing imputation error and uncertainty (Zahmatkesh and Zech,

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

2026). Under the first law of geography, nearby areas tend to share similar structural characteristics (Goodchild, 2009), and regional time series usually show regular temporal dynamics (Song et al., 2024), both of which are critical for long-term geospatial data (Song et al., 2020; Song et al., 2022). Accordingly, spatiotemporal models, especially those explicitly coupling spatial and temporal processes, are well suited to missing-value imputation because they can recover missing values from the variables' own structural patterns even without extra samples or auxiliary information (Song et al., 2018). Yet for official socioeconomic datasets, few studies have combined satellite-derived auxiliary variables with spatiotemporal models to further improve imputation accuracy.

In China, subnational socioeconomic data also exhibit substantial spatiotemporal missingness. Although official sources, such as the China City Statistical Yearbook and the China County Statistical Yearbook, cover broad geographic areas and are publicly available, long-term statistical data remain missing for some regions and reporting of certain indicators has been discontinued. Academic datasets are likewise limited. Existing products provide fine spatial resolution at the city, county, or grid level, but most focus on single indicators, such as electricity consumption (Zhou et al., 2025), GDP (Zhao et al., 2017), human development index (Gong et al., 2025), or hospital accessibility (Ye et al., 2024). For multi-indicator socioeconomic data, at present only the county-level dataset by Song et al. exists. It contains only about 20 absolute indicators, and covers just 2002-2011, making it insufficient for large-scale, long-term analyses. Moreover, this dataset has not been publicly released, further restricting its use. Consequently, China still lacks a long-term, multi-indicator subnational socioeconomic dataset. This gap is especially consequential for the Healthy Cities initiative and for monitoring progress toward SDG 11. As urbanization accelerates, cities have become central units of socioeconomic development and key spatial units for SDG implementation, making city-level socioeconomic data increasingly important.

In summary, this study focuses on the city scale and integrates remote-sensing-derived auxiliary variables with spatiotemporal modeling to improve missing-value prediction and generate a long-term, broad-coverage socioeconomic dataset for China. Specifically, we use satellite-derived proxies (e.g., nighttime lights, $PM_{2.5}$, and temperature) as auxiliary variables, apply the Bayesian spatiotemporal interacting varying intercepts (BSTIVI) model (Song et al., 2022), and incorporate spatiotemporal interaction structures in the target variables to enhance predictive performance. This framework enables the production and validation of the first city-level socioeconomic dataset for China. Based on the completed indicators, we further construct a composite index of socioeconomic development and evaluate spatiotemporal inequality. Overall, the resulting annual city-level dataset for 2000-2021 provides a valuable basis for long-term regional research in China and offers a transferable solution for imputing missing values in small-area socioeconomic datasets worldwide, thereby supporting progress toward the SDGs.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

2 Materials and methods

2.1 Data collection and pre-processing

2.1.1 Socioeconomic data

To construct a multidimensional socioeconomic dataset, we collected and organized 35 official socioeconomic indicators for 366 Chinese cities from 2000 to 2021. The primary source was the annual China City Statistical Yearbook, supplemented by municipal statistical bulletins and other official local releases to fill missing records, such as Chengdu's local statistical bulletin. After integrating and checking these sources, missingness across the 35 indicators over the 22-year period ranged from 20% to 50%. **Table 1** summarizes each indicator, including its code, name, unit, and missing-value proportion.

The indicators used in this study are relative measures derived from official socioeconomic statistics, calculated mainly by standardizing raw values by year-end population or administrative area reported in the statistical yearbooks. For example, local government budgetary revenue/expenditure per capita was obtained by dividing revenue/expenditure by year-end population, the density of primary school students was defined as the number of primary school students per unit area. For the health and education dimensions, beyond conventional per capita and area-based measures, we followed Zhao (Zhao et al., 2022) and Wang (Wang et al., 2023) to construct three health resource density index (HRDI) indicators, namely the doctor, hospital bed, and hospital resource density index, and four education resource density index (ERDI) indicators, namely the primary school teacher, primary school, regular secondary school teacher, and regular secondary school resource density indices. These correspond to Y11, Y12, Y13, Y16, Y17, Y21, and Y22 in **Table 1**. The calculation procedures of all indicators are provided in **Table A1**.

Table 1: Detailed information on the 35 socioeconomic indicators. In the main text, Y1-Y35 denote these indicators. HRDI denotes the Health Resource Density Index, and ERDI denotes the Education Resource Density Index.

Number	Socioeconomic variable	Unit	Missing percentage
Y1	Number of doctors per 1,000 population	person	40.00%
Y2	Share of tertiary sector value added in GDP	%	30.22%
Y3	Number of full-time primary school teachers per student	person	33.59%
Y4	Local government budgetary expenditure per capita	yuan	22.81%
Y5	Total public library book holdings per capita	book	32.00%
Y6	Share of primary sector value added in GDP	%	29.76%
Y7	Share of secondary sector value added in GDP	%	30.22%
Y8	Local government budgetary revenue per capita	yuan	25.92%
Y9	Number of hospital beds per 1,000 population	number	26.53%
Y10	Number of hospitals per 1,000 population	number	48.57%
Y11	HRDI for doctors	/	40.00%
Y12	HRDI for hospital beds	/	26.53%

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Number	Socioeconomic variable	Unit	Missing percentage
Y13	HRDI for hospitals	/	48.53%
Y14	Density of primary school students	person/m^2	25.47%
Y15	Number of primary schools per student	number	27.15%
Y16	ERDI for full-time primary school teachers	/	33.61%
Y17	ERDI for primary schools	/	26.97%
Y18	Density of regular secondary school students	person/m^2	28.86%
Y19	Number of full-time regular secondary school teachers per student	person	32.20%
Y20	Number of regular secondary schools per student	number	30.51%
Y21	ERDI for full-time regular secondary school teachers	/	30.65%
Y22	ERDI for regular secondary schools	/	30.51%
Y23	GDP per capita	yuan	21.63%
Y24	Total retail sales of social consumer goods per capita	yuan	26.59%
Y25	Total sales of wholesale and retail trade above the designated threshold per capita	yuan	34.41%
Y26	Density of industrial enterprises above a designated size	number/m^2	31.37%
Y27	Total wage of employed workers per capita	yuan	27.32%
Y28	Average salary of employed workers	yuan	21.75%
Y29	Employee density of enterprises and institutions	person/m^2	33.78%
Y30	Year-end loan balance per capita from financial institutions	yuan	31.94%
Y31	Year-end resident deposit balances per capita	yuan	23.26%
Y32	Total fixed-asset investment per capita	yuan	23.35%
Y33	Year-end mobile phone users per capita	person	32.48%
Y34	Year-end broadband users per capita	person	29.56%
Y35	Population density	person/m^2	25.43%

During the collection and entry, unavoidable human and natural factors may introduce outliers, i.e., observations with obvious extreme values in each city's time series, into the dataset. Such outliers can negatively affect the predictive accuracy of the imputation model. Therefore, to ensure model prediction accuracy and optimize the quality of the final socioeconomic dataset, we carefully examined the time series of each socioeconomic indicator for each city after converting all raw indicators into relative indicators, with the aim of identifying and removing outliers and improving data quality.

2.1.2 Remote Sensing data

After reviewing the literature, we collected 10 remote-sensing-derived variables associated with socioeconomic conditions as auxiliary inputs for imputation. These variables cover meteorological and climatic conditions, vegetation, air pollution, and

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

socioeconomic remote-sensing products. Specifically, meteorological, climatic, and vegetation data were retrieved from the Giovanni platform; nighttime light data came from the global nighttime light dataset published by Li (Li et al., 2020) on Figshare; air pollution data were taken from the high-resolution raster datasets for China available on Zenodo (Wei et al., 2019; Wei et al., 2021); and population density data were obtained from the WorldPop repository. Details on spatial resolution, data source, and temporal coverage are listed in **Table 2**.

Table 2: Remote-sensing-derived auxiliary factors collected in this study, including spatial resolution, time range, and data source information.

Number	Auxiliary variable	Spatial resolution	Time range	Source
X1	Population Density by Raster Type	1 × 1km	2000-2020	WorldPop (https://hub.worldpop.org/)
X2	Nighttime Light	30 arc-seconds	2000-2021	Figshare (https://figshare.com/)
X3	PM_1	1 × 1km	2000-2021	Zenodo (https://zenodo.org/)
X4	$PM_{2.5}$	0.5 × 0.625°	2000-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)
X5	PM_{10}	1 × 1km	2000-2021	Zenodo (https://zenodo.org/)
X6	2-meter Humidity	0.5 × 0.625°	2000-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)
X7	Precipitation	0.5 × 0.625°	2000-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)
X8	2-meter Air Temperature	0.5 × 0.625°	2000-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)
X9	Wind Speed	0.5 × 0.625°	2000-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)
X10	Normalized Difference Vegetation Index (NDVI)	0.05 × 0.05°	2001-2021	Giovanni – NASA (https://giovanni.gsfc.nasa.gov/giovanni/)

For raster-format remote-sensing variables, ArcGIS Pro was used to reproject all datasets to the Albers-Beijing 1954 coordinate system. Subsequently, to ensure consistency and comparability, we resampled each remote-sensing variable to a spatial resolution of 1,000 meters and then aggregated raster statistics to the city scale. Meanwhile, for reducing potential effects arising from differences in variable magnitudes, all remote-sensing auxiliary variables were standardized.

In addition, to assess multicollinearity among the selected remote-sensing factors, we constructed a Pearson correlation matrix and calculated the variance inflation factor (VIF). This screening was performed separately for each socioeconomic

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026


Open Access Earth System Science Data Discussions

indicator. Specifically, for each indicator, variables were iteratively removed based on their VIF values until all remaining variables exhibited VIF values below 5. Accordingly, the auxiliary predictors used in subsequent imputation modeling were indicator-specific, and the resulting auxiliary dataset represents the collection (union) of auxiliary-variable subsets retained across all indicators, which was used to improve prediction accuracy and model reliability.

2.2 Methods

Figure 1 summarizes the workflow for generating a complete city-level socioeconomic dataset and a composite index for China. We first pre-processed 35 socioeconomic indicators and 10 remote-sensing covariates, including auxiliary-variable screening. We then applied a Bayesian spatiotemporal model to impute missing values for each socioeconomic indicator by modeling target-variable spatiotemporal interactions while incorporating remote-sensing information. A multivariable linear regression model served as the benchmark, and model fit and predictive performance were evaluated using multiple criteria. Imputation quality was further examined through spatial error maps and temporal trends. Finally, composite-index weights were derived with the entropy weighting approach, and the resulting index was used to assess spatiotemporal inequality from multiple perspectives. Detailed procedures are provided in the following subsections. Data processing, imputation, index construction, and model evaluation were conducted primarily in R 4.4.1, whereas mapping and visualization were performed in ArcGIS Pro 3.0.0.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

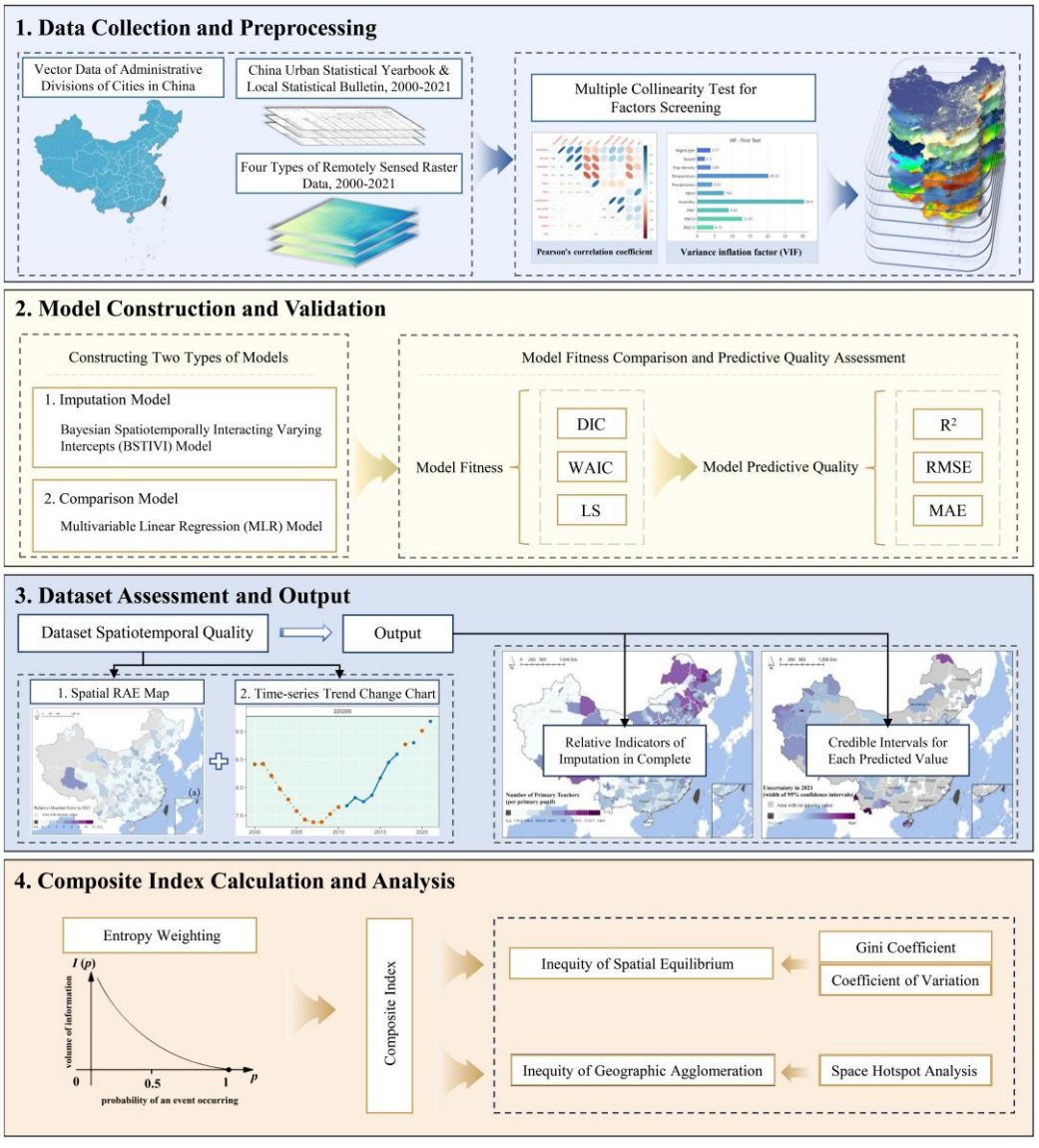


Figure 1: Overview of the methodological workflow for data pre-processing, model evaluation, missing-value imputation, and composite index construction and analysis.

2.2.1 Model construction and validation

To maximize imputation accuracy, we adopted the Bayesian spatiotemporal interacting varying intercepts (BSTIVI) model as the missing-value imputation framework. Designed for spatiotemporal heterogeneity analysis, BSTIVI captures spatial structure, temporal structure, and their coupled autocorrelation in the target variable. Incorporating these dynamic spatiotemporal dependencies into imputation improves predictive reliability and overall accuracy. The BSTIVI model used in this study is formulated as follows:

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

$$log(y_{it}) = \sum_{l=1}^{L} \beta_l X_{itl} + f_{ST}(\omega_{it}) + \alpha + \varepsilon_{it}, \quad (1)$$

$$\omega \sim N_{st}(0, [\tau_\omega R_{st}]^-), \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2), \quad (2)$$

where y_{it} is the observed socioeconomic indicator for unit i at time t. L denotes the total count of remote sensing auxiliary predictors; X_{itl} represents the value of the l-th factor; β_l is the global coefficient of the remote sensing auxiliary factor l. ω_{it} represents the spatiotemporal interaction intercept at the city scale; α represents the global intercept; ε_{it} is the model residual.

For comparison, a Bayesian multivariate linear regression (MLR) model was constructed. Unlike the BSTIVI model, the MLR approach accounts only for stationary contributions from remote-sensing auxiliary factors without incorporating structured spatiotemporal autocorrelations (Wan et al., 2022). Comparing these two models under the same predictor set allows us to quantify the performance gains derived from modeling the target variable's spatiotemporal interactions. The mathematical formulation of the MLR model is given below, and the definitions of its parameters are consistent with those described above:

$$log(y_{it}) = \sum_{l=1}^{L} \beta_l X_{itl} + \alpha + \varepsilon_{it}\ , \quad (3)$$

To evaluate missing-value imputation performance within the Bayesian framework, we employed DIC, WAIC, and Logarithmic Score (LS) across all socioeconomic metrics. DIC and WAIC are widely used criteria for Bayesian model comparison, as they jointly account for model fit and complexity (Du et al., 2023). LS, derived from the conditional predictive ordinate (CPO), serves as a robust measure of model predictive accuracy (Song et al., 2020). For all three metrics, diminished values indicate superior model performance, providing a quantitative basis for the comparative assessment between the BSTIVI and MLR models. Their mathematical formulations are as follows:

$$DIC = \overline{D} + p_D, \quad (4)$$

$$WAIC = -2(\sum_{i=1}^{n} log\, \overline{y_i} - \sum_{i=1}^{n} V(y_i)), \quad (5)$$

$$LS = \frac{-\sum_{i=1}^{n} log(p(\widehat{y_i}|y_i))}{n}, \quad (6)$$

where $\overline{D}$ denotes the mean deviation of the posterior probability distribution, while p_D represents the count of effective parameters. For n training samples, $\overline{y_i}$ and $V(y_i)$ signify the mean and variance of all observations, respectively. $\widehat{y_i}$ indicates the predicted value, with y_i serving as the sampled observed values for model fitting.

To assess predictive performance and generalizability across socioeconomic indicators, a five-fold cross-validation was implemented by partitioning all samples into five equal subsets. In each iteration, a single subset was reserved for testing while the remaining four were integrated for training. Model performance was subsequently recorded and evaluated across all folds. This strategy effectively reduces evaluation bias caused by uneven data partitioning through multiple rounds of training and testing, thereby improving the accuracy and reliability of model evaluation (Shao, 1993; Arlot and Celisse, 2010).

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Model predictive performance was further benchmarked using R^2, RMSE, and MAE. R^2 was employed to assess explanatory power, while RMSE and MAE quantified the magnitude of prediction errors. For these metrics, higher R^2 and lower error values (RMSE/MAE) denote superior predictive fidelity. Detailed calculation protocols for these indicators follow established statistical standards (Mckeen et al., 2023; Pezzulo et al., 2023).

2.2.2 Dataset assessment and output

To evaluate the actual performance at the dataset level, relative absolute error (RAE) was employed to quantify discrepancies between actual and predicted values. The calculation is formulated as:

$$RAE_{ij} = \left|\frac{(y_{ij} - \widehat{y_{ij}})}{y_{ij}}\right|, \quad (7)$$

where RAE_{ij} denotes the error for socioeconomic indicator j in spatial unit i, while y_{ij} and $\widehat{y_{ij}}$ signify the ground-truth observation and model-derived prediction, respectively.

Beyond spatial analysis, temporal imputation fidelity was verified by constructing 22-year time-series plots for representative indicators. These visualizations, overlaying original yearbook data with model-imputed series, facilitate the identification of potential anomalies. Following this, the predicted values for missing entries were integrated into the panel dataset, accompanied by their 50% and 95% Bayesian credible intervals.

2.2.3 Composite index calculation and analysis

The entropy weight method, grounded in information entropy theory (Pamucar et al., 2022), was utilized to objectively determine indicator weights. By quantifying data variability and information content, this approach assigns greater weights to indicators with lower information entropy, thereby minimizing subjective bias and enhancing the consistency of the evaluation results. Using 35 fully imputed socioeconomic indicators, we applied entropy weighting to derive indicator-specific weights and construct a city-level composite index of socioeconomic development. After standardization, information entropy was computed for each indicator, weights were estimated accordingly, and the weighted sum of standardized values yielded the final index.

The Gini coefficient quantifies inequality within and across regions for variables of interest (Martin and Conway, 2025), ranging from 0 to 1, with lower values indicating greater equality and higher values indicating stronger inequality. In general, a Gini coefficient below 0.3 indicates relatively equitable conditions between regions; values between 0.3 and 0.4 are considered moderately reasonable; and values above 0.4 indicate a high degree of inequality. The coefficient of variation, a standardized measure of dispersion, effectively reflects the magnitude of differences in socioeconomic development levels between regions, indicating the severity of regional socioeconomic inequality. Higher values indicate greater inequality.

Spatial hotspot analysis identifies local clusters of high and low values in geographic data (Fischer et al., 2010). We computed the Getis-Ord Gi* statistic for each spatial unit and evaluated clustering using z-scores and p-values. Significantly

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

positive z-scores indicate hotspots, whereas significantly negative z-scores indicate cold spots. Significance was classified at the 90%, 95%, and 99% confidence levels, with higher levels indicating stronger statistical evidence.

Although the Gini coefficient and the coefficient of variation can quantify overall disparity, they cannot identify the specific spatial distribution of inequality. By contrast, spatial hotspot analysis reveals spatial clustering patterns but does not quantify overall inequality levels. We therefore combined these methods to assess socioeconomic inequality among Chinese cities across and within administrative scales. The Gini coefficient and coefficient of variation were used to quantify overall disparity, whereas hotspot analysis identified spatiotemporal clustering in socioeconomic development. This integrated framework evaluates inequality from both distributional and geographic-clustering perspectives. Specifically, the Gini coefficient was estimated using the *ineq* package in R 4.4.1, the coefficient of variation was computed in R 4.4.1, and spatial hotspot analysis was performed in ArcGIS Pro 3.0.0.

3 Results

3.1 Evaluation of models and imputation performance

3.1.1 Screening of remote sensing auxiliary factors

The Pearson correlation coefficient matrix for the 35 socioeconomic relative indicators is provided in **Fig. A1**. Using the local government budgetary expenditures per capita (Y4) as an example, **Table 3** summarizes the screening of remote-sensing covariates. For Y4, four rounds of VIF screening were performed, with a threshold of 5. During this process, two-meter humidity (X6), PM_{10} (X5), and NDVI (X10) were removed in the first three rounds. The same iterative screening procedure was applied to each socioeconomic indicator. Ultimately, an auxiliary dataset consisting of seven variables was retained for model construction, including population density (X1), nighttime lights (X2), PM_1 (X3), $PM_{2.5}$ (X4), precipitation (X7), two-meter air temperature (X8), and wind speed (X9).

Table 3: Variance inflation factor (VIF)-based multicollinearity screening results for indicator Y4 (local government budgetary expenditures per capita).

Remote Sensing Factors	First Test	Second Test	Third Test	Fourth Test
Pop_density	4.22	4.10	4.08	4.06
Nighttime Light	4.28	4.19	4.18	3.86
2-meter Air Temperature	21.47	3.71	3.71	2.27
Precipitation	4.24	3.05	2.93	2.88
Wind Speed	2.26	2.21	2.20	1.52
2-meter Humidity	32.22	×	×	×
$PM_{2.5}$	4.26	4.00	2.77	1.30

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

PM_{10}	17.57	16.77	×	×
PM_1	11.93	11.92	1.39	1.35
NDVI	7.92	7.91	7.15	×

3.1.2 Comparison of missing-value imputation models

Because the GDP shares of the primary, secondary, and tertiary sectors (Y6, Y7, and Y2) sum to 100% in official statistics, and Y7 showed the poorest imputation performance in the experiment, we imputed Y6 and Y2 using the model-based approach and derived Y7 as 100% − Y6 − Y2. Accordingly, Y7 was excluded from the model-comparison results in this subsection and from the cross-validation results in **Sect. 3.1.3**.

Figure 2 presents the model comparison results for socioeconomic indicators Y1-Y6, while the corresponding results for the remaining indicators are provided in **Fig. A2**. Across all three global evaluation metrics, the BSTIVI model consistently yields lower DIC and WAIC values than the conventional MLR model, indicating better model fit. In addition, the LS of the BSTIVI model is also markedly lower than that of the MLR model, further supporting superior predictive performance. Overall, by accounting for temporal, spatial, and spatiotemporal interaction autocorrelation in the target variable, BSTIVI substantially outperformed the traditional MLR model. These findings demonstrate that incorporating structured autocorrelation information of the target variable is essential for improving imputation performance.

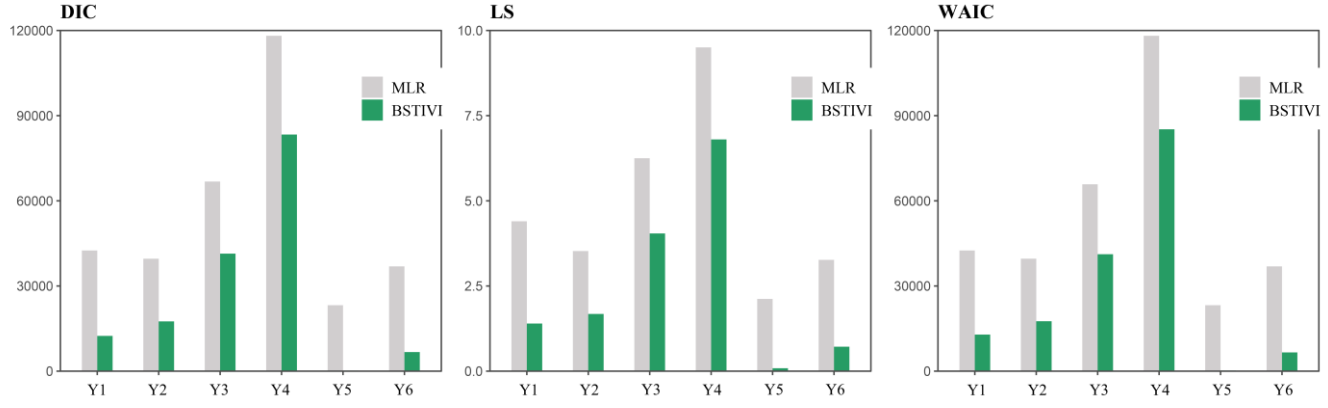


Figure 2: Model performance comparison between the multivariate linear regression (MLR) and Bayesian spatiotemporal interacting varying intercepts (BSTIVI) models for indicators Y1-Y6, based on three Bayesian evaluation metrics (DIC, WAIC, and LS). Indicator definitions: Y1, number of doctors per 1,000 population; Y2, share of tertiary sector value added in GDP; Y3, number of full-time primary school teachers per student; Y4, local government budgetary expenditures per capita; Y5, total public library book holdings per capita; Y6, share of primary sector value added in GDP.

3.1.3 Cross-validation

To evaluate imputation accuracy, we calculated evaluation metrics for model-fitting and validation data and examined scatterplots of observed versus predicted values. Given the large number of socioeconomic indicators, **Fig. 3** presents cross-validation results for the first six indicators, with **Fig. 3(a)** showing the model-fitting subset and **Fig. 3(b)** the validation

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026


Open Access Earth System Science Data Discussions

subset, while results for the remaining 28 indicators are shown in **Fig. A3** and **A4**. Results from model fitting indicate consistently high R^2 values and relatively low MAE and RMSE across all socioeconomic indicators. The close correspondence between observed and predicted values further supports the good fit of the BSTIVI model.

For the validation subset, observed and predicted values were less tightly clustered than in the model-fitting subset, although the fitted lines still closely followed the 1:1 reference. Among the 34 socioeconomic indicators, 32 achieved $R^2 >= 0.95$, and the remaining two also performed well, with R^2 of 0.93 for total retail sales of social consumer goods per capita (Y24) and 0.89 for broadband users per capita (Y34). This pattern indicates robust out-of-sample predictive performance. Validation-subset MAE and RMSE were uniformly low, and their small differences from the model-fitting-subset values further suggest that overfitting was minimal. Overall, BSTIVI appears effective for missing-value imputation in long-term socioeconomic time-series data while maintaining high predictive accuracy.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

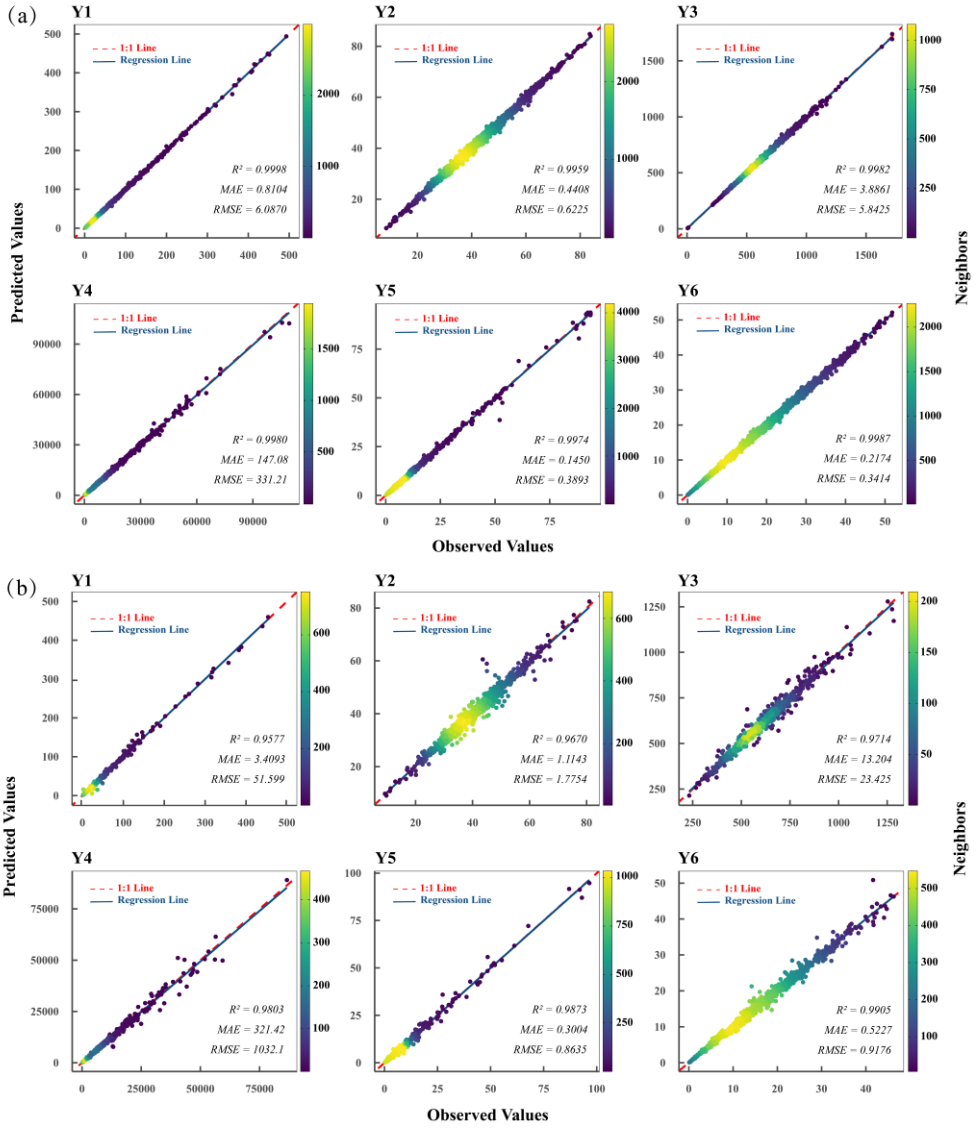


Figure 3: Observed versus predicted scatter plots and evaluation metrics for indicators Y1-Y6 based on five-fold cross-validation: (a) training set and (b) test set. Indicator definitions: Y1, number of doctors per 1,000 population; Y2, share of tertiary sector value added in GDP; Y3, number of full-time primary school teachers per student; Y4, local government budgetary expenditures per capita; Y5, total public library book holdings per capita; Y6, share of primary sector value added in GDP.

3.1.4 Dataset imputation quality assessment

In addition to five-fold cross-validation, we calculated the relative absolute error (RAE) for each city using all available samples during imputation and mapped the results to show city-level variation in prediction error. The first six socioeconomic indicators were used as representative examples. As shown in **Fig. 4**, panels (a)-(f) present the spatial distributions of RAE for Y1-Y6 in 2021, respectively.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

The spatial maps indicate generally high predictive accuracy for Y1-Y6. Except for Y5, prediction errors were below 1% in many cities and below 2% in nearly all cities, indicating close correspondence between estimated and observed values and thus strong model performance. Although Y5 performed slightly worse than the other indicators, most cities still had errors below 10%, and only four exceeded 10%, all remaining under 15%. Overall, prediction errors stayed within acceptable limits, supporting the effectiveness of the spatiotemporal imputation model used in this study.

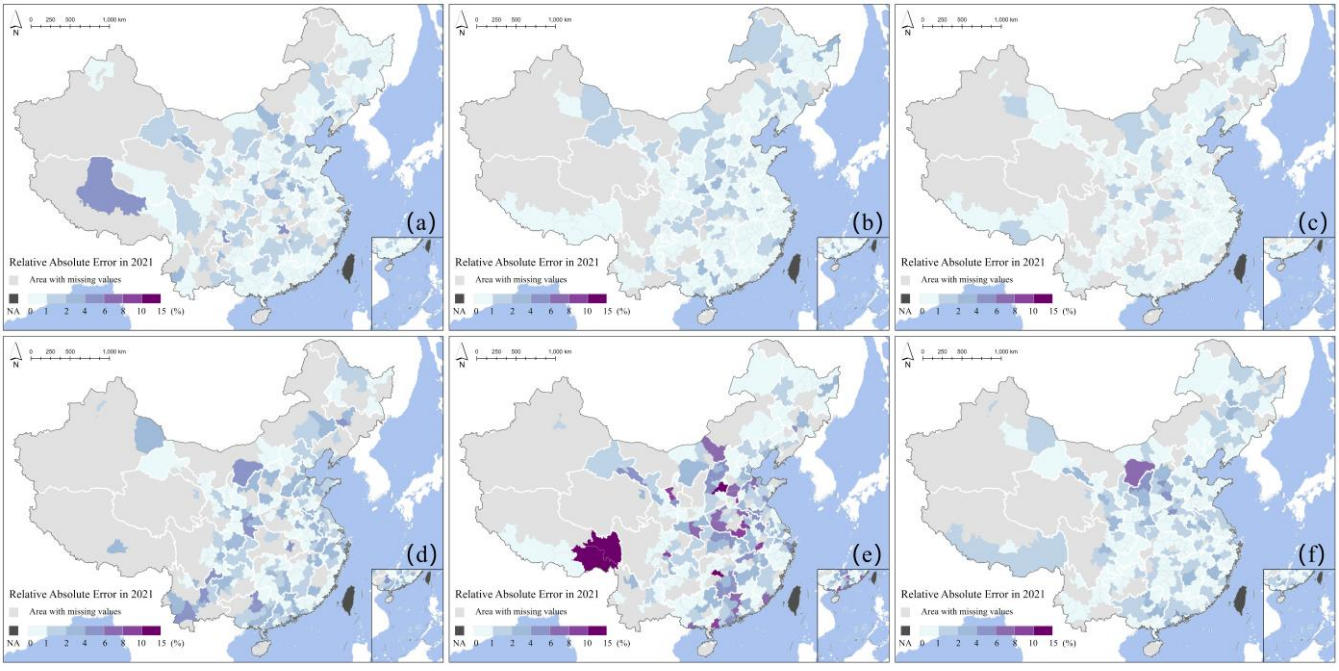
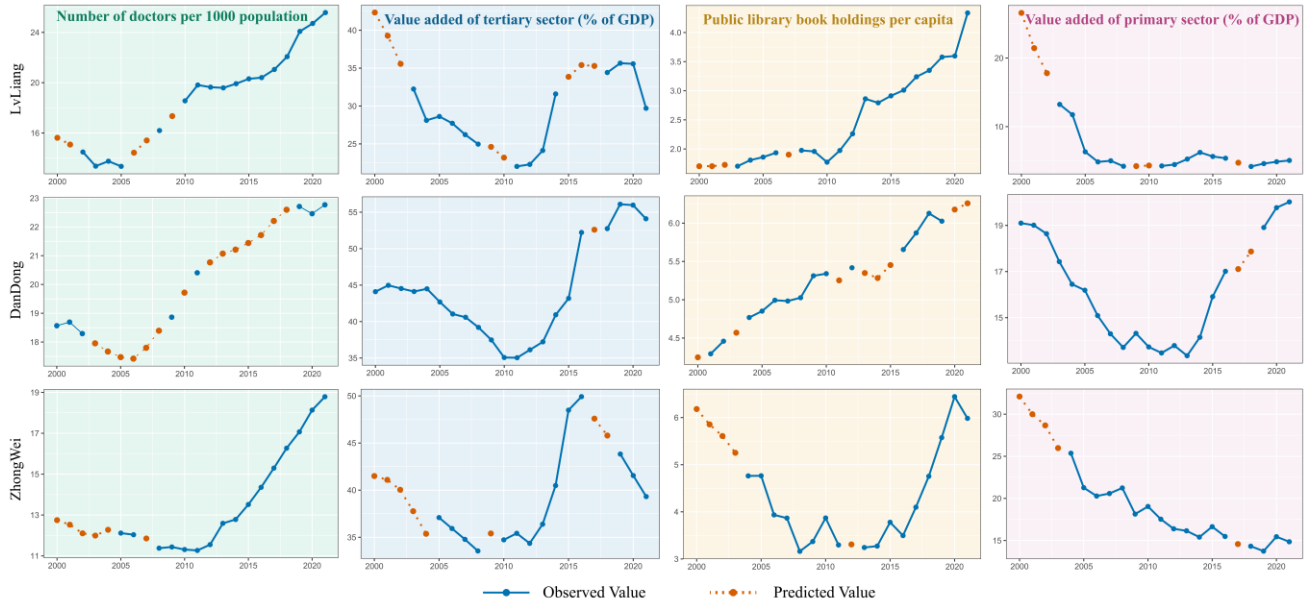


Figure 4: City-level spatial distributions of relative absolute error (RAE) for indicators Y1-Y6 in 2021: (a) Y1, (b) Y2, (c) Y3, (d) Y4, (e) Y5, and (f) Y6. Indicator definitions: Y1, number of doctors per 1,000 population; Y2, share of tertiary sector value added in GDP; Y3, number of full-time primary school teachers per student; Y4, local government budgetary expenditures per capita; Y5, total public library book holdings per capita; Y6, share of primary sector value added in GDP.

To further assess time-series imputation quality, we randomly selected three cities with missing records and plotted post-imputation temporal trajectories for four indicators, namely the number of doctors per 1,000 population (Y1), the share of tertiary sector value added in GDP (Y2), the total public library book holdings per capita (Y5), and the share of primary sector value added in GDP (Y6) (**Fig. 5**). Across indicators and cities, the imputed values (orange points) showed no obvious outliers and closely tracked the temporal trajectories of the original yearbook data (blue points), further supporting the robustness of the proposed method for time-series imputation.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026



305 **Figure 5: Time-series profiles of imputed relative indicators Y1, Y2, Y5, and Y6 for three cities (Lvliang, Dandong, and Zhongwei). Indicator definitions: Y1, number of doctors per 1,000 population; Y2, share of tertiary sector value added in GDP; Y5, total public library book holdings per capita; Y6, share of primary sector value added in GDP.**

3.1.5 Construction of complete socioeconomic dataset

In the final socioeconomic data product, all indicators are complete and reported as directly comparable relative measures. 310 These indicators can thus be used directly without further pre-processing. In addition, whereas frequentist imputation methods do not explicitly characterize uncertainty in predicted values, the Bayesian framework adopted here naturally yields credible intervals that provide interpretable measures of prediction uncertainty (Kaplan, 2025). The released dataset therefore contains not only complete relative indicators but also both wide (95%) and narrow (50%) credible intervals for each predicted value. For illustration, **Fig. 6** shows the 2021 predictions for the 40 cities with missing Y3 values, together with 315 their corresponding wide and narrow credible intervals.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

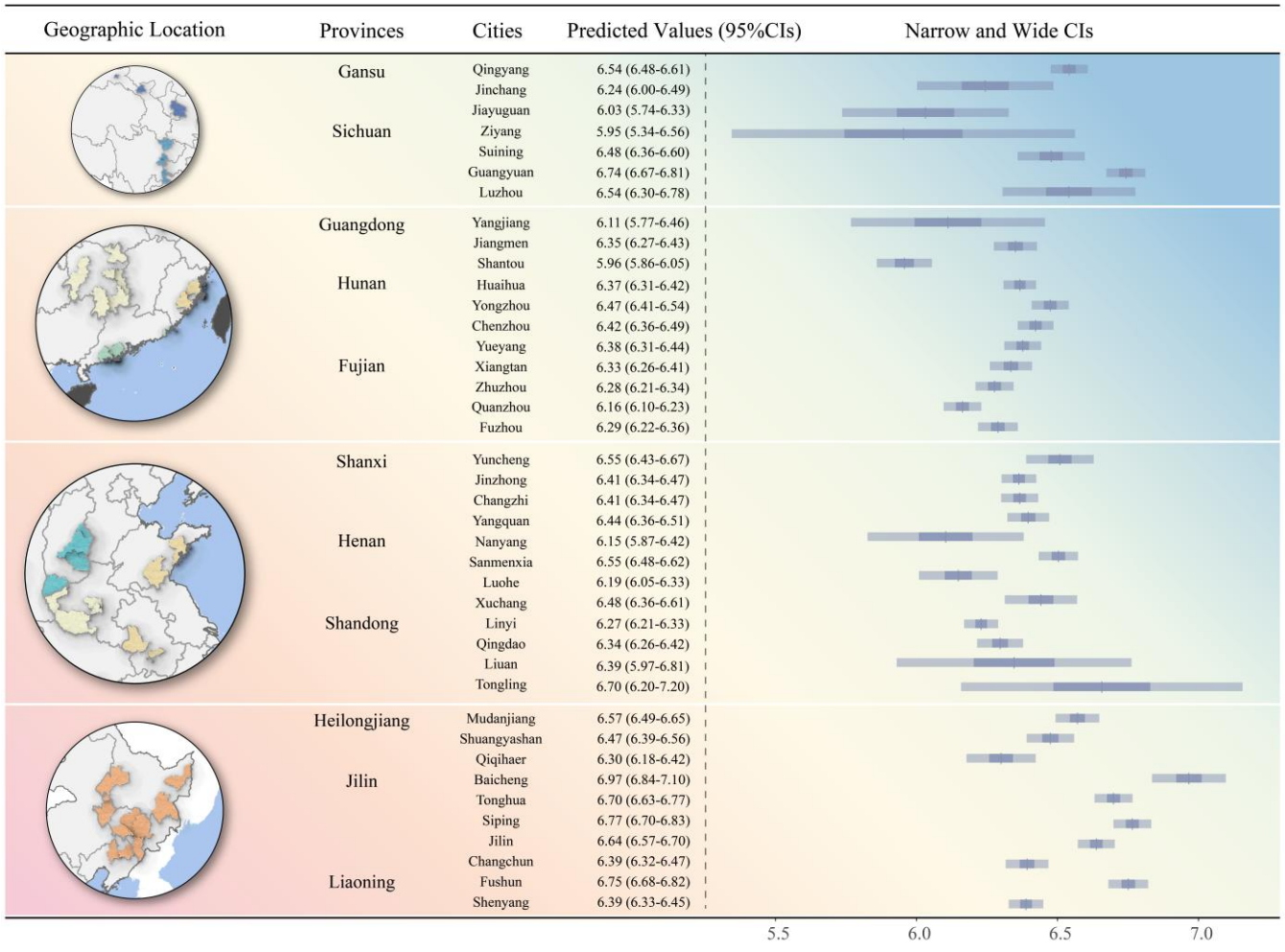


Figure 6: Predicted values and Bayesian credible intervals in 2021 for the 40 cities with missing observations for indicator Y3 (number of full-time primary school teachers per student). Predictions are generated on the log scale, back-transformed values represent the number of full-time primary school teachers per 10,000 pupils. Credible intervals are reported as narrow (50%) and wide (95%).

3.2 Assessment of socioeconomic development level

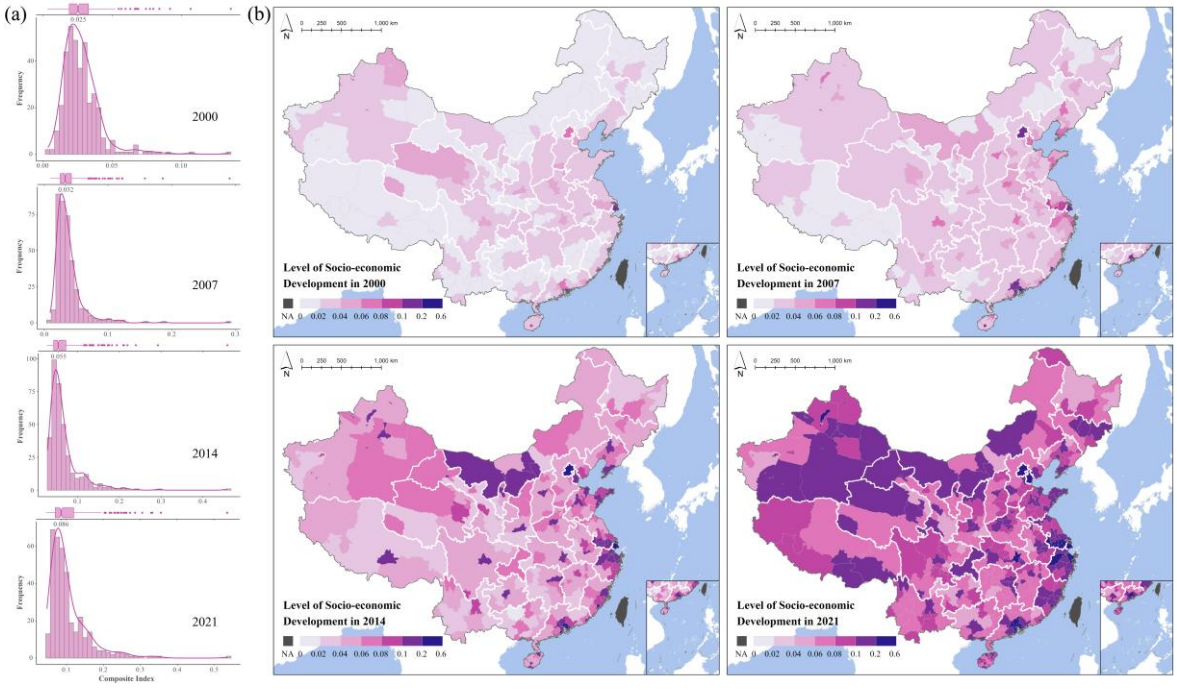
3.2.1 Socioeconomic development index calculation and spatiotemporal patterns

Because several indicators within the same domain were highly correlated, such as the number of hospitals per 1,000 population (Y10) versus the HRDI for hospitals (Y13), including them simultaneously could introduce redundancy and overweight overlapping information in entropy weighting. We therefore compared similar indicators in terms of spatial pattern, substantive meaning, and coverage, and removed 10 accordingly. Entropy weights were then estimated for the remaining 25 socioeconomic indicators. The largest weights were assigned to the HRDI for hospital beds (Y12), the HRDI for doctors (Y11), and the year-end loan balance per capita from financial institutions (Y30), whereas the smallest were

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

assigned to the density of industrial enterprises above a designated size (Y26) and employee density of enterprises and institutions (Y29). Detailed weight calculation results for the 25 selected socioeconomic indicators are provided in **Table B1**. Using the weights reported in **Table B1**, we derived the composite index and mapped its city-level distribution for 2000, 2007, 2014, and 2021. According to the data characteristics, index values were grouped into seven classes (**Fig. 7**). In terms of magnitude, the median index increased from 0.025 (range: 0.003-0.135) in 2000 to 0.086 (range: 0.046-0.535) in 2021, indicating a marked rise in the overall socioeconomic development level of Chinese cities. The annual frequency curves further show that the modal frequency increased from about 50 in 2000 to around 100 in 2014, then declined to about 70 in 2021. Over the same period, the distribution first widened, then narrowed, and later widened again, suggesting a transition from a relatively even pattern to stronger polarization between low- and high-development cities, followed by partial rebalancing. Together, these changes reflect the dynamic evolution of regional socioeconomic inequality among Chinese cities.

Spatially, city-level socioeconomic development showed persistent regional disparities throughout the study period. The highest development levels during 2000-2021 were concentrated in major metropolitan areas, including Beijing, Shanghai, Guangzhou, Shenzhen, and Dongguan, all far above the national average. More broadly, eastern China consistently exhibited higher development levels than western China, with high-value cities forming clear spatial clusters, especially in coastal provinces. In many other provinces, provincial capitals generally served as local high-value centers and substantially outperformed other cities within the same province. By contrast, several cities in Inner Mongolia, Yunnan, and the Tibet Autonomous Region remained in the relatively low-development group over the 22-year period, although their development levels improved steadily and their gap with other cities gradually narrowed.



https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Figure 7: Composite index for 2000, 2007, 2014, and 2021: (a) summary statistics and frequency distribution plots and (b) city-level spatial distribution maps.

3.2.2 Spatiotemporal inequality analysis of socioeconomic development level

For city-level spatiotemporal panel data of the composite index, we examined national- and provincial-scale time-series of the Gini coefficient and coefficient of variation, together with provincial maps of the Gini coefficient, to characterize spatial disparities in socioeconomic development.

At the national level, the Gini coefficient trend (**Fig. 8(a)**) rose initially and then declined over 2000-2021, indicating that nationwide socioeconomic inequality first worsened and later eased. Provincial trajectories, however, were more heterogeneous and displayed both linear and nonlinear patterns. Guangdong was the only province with Gini coefficients above 0.4 during 2005-2019, indicating pronounced internal inequality. Trends in the coefficient of variation (**Fig. B1**) closely matched those of the Gini coefficient at both national and provincial scales, suggesting that the two measures yield consistent assessments of inequality. At the same time, differences between national and provincial trajectories show that national summaries can obscure substantial subnational heterogeneity, underscoring the value of the city-level dataset compiled here.

Figure 8(b) maps provincial Gini coefficients for 2000 and 2021 using six classes, with darker shades indicating greater spatial inequality. In both years, Gini coefficients were consistently higher in southeastern coastal and northwestern inland provinces than in central China. Specifically, provinces with relatively high spatial inequality included the economically developed Guangdong Province as well as the less developed Tibet, Inner Mongolia, and Xinjiang region. Among them, Guangdong exhibited higher Gini coefficients than the other three provinces in both years. Nevertheless, time-series analysis reveals a clear declining trend in Gini coefficients for all four provinces, indicating that provincial-level spatial inequality had declined substantially by 2021.

Figure 8(c) shows hotspot patterns of city-level socioeconomic development in 2000, 2007, 2014, and 2021. The results indicate persistent spatial clustering, with distinct hotspot and cold-spot agglomerations and a broadly stable overall configuration. Hotspots were concentrated mainly along the southeastern coast and in several cities in Xinjiang, whereas cold spots were located primarily in southwestern and northeastern China. In addition, comparing the hotspot and cold-spot distributions between 2000 and 2021 shows a notable reduction in the number of both hotspot and cold-spot cities, suggesting a gradual mitigation of geographically clustered inequality.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

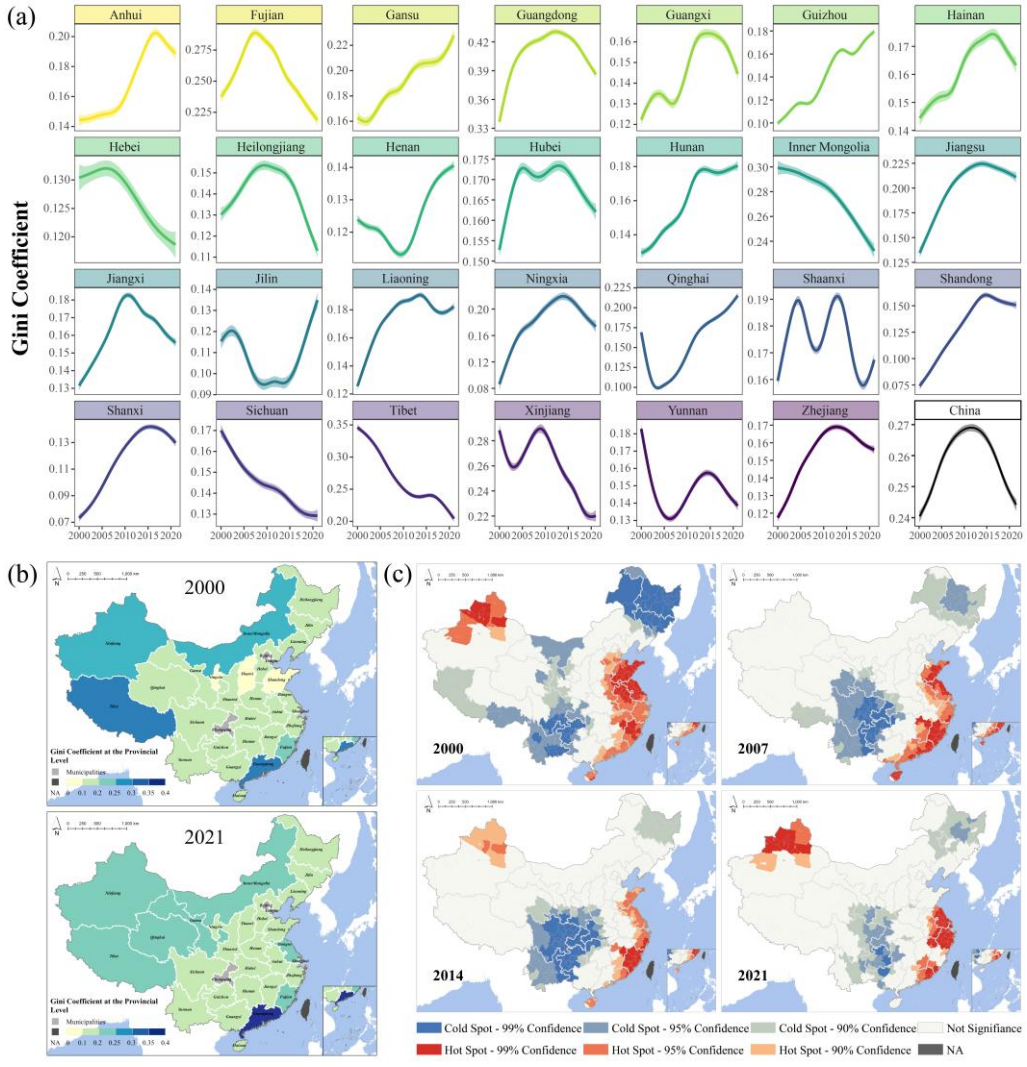


Figure 8: Spatiotemporal inequality patterns of the composite index: (a) temporal trends in the Gini coefficient (national and provincial scales), (b) provincial-level spatial distributions of the Gini coefficient in 2000 and 2021, and (c) hotspot analysis of the composite index in 2000, 2007, 2014, and 2021.

4 Discussion

The growing prominence of urban geography and urban economics, together with initiatives such as the WHO Healthy Cities Programme and SDG 11, highlights the city as a critical scale for spatially targeted policymaking and resource allocation. Although the UN has promoted the collection and utilization of socioeconomic data worldwide through various collaborations and agendas (Murphy, 2022), the availability and timeliness of official statistics at subnational scales remain major challenges (Allen et al., 2021). These limitations continue to constrain the temporal continuity and spatial coverage of officially released socioeconomic datasets. In this study, without relying on additional samples or conventional survey-based

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

auxiliary data, we addressed spatiotemporal missingness in China's official socioeconomic statistics by integrating satellite remote-sensing data into a spatiotemporal modeling framework. This approach enabled the completion of 35 socioeconomic indicators for 366 cities over a 22-year period. Multiple validation analyses confirm the reliability and stability of the proposed method. Using the completed dataset, we further constructed and analyzed a composite index of city-level socioeconomic development, allowing a systematic assessment of spatiotemporal inequality among Chinese cities over the study period. In summary, this study makes three main contributions: (1) providing the first complete long-term, multi-indicator urban socioeconomic dataset for China; (2) developing a composite index for assessing urban socioeconomic development and inequality; (3) proposing a generalizable imputation paradigm for addressing missing values in small-area socioeconomic data. These contributions are discussed below.

First, the urban socioeconomic dataset developed in this study constitutes the most complete publicly available long-term, multidimensional city-level socioeconomic data product currently available for China. Although the National Bureau of Statistics of China provides relatively complete statistics at the provincial level, official data at finer administrative scales, particularly cities, remain severely affected by spatiotemporal missingness, and no public dataset has previously offered comprehensive multi-indicator coverage. The dataset assembled here therefore provides an important foundation for long-term, multidisciplinary research on Chinese cities, facilitates cross-regional monitoring of socioeconomic development, and supports the design and evaluation of localized policy interventions. It also strengthens the evidence base for translating macro-level initiatives, such as the Healthy China Strategy and the SDGs, into operational urban policy actions. In addition, all indicators are expressed as relative measures, which improves interpretability and comparability across space and time, reduces user-side pre-processing, and enhances overall usability and accessibility.

Second, beyond providing complete values for 35 socioeconomic indicators together with uncertainty estimates, this study also develops a composite index. By integrating multiple dimensions, including economic activity, health resources, and educational conditions, the index provides a more comprehensive basis for assessing development levels and regional disparities among cities. Indicator weights were determined using an objective entropy-based approach, thereby limiting subjective influence, and more closely reflecting underlying data variability. Like the individual indicators, the composite index is a relative measure, enabling direct comparison across cities and over time. Owing to its comprehensiveness, objectivity, and comparability, the index offers a practical and efficient tool for analyzing urban socioeconomic dynamics and informing evidence-based decision-making.

Third, this study proposes a transferable imputation framework for addressing missing socioeconomic data in small-area settings worldwide. In many countries, filling spatiotemporal gaps in official socioeconomic statistics through additional surveys or auxiliary census data is often infeasible because of financial and institutional constraints. Satellite remote sensing offers an alternative source of auxiliary information by capturing spatial signatures associated with socioeconomic conditions (Palacios-Lopez et al., 2019; Leasure et al., 2020). Yet remote-sensing data have rarely been used to directly impute missing values in official socioeconomic datasets. For long-term geographic data with spatiotemporal missingness, effective imputation requires jointly leveraging the target variable's structured spatiotemporal dependence and the

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

explanatory information provided by auxiliary covariates. Previous studies have shown that both components can substantially improve imputation accuracy (Song et al., 2018). By integrating remote-sensing-derived auxiliary data with a spatiotemporal imputation model, this study helps bridge this methodological gap and extends both the technical workflow and its practical utility. Overall, this method offers a broadly applicable solution for gap-filling small-area socioeconomic statistics and constructing complete socioeconomic time-series datasets.

Despite the strong predictive performance of the proposed method, several limitations should be acknowledged. First, the current spatiotemporal model incorporates autocorrelation in the target variables but does not explicitly model spatiotemporal heterogeneity (i.e., non-stationarity) in the remote-sensing-derived covariates. This omission may reduce predictive accuracy for cities with extreme values. Future work will therefore explore the Bayesian spatiotemporal interacting varying coefficient (BSTIVC) model to incorporate covariate non-stationarity into the imputation framework and compare its performance. Second, this study generates socioeconomic data products at the city scale. Although this scale is well suited to many macro-level policy analyses and multidisciplinary applications, socioeconomic and geographic structures within Chinese cities are often highly complex and heterogeneous. City-level data may therefore fail to capture finer intra-urban variation, limiting the precise identification and targeted mitigation of within-city development disparities. Future research should thus prioritize the development of higher-resolution socioeconomic datasets to support more nuanced and spatially targeted policies for equitable and coordinated regional development.

5 Conclusions

This study addresses spatiotemporal missingness in officially released small-area socioeconomic indicators by integrating satellite remote-sensing data with a Bayesian spatiotemporal interacting varying intercepts model, thereby generating a complete city-level socioeconomic dataset for China containing 35 indicators for 2000-2021. To compare the performance of spatiotemporal and conventional global imputation models, we conducted a systematic evaluation of both approaches. The results show that incorporating structured temporal and spatial autocorrelation in the target variables is essential for improving imputation performance. Comprehensive assessment of model accuracy and completed-data quality further indicates that the proposed spatiotemporal framework achieves high predictive accuracy in Chinese urban settings. Using the fully imputed indicators, we also constructed a composite index, included in the final data product as the 36th indicator. Analyses based on the Gini coefficient, the coefficient of variation, and hotspot detection reveal marked spatiotemporal inequality in socioeconomic development across Chinese cities, with clear regional heterogeneity. Overall, under severe missingness in official small-area statistics, this study provides a practical and transferable framework for informing targeted regional policymaking and advancing progress toward the SDGs.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Open Access Earth System Science Data Discussions

6 Appendices

6.1 Appendix A

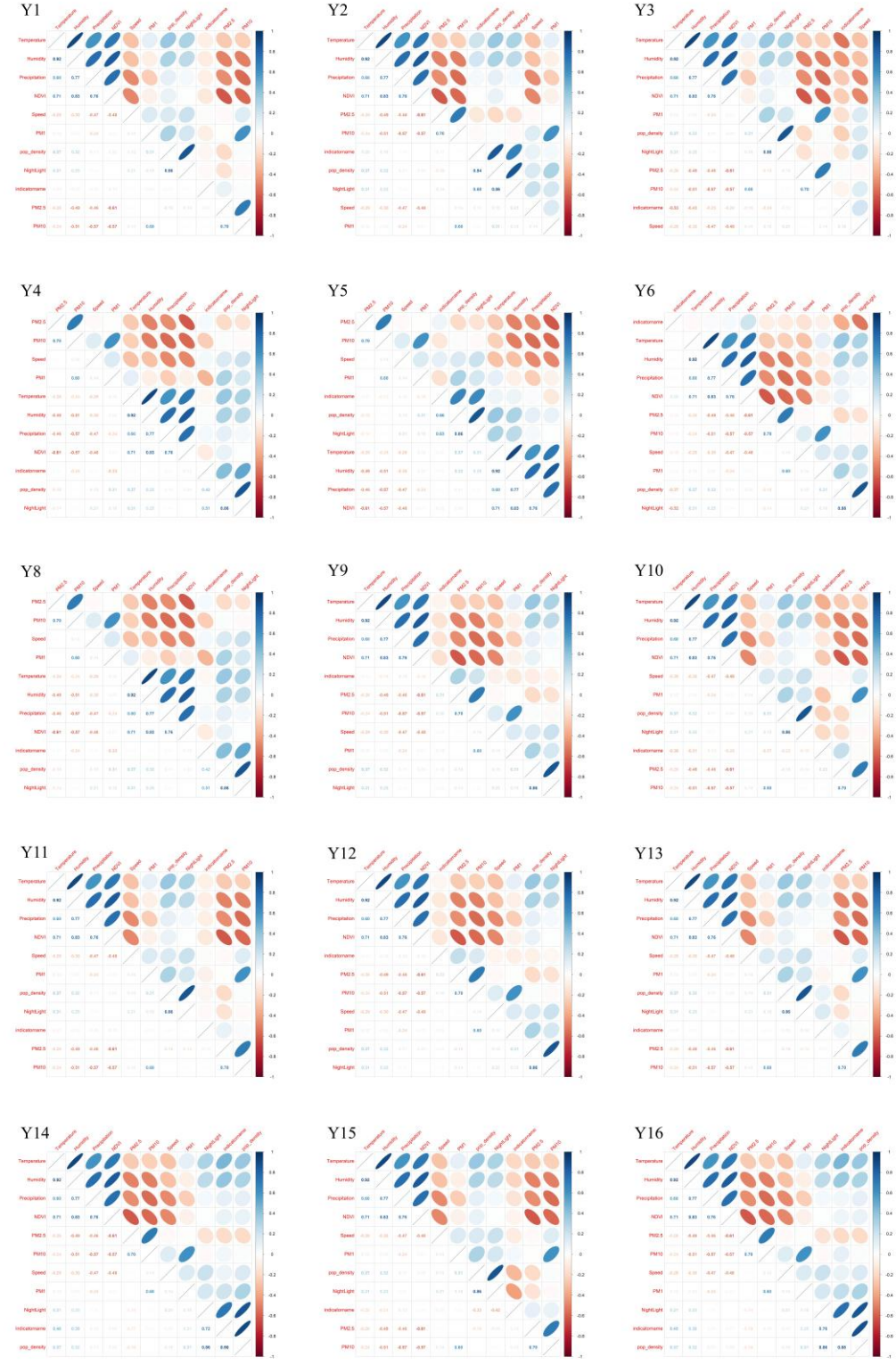
Table A1: Calculation methods for the 35 relative socioeconomic indicators for Chinese cities.

Number	Socioeconomic variable	Calculation method
Y1	Number of doctors per 1,000 population	Number of doctors / the population
Y2	Share of tertiary sector value added in GDP	Untreated
Y3	Number of full-time primary school teachers per student	Number of full-time primary school teachers / number of primary school students
Y4	Local government budgetary expenditures per capita	Local government general budgetary expenditure / the population
Y5	Total public library book holdings per capita	The number of books in public libraries / the population
Y6	Share of primary sector value added in GDP	Untreated
Y7	Share of secondary sector value added in GDP	Untreated
Y8	Local government budgetary revenue per capita	Local government general budgetary revenue / the population
Y9	Number of hospital beds per 1,000 population	Number of hospital beds / the population
Y10	Number of hospitals per 1,000 population	Number of hospitals / the population
Y11	HRDI for doctors	((Number of doctors / the population) * (Number of doctors / the administrative area))$^{1/2}$
Y12	HRDI for hospital beds	((Number of hospital beds / the population) * (Number of hospital beds / the administrative area))$^{1/2}$
Y13	HRDI for hospitals	((Number of hospitals / the population) * (Number of hospitals / the administrative area))$^{1/2}$
Y14	Density of primary school students	Number of primary school students / the administrative area
Y15	Number of primary schools per student	Number of primary schools / number of primary school students
Y16	ERDI for full-time primary school teachers	((Number of full-time primary school teachers / number of primary school students) * (Number of full-time primary school teachers / the administrative area))$^{1/2}$
Y17	ERDI for primary schools	((Number of primary schools / number of primary school students) * (Number of primary schools / the administrative area))$^{1/2}$
Y18	Density of regular secondary school students	Number of regular secondary school students / the administrative area
Y19	Number of full-time regular secondary school teachers per student	Number of full-time regular secondary school teachers / number of regular secondary school students
Y20	Number of regular secondary schools per student	Number of regular secondary schools / number of regular secondary school students
Y21	ERDI for full-time regular secondary school teachers	((Number of full-time regular secondary school teachers / number of regular secondary school students) * (Number of full-time regular

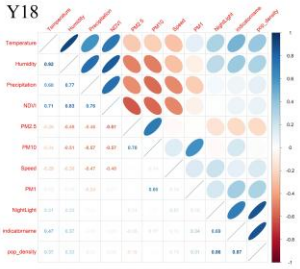
https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

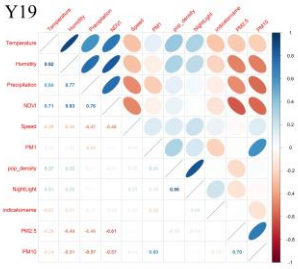
Number	Socioeconomic variable	Calculation method
		secondary school teachers / the administrative area))$^{1/2}$
Y22	ERDI for regular secondary schools	((Number of regular secondary schools / number of regular secondary school students) * (Number of regular secondary schools / the administrative area))$^{1/2}$
Y23	GDP per capita	Untreated
Y24	Total retail sales of social consumer goods per capita	Total retail sales of social consumer goods / the population
Y25	Total sales of wholesale and retail trade above the designated threshold per capita	Total value of goods sold in wholesale and retail trade above the threshold / the population
Y26	Density of industrial enterprises above a designated size	Number of industrial enterprises above a designated size / the administrative area
Y27	Total wage of employed workers per capita	Total wage of employed workers / the population
Y28	Average salary of employed workers	Untreated
Y29	Employee density of enterprises and institutions	Number of employed personnel at year-end in the organization / the administrative area
Y30	Year-end loan balance per capita from financial institutions	Year-end loan balance from financial institutions / the population
Y31	Year-end resident deposit balances per capita	Year-end savings balance for residents / the population
Y32	Total fixed-asset investment per capita	Total fixed asset investment / the population
Y33	Year-end mobile phone users per capita	Year-end mobile phone user count / the population
Y34	Year-end broadband users per capita	Year-end broadband user count / the population
Y35	Population density	Total population at year-end / the administrative area

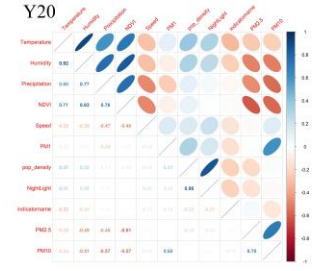
https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

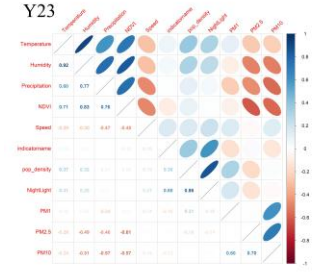


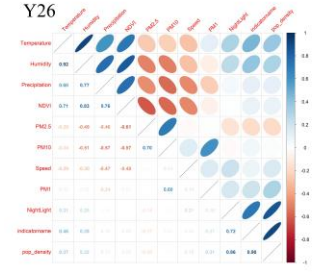
https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

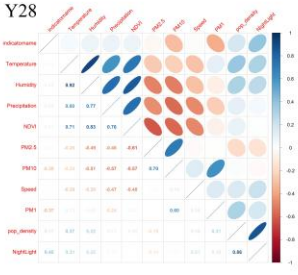


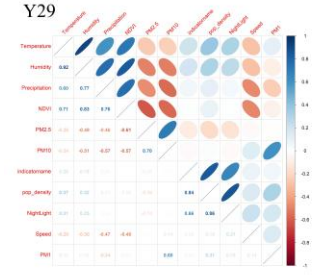


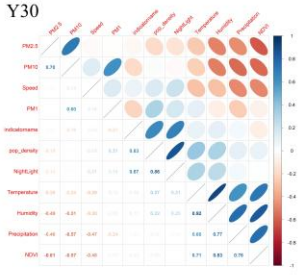


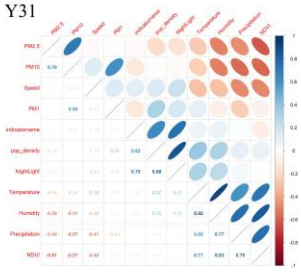












455

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

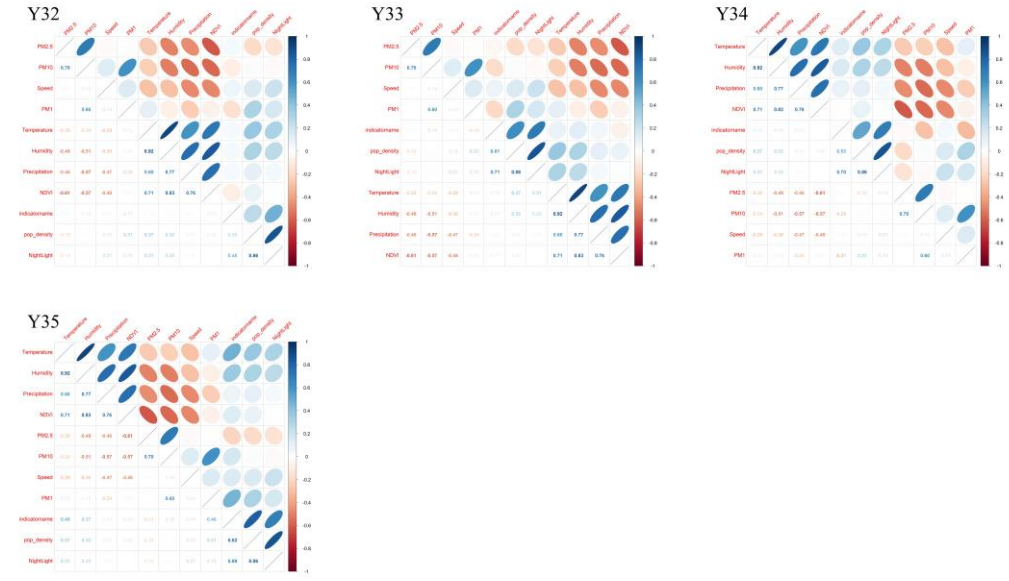
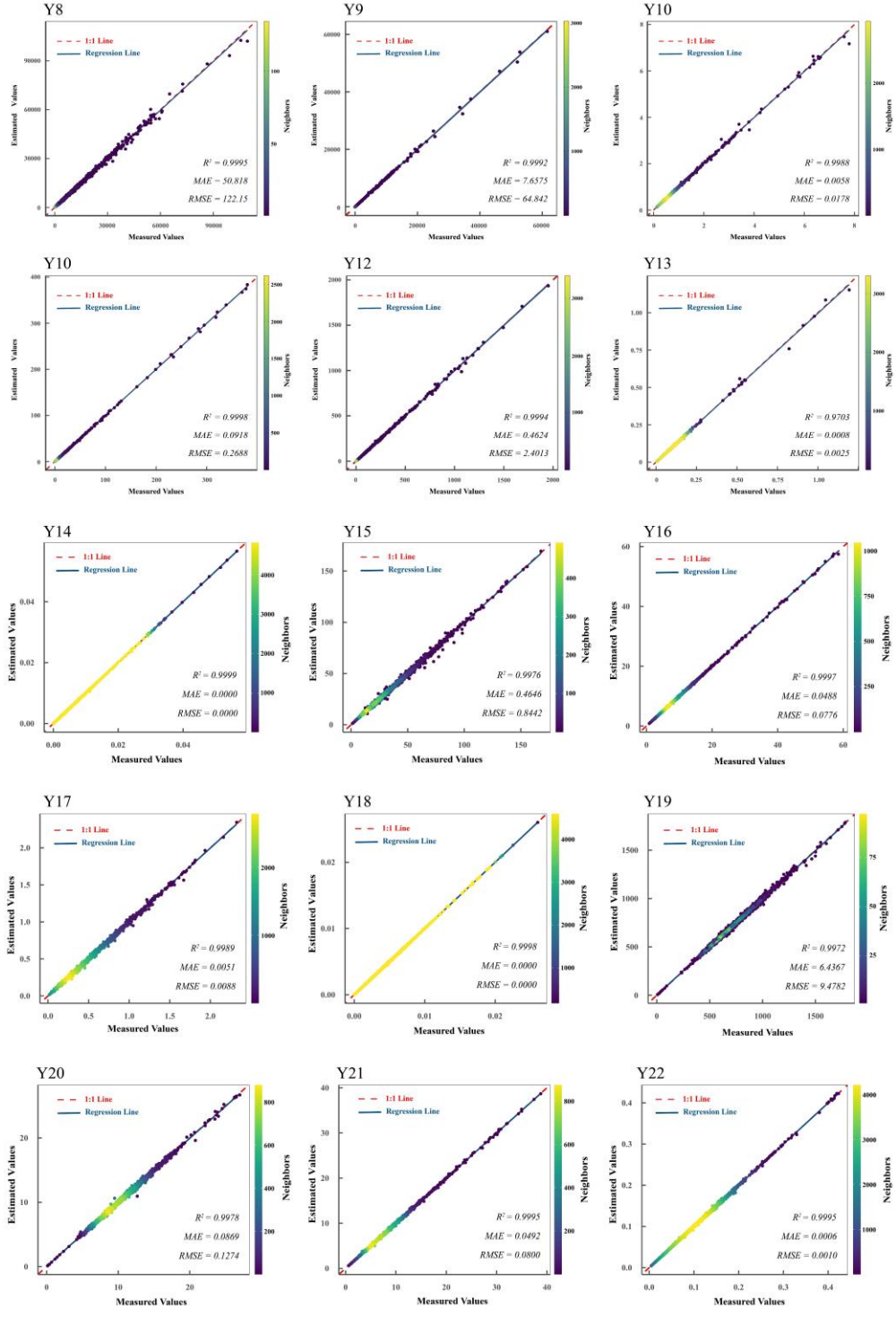


Fig. A1 Pearson correlation matrices among the 34 relative socioeconomic indicators (excluding Y7) and the 10 remote-sensing-derived auxiliary variables. *Indicatorname* in the figure denotes the corresponding socioeconomic relative indicator.

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

460 **Fig. A2 Model performance comparison between the multivariate linear regression (MLR) and Bayesian spatiotemporal interacting varying intercepts (BSTIVI) models for the 34 evaluated indicators, based on three Bayesian evaluation metrics: (a) DIC, (b) LS, and (c) WAIC.**

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026



https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

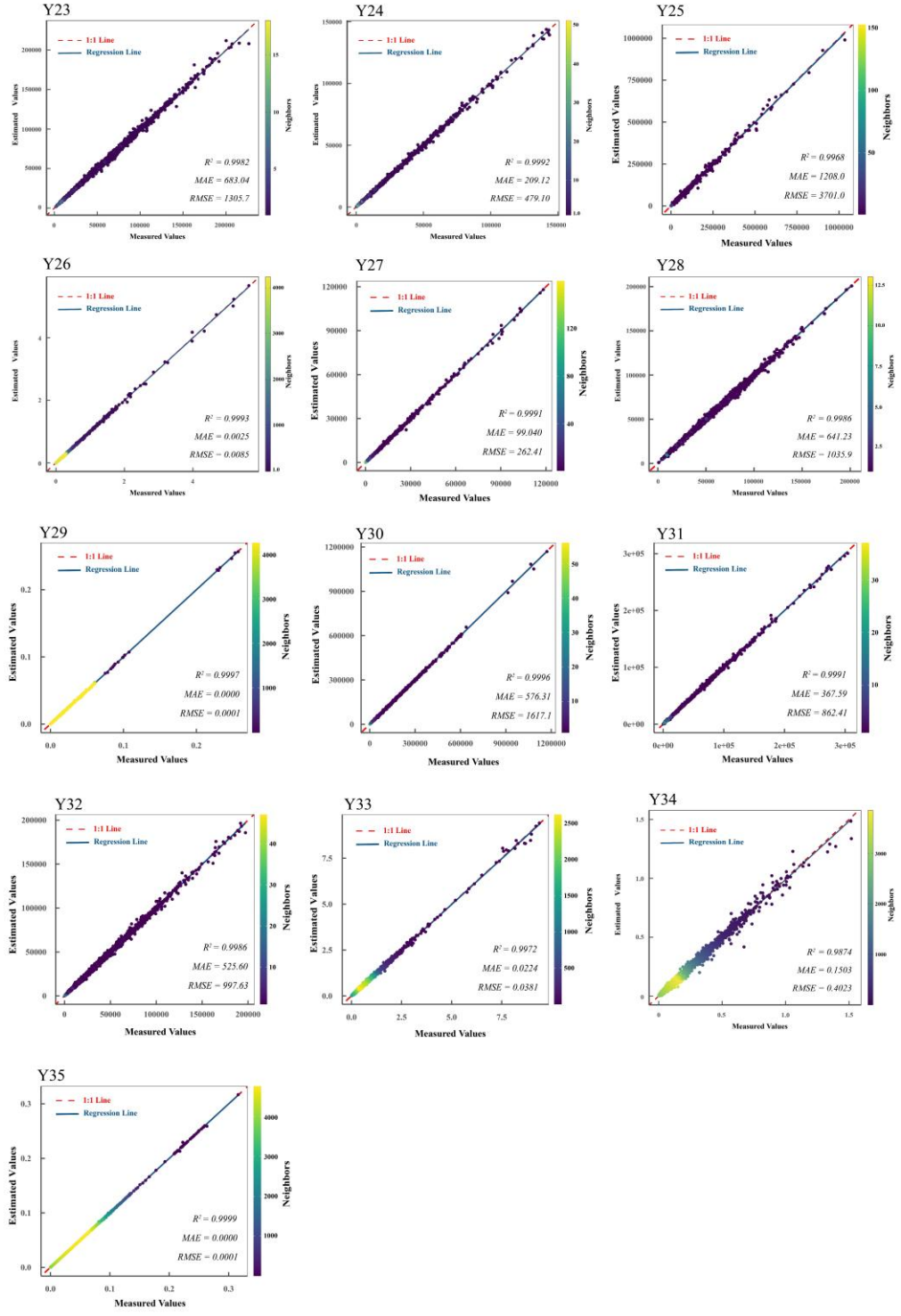
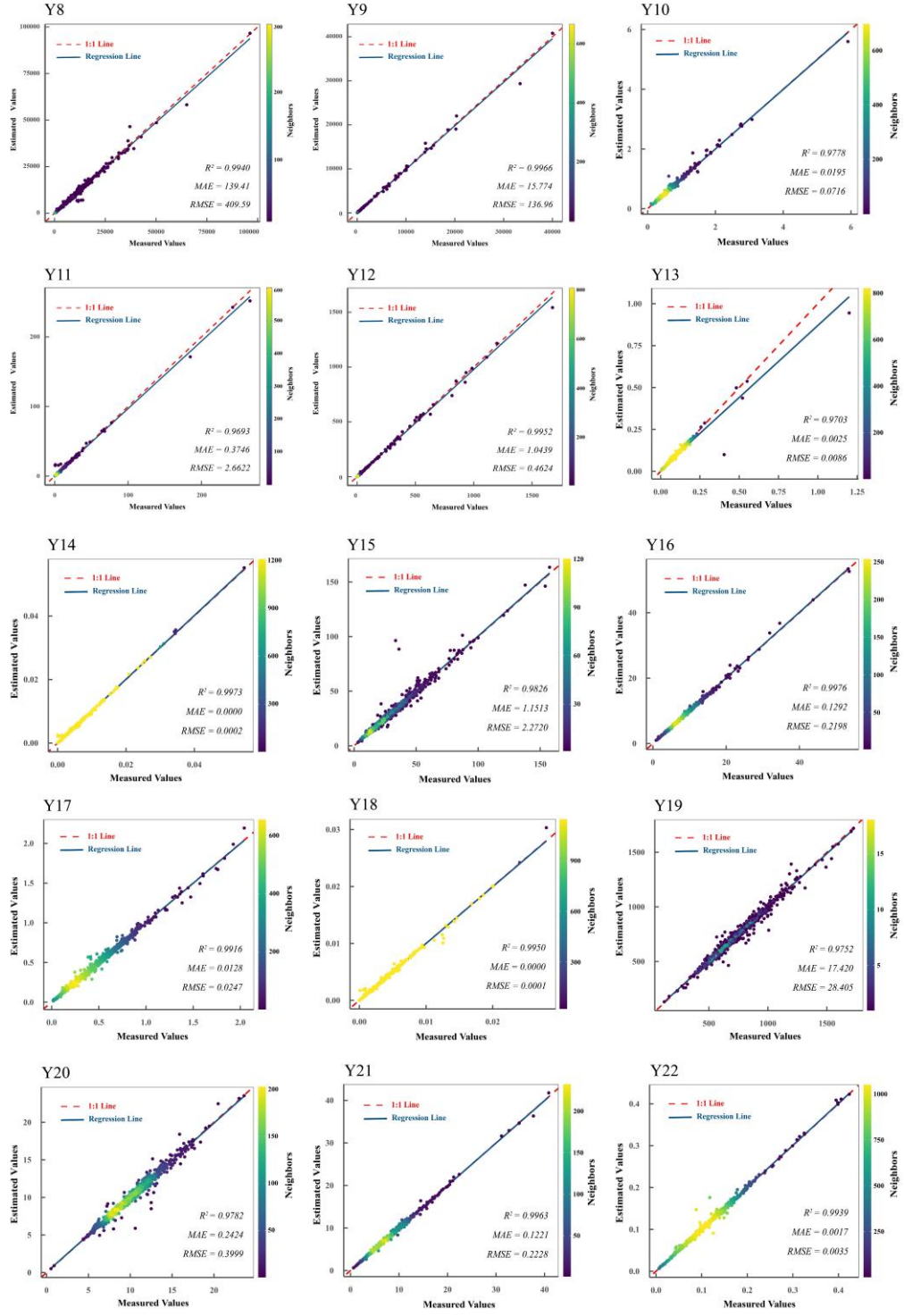


Fig. A3 Observed versus predicted scatter plots for the training set across socioeconomic indicators Y8-Y35, together with evaluation metrics (R^2, MAE, and RMSE).

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026



https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

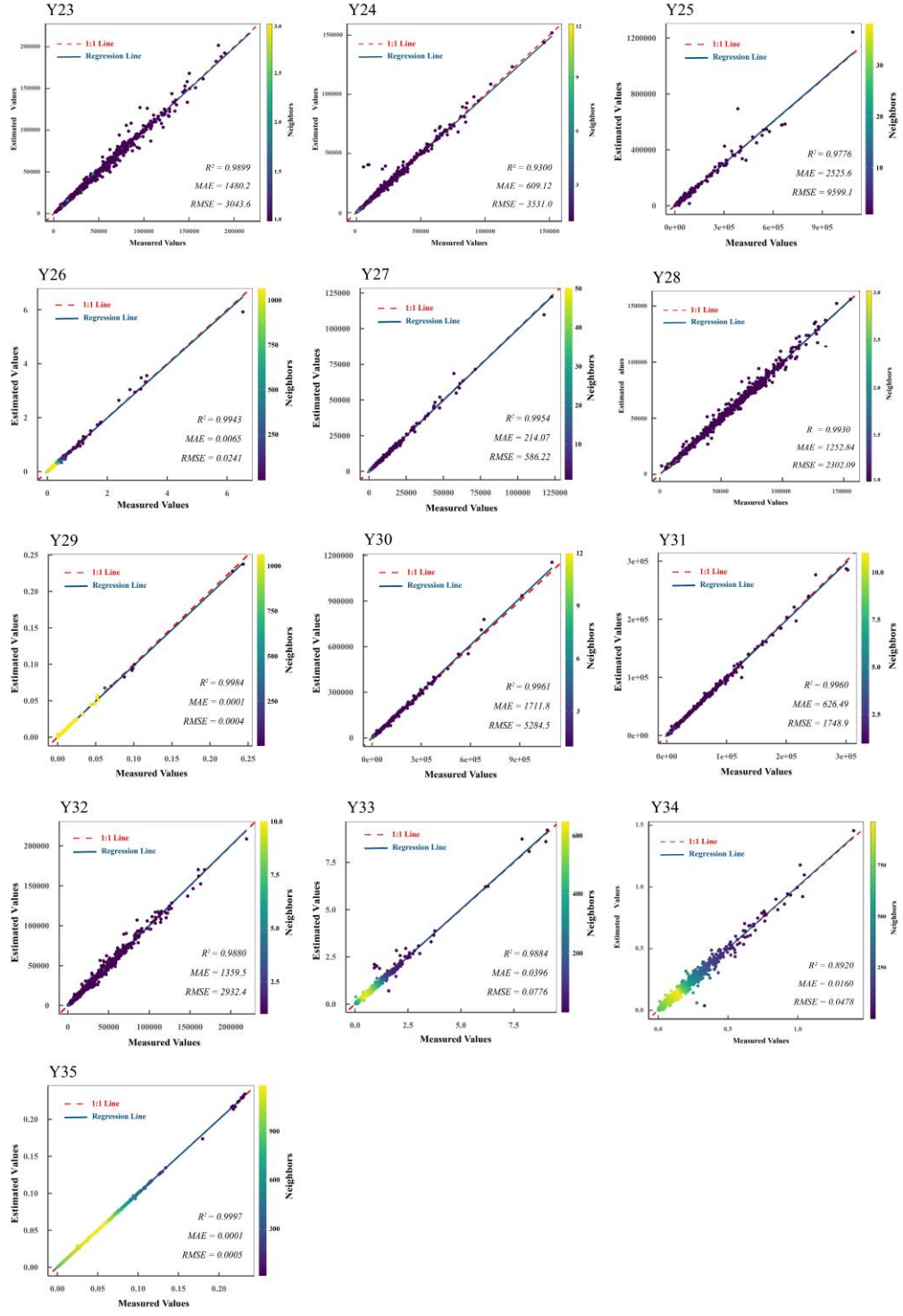


Fig. A4 Observed versus predicted scatter plots for the test set across socioeconomic indicators Y8-Y35, together with evaluation metrics (R^2, MAE, and RMSE).

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

6.2 Appendix B

Table B1: Entropy-based weight calculation results for the 25 selected relative socioeconomic indicators.

Goal level	Guideline level	Indicator level	Weights
Level of urban socio-economic development	Economic development	Y2	0.00481
		Y4	0.04608
		Y8	0.06951
		Y23	0.03339
		Y24	0.04673
		Y26	0.00012
		Y30	0.07759
		Y31	0.04365
		Y32	0.04561
	People's lives	Y27	0.06488
		Y28	0.02068
		Y33	0.03299
		Y34	0.04635
	Public services	Y5	0.04744
		Y11	0.09149
		Y12	0.16250
		Y13	0.01948
		Y14	0.03551
		Y16	0.01020
		Y17	0.02130
		Y18	0.02815
		Y21	0.00958
		Y22	0.00846
	Population	Y29	0.00016
		Y35	0.03334

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

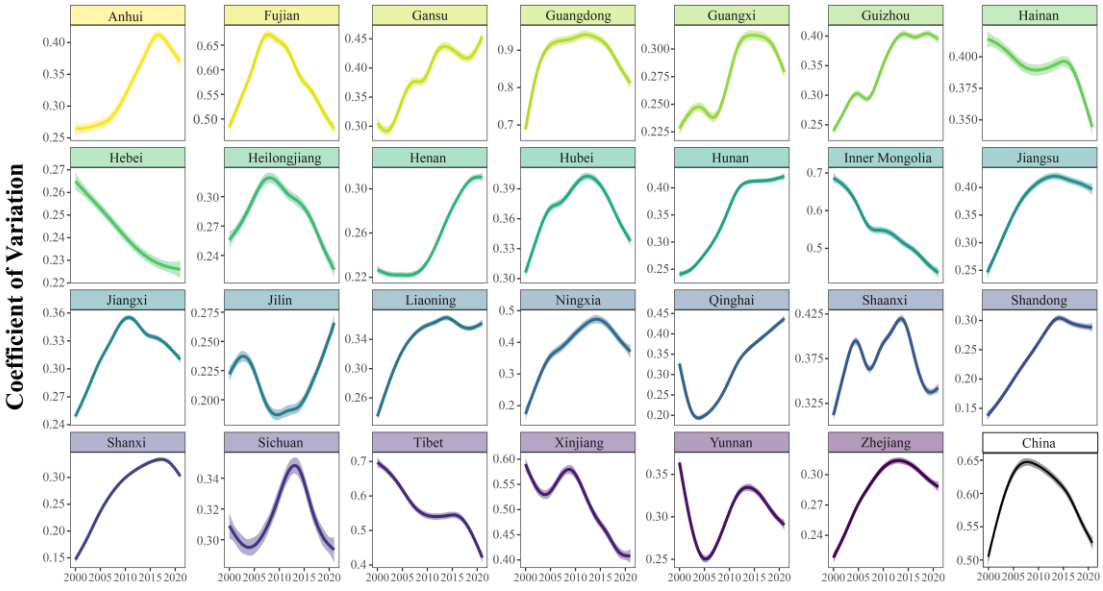


Fig. B1 Temporal trends in the coefficient of variation of the composite index at the national and provincial scales in China, 2000-2021.

Code and data availability

We have shared a dataset of 35 socioeconomic relative indicators and one composite index for Chinese cities covering 2000-2021 on the Zenodo platform (URL: https://zenodo.org/records/18217116) (Tang et al., 2026). The database will be regularly updated to the most recent years, maintained at an academic standard, and shared with all who need it.

We have made our self-developed BSTVC-R package freely available as open-source software at https://github.com/songbi123/BSTVC (Song and Tang, 2025). This package contains multiple model methods, but the BSTIVI method employed in the current study has not yet been integrated. However, the core method in this package, BSTVC, differs from BSTIVI only in minor parameter specifications, so the existing code can serve as a reference at this stage. Meanwhile, we will incorporate the method used in this study into future releases of the package in accordance with our update cycle.

Author contributions

Zhangying Tang: Writing (original draft preparation), Conceptualization, Supervision, Methodology, Project administration. **Xianteng Tang:** Writing (original draft preparation), Data curation, Formal analysis, Software, Validation, Visualization. **Lingfeng Liao:** Writing (review and editing), Data curation. **Guoqiang Yan:** Writing (review and editing), Data curation. **Zhenyan Wang:** Writing (review and editing), Data curation. **Yuju Wu:** Writing (review and editing), Validation. **Mingyu Xie:** Writing (review and editing), Visualization. **Yumeng Zhang:** Writing (review and editing), Software. **Chengwu Wang:**

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Writing (review and editing), Project administration. **Zhoufeng Wang:** Writing (review and editing), Resources. **Yangting Zeng:** Writing (review and editing), Visualization. **Chao Song:** Writing (review and editing), Conceptualization, Formal analysis, Funding acquisition, Methodology. **Jay Pan:** Writing (review and editing), Resources.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support

This work was supported by the General Program of the Sichuan Provincial Natural Science Foundation (2026NSFSC0215), the National Natural Science Foundation of China (42071379), the China Medical Board (25-614), the Sichuan Provincial Medical Research Youth Innovation Project (Q20250008), and the Open Fund of Sichuan Oil and Gas Development Research Center (2025SY006).

References

Alegana, V. A., Atkinson, P. M., Pezzulo, C., Sorichetta, A., Weiss, D., Bird, T., Erbach-Schoenberg, E., and Tatem, A. J.: Fine resolution mapping of population age-structures for health and development applications, Journal of The Royal Society Interface, 12, 10.1098/rsif.2015.0073, 2015.

Allen, C., Smith, M., Rabiee, M., and Dahmm, H.: A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals, Sustainability Science, 16, 1701-1716, 10.1007/s11625-021-00982-3, 2021.

Amaral, L. A. N., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J.: Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data, Plos One, 10, 10.1371/journal.pone.0107042, 2015.

Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, Statistics Surveys, 4, 10.1214/09-ss054, 2010.

de Silva, H. and Perera, A. S.: Missing data imputation using Evolutionary k- Nearest neighbor algorithm for gene

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

expression data, 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 10.1109/icter.2016.7829911, 2016.

de Souto, M. C. P., Jaskowiak, P. A., and Costa, I. G.: Impact of missing data imputation methods on gene expression clustering and classification, BMC Bioinformatics, 16, 10.1186/s12859-015-0494-3, 2015.

Du, H., Keller, B., Alacam, E., and Enders, C.: Comparing DIC and WAIC for multilevel models with missing data, Behavior Research Methods, 56, 2731-2750, 10.3758/s13428-023-02231-0, 2023.

Ferreira, L. Z., Blumenberg, C., Utazi, C. E., Nilsen, K., Hartwig, F. P., Tatem, A. J., and Barros, A. J. D.: Geospatial estimation of reproductive, maternal, newborn and child health indicators: a systematic review of methodological aspects of studies based on household surveys, International Journal of Health Geographics, 19, 10.1186/s12942-020-00239-9, 2020.

Fischer, M., Getis, A., and Eds: Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, Springer Berlin, Heidelberg, 10.1007/978-3-642-03647-7, 2010.

Gong, P., Zhu, S., Jiang, M., Zhu, B., and Yang, Y.: A new dataset of province- and prefecture-level human development index in China, Scientific Data, 12, 10.1038/s41597-025-05745-8, 2025.

Goodchild, M. F.: First Law of Geography, in: International Encyclopedia of Human Geography, edited by: Kitchin, R., and Thrift, N., Elsevier, Oxford, 179-182, 10.1016/B978-008044910-4.00438-7, 2009.

India State-Level Disease Burden Initiative Child Mortality, C.: Subnational mapping of under-5 and neonatal mortality trends in India: the Global Burden of Disease Study 2000-17, Lancet, 395, 1640-1658, 10.1016/S0140-6736(20)30471-2, 2020.

James, W. H. M., Tejedor-Garavito, N., Hanspal, S. E., Campbell-Sutton, A., Hornby, G. M., Pezzulo, C., Nilsen, K., Sorichetta, A., Ruktanonchai, C. W., Carioli, A., Kerr, D., Matthews, Z., and Tatem, A. J.: Gridded birth and pregnancy datasets for Africa, Latin America and the Caribbean, Scientific Data, 5, 10.1038/sdata.2018.90, 2018.

Kaplan, D.: On the Quantification of Model Uncertainty: A Bayesian Perspective, Psychometrika, 86, 215-238, 10.1007/s11336-021-09754-5, 2025.

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., and Tatem, A. J.: National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty, Proceedings of the National Academy of Sciences, 117, 24173-24179, 10.1073/pnas.1913050117, 2020.

Li, X., Zhou, Y., Zhao, M., and Zhao, X.: A harmonized global nighttime light dataset 1992-2018, Sci Data, 7, 168, 10.1038/s41597-020-0510-y, 2020.

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., and Tatem, A. J.: Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets, Big Earth Data, 3, 108-139, 10.1080/20964471.2019.1625151, 2019.

Lozano, R., Fullman, N., Abate, D., Abay, S. M., Abbafati, C., Abbasi, N., and al., e.: Measuring progress from 1990 to 2017 and projecting attainment to 2030 of the health-related Sustainable Development Goals for 195 countries and territories:

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

a systematic analysis for the Global Burden of Disease Study 2017, The Lancet, 392, 2091-2138, 10.1016/s0140-6736(18)32281-5, 2018.

Martin, A. J. F. and Conway, T. M.: Using the Gini Index to quantify urban green inequality: A systematic review and recommended reporting standards, Landscape and Urban Planning, 254, 10.1016/j.landurbplan.2024.105231, 2025.

McKeen, T., Bondarenko, M., Kerr, D., Esch, T., Marconcini, M., Palacios-Lopez, D., Zeidler, J., Valle, R. C., Juran, S., Tatem, A. J., and Sorichetta, A.: High-resolution gridded population datasets for Latin America and the Caribbean using official statistics, Scientific Data, 10, 10.1038/s41597-023-02305-w, 2023.

Murphy, E., Banerjee, A., Walsh P.P.: Partnerships and the Sustainable Development Goals, Springer Cham, Cham, Switzerland, 10.1007/978-3-031-07461-5, 2022.

Neal, S., Ruktanonchai, C. W., Chandra-Mouli, V., Harvey, C., Matthews, Z., Raina, N., and Tatem, A.: Using geospatial modelling to estimate the prevalence of adolescent first births in Nepal, BMJ Global Health, 4, 10.1136/bmjgh-2018-000763, 2019.

Oh, D., Cogen, R. M., Mullany, E. C., McLaughlin, S., Abiodun, O., Adamu, L. H., and al., e.: Mapping heterogeneity in family planning indicators in Burkina Faso, Kenya, and Nigeria, 2000–2020, BMC Medicine, 22, 10.1186/s12916-023-03214-w, 2024.

Palacios-Lopez, D., Bachofer, F., Esch, T., Heldens, W., Hirner, A., Marconcini, M., Sorichetta, A., Zeidler, J., Kuenzer, C., Dech, S., Tatem, A. J., and Reinartz, P.: New Perspectives for Mapping Global Population Distribution Using World Settlement Footprint Products, Sustainability, 11, 10.3390/su11216056, 2019.

Pamucar, D., Wu, R. M. X., Zhang, Z., Yan, W., Fan, J., Gou, J., Liu, B., Gide, E., Soar, J., Shen, B., Fazal-e-Hasan, S., Liu, Z., Zhang, P., Wang, P., Cui, X., Peng, Z., and Wang, Y.: A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement, Plos One, 17, 10.1371/journal.pone.0262261, 2022.

Pati, S. K. and Das, A. K.: Missing value estimation for microarray data through cluster analysis, Knowledge and Information Systems, 52, 709-750, 10.1007/s10115-017-1025-5, 2017.

Pezzulo, C., Tejedor-Garavito, N., Chan, H. M. T., Dreoni, I., Kerr, D., Ghosh, S., Bonnie, A., Bondarenko, M., Salasibew, M., and Tatem, A. J.: A subnational reproductive, maternal, newborn, child, and adolescent health and development atlas of India, Scientific Data, 10, 10.1038/s41597-023-01961-2, 2023.

Purwar, A. and Singh, S. K.: Hybrid prediction model with missing value imputation for medical data, Expert Systems with Applications, 42, 5621-5631, 10.1016/j.eswa.2015.02.050, 2015.

Seu, K., Kang, M.S., Lee, H.M.: An Intelligent Missing Data Imputation Techniques: A Review, International Journal on Informatics Visualization, 6, 278–283, 10.30630/joiv.6.1-2.935, 2022.

Shao, J.: Linear Model Selection by Cross-Validation, Journal of the American Statistical Association, 88, 486-494, 10.2307/2290328, 1993.

Song, C., Shi, X., and Wang, J.: Spatiotemporally Varying Coefficients (STVC) model: a Bayesian local regression to detect spatial and temporal nonstationarity in variables relationships, Annals of GIS, 26, 277-291,

10.1080/19475683.2020.1782469, 2020.

Song, C., Yang, X., Shi, X., Bo, Y., and Wang, J.: Estimating missing values in China's official socioeconomic statistics using progressive spatiotemporal Bayesian hierarchical modeling, Scientific Reports, 8, 10.1038/s41598-018-28322-z, 2018.

Song, C., Yin, H., Shi, X., Xie, M., Yang, S., Zhou, J., Wang, X., Tang, Z., Yang, Y., and Pan, J.: Spatiotemporal disparities in regional public risk perception of COVID-19 using Bayesian Spatiotemporally Varying Coefficients (STVC) series models across Chinese cities, Int J Disaster Risk Reduct, 77, 103078, 10.1016/j.ijdrr.2022.103078, 2022.

Song, C., Fang, L., Xie, M., Tang, Z., Zhang, Y., Tian, F., Wang, X., Lin, X., Liu, Q., Xu, S., and Pan, J.: Revealing spatiotemporal inequalities, hotspots, and determinants in healthcare resource distribution: insights from hospital beds panel data in 2308 Chinese counties, BMC Public Health, 24, 423, 10.1186/s12889-024-17950-y, 2024.

Song, C. and Tang, X.: User's Guide for the BSTVC R Package, Github [code], https://github.com/songbi123/BSTVC, 2025.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J.: High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020, Scientific Data, 2, 10.1038/sdata.2015.45, 2015.

Tatem, A. J., Campbell, J., Guerra-Arias, M., Bernis, L.D., Moran, A., Matthews, Z.: Mapping for maternal and newborn health the distributions of women of childbearing age, pregnancies and births, International Journal of Health Geographics, 13, 2, 10.1186/1476-072X-13-2, 2014.

Tang, Z., Tang, X., Liao, L., Yan, G., Wang, Z., Wu, Y., Xie, M., Zhang, Y., Wang, C., Wang, Z., Zeng, Y., Song, C., and Pan, J.: City-Level Socioeconomic Indicators and Composite Development Index for China, 2000-2021, Zenodo [data set], https://doi.org/10.5281/zenodo.18217116, 2026.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J.: High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries, Vaccine, 36, 1583-1591, 10.1016/j.vaccine.2018.02.020, 2018.

Wan, Q., Tang, Z., Pan, J., Xie, M., Wang, S., Yin, H., Li, J., Liu, X., Yang, Y., and Song, C.: Spatiotemporal heterogeneity in associations of national population ageing with socioeconomic and environmental factors at the global scale, Journal of Cleaner Production, 373, 10.1016/j.jclepro.2022.133781, 2022.

Wang, Y., Li, X., Zhou, M., Luo, S., Liang, J., Liddell, C. A., Coates, M. M., Gao, Y., Wang, L., He, C., Kang, C., Liu, S., Dai, L., Schumacher, A. E., Fraser, M. S., Wolock, T. M., Pain, A., Levitz, C. E., Singh, L., Coggeshall, M., Lind, M., Li, Y., Li, Q., Deng, K., Mu, Y., Deng, C., Yi, L., Liu, Z., Ma, X., Li, H., Mu, D., Zhu, J., Murray, C. J., and Wang, H.: Under-5 mortality in 2851 Chinese counties, 1996-2012: a subnational assessment of achieving MDG 4 goals in China, Lancet, 387, 273-283, 10.1016/S0140-6736(15)00554-1, 2016.

Wang, Z., Dong, L., Xing, X., Liu, Z., and Zhou, Y.: Disparity in hospital beds' allocation at the county level in China: an analysis based on a Health Resource Density Index (HRDI) model, BMC Health Services Research, 23, 10.1186/s12913-023-10266-4, 2023.

Wei, J., Li, Z., Guo, J., Sun, L., Huang, W., Xue, W., Fan, T., and Cribb, M.: Satellite-Derived 1-km-Resolution PM1

https://doi.org/10.5194/essd-2026-306
Preprint. Discussion started: 8 May 2026

Concentrations from 2014 to 2018 across China, Environmental Science & Technology, 53, 13265-13274, 10.1021/acs.est.9b03258, 2019.

Wei, J., Li, Z., Xue, W., Sun, L., Fan, T., Liu, L., Su, T., and Cribb, M.: The ChinaHighPM10 dataset: generation, validation, and spatiotemporal variations from 2015 to 2019 across China, Environment International, 146, 10.1016/j.envint.2020.106290, 2021.

Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., and Ning, G.: Adjusted weight voting algorithm for random forests in handling missing values, Pattern Recognition, 69, 52-60, 10.1016/j.patcog.2017.04.005, 2017.

Ye, P., Ye, Z., Xia, J., Zhong, L., Zhang, M., Lv, L., Tu, W., Yue, Y., and Li, Q.: National-scale 1-km maps of hospital travel time and hospital accessibility in China, Scientific Data, 11, 10.1038/s41597-024-03981-y, 2024.

Zahmatkesh, S. and Zech, P.: Spatio-Temporal Missing Data Imputation: A Systematic Literature Review with a Focus on Statistical and Machine Learning-Based Approaches, ACM Computing Surveys, 10.1145/3797903, 2026.

Zhang, H., Dong, G., Li, B., Xie, Z., Miao, C., Yang, F., Gao, Y., Meng, X., Yang, D., Liu, Y., Zhang, H., Wu, L., Shi, F., Chen, Y., Wu, W., Laszkiewicz, E., Liang, Y., Lu, B., Yao, J., and Li, X.: Developing an annual global Sub-National scale economic data from 1992 to 2021 using nighttime lights and deep learning, International Journal of Applied Earth Observation and Geoinformation, 133, 10.1016/j.jag.2024.104086, 2024.

Zhao, J., Yang, Y., and Ogasawara, K.: Measuring the Inequalities in the Distribution of Public Healthcare Resources by the HRDI (Health Resources Density Index): Data Analysis from 2010 to 2019, Healthcare, 10, 10.3390/healthcare10081401, 2022.

Zhao, N., Liu, Y., Cao, G., Samson, E. L., and Zhang, J.: Forecasting China's GDP at the pixel level using nighttime lights time series and population images, GIScience & Remote Sensing, 54, 407-425, 10.1080/15481603.2016.1276705, 2017.

Zhou, K., Hu, R., Lu, X., Yang, Z., and Gao, Y.: Multi-resolution dataset of electricity consumption in Chinese cities, Scientific Data, 12, 10.1038/s41597-025-06256-2, 2025.