International Journal of Geo-Information

MDPI

Article

Multiple Global Population Datasets: Differences and Spatial Distribution Characteristics

Ruxia Chen [1,2], Huimin Yan [1,2,*], Fang Liu [1], Wenpeng Du [1,2] and Yanzhao Yang [1,2]

[1] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; chenrx.18s@igsnrr.ac.cn (R.C.); liuf.08s@igsnrr.ac.cn (F.L.); duwp.18b@igsnrr.ac.cn (W.D.); yangyz@igsnrr.ac.cn (Y.Y.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: yanhm@igsnrr.ac.cn; Tel.: +86-10-648-89467

Received: 14 September 2020; Accepted: 22 October 2020; Published: 27 October 2020

Abstract: Spatial data of regional populations are indispensable in studying the impact of human activities on resource utilization and the ecological environment. Because the differences between datasets and their spatial distribution are still unclear, this has become a puzzle in data selection and application. This study is based on four mainstream spatialized population datasets: the History Database of the Global Environment version 3.2.000 (HYDE), Gridded Population of the World version 4 (GPWv4), Global Human Settlement Layer (GHSL), and WorldPop. In view of possible influences of geographical factors, this study analyzes the differences in accuracy of population estimation by computing relative errors and population spatial distribution consistency in different regions by comparing datasets pixel by pixel. The results demonstrate the following: (1) Source data, spatialization methods, and case area features affect the precision of datasets. As the main data source is statistical data and the spatialization method maintains the population in the administrative region, the populations of GPWv4 and GHSL are closest to the statistical data value. (2) The application of remote sensing, mobile communication, and other geospatial data makes the datasets more accurate in the United Kingdom, with rich information, and the absolute value of relative errors is less than 4%. In the Tibet Autonomous Region of China, where data are hard to obtain, the four datasets have larger relative errors. However, the area where the four datasets are completely consistent is as high as 84.73% in Tibet, while in the UK it is only 66.76%. (3) The areas where the spatial patterns of the four datasets are completely consistent are mainly distributed in areas with low population density, or with developed urbanization and concentrated population distribution. Areas where the datasets have poor consistency are mainly distributed in medium population density areas with high urbanization levels. Therefore, in such areas, a more careful assessment should be made during the data application process, and more emphasis should be placed on improving data accuracy when using spatialization methods.

Keywords: spatial population dataset; spatial distribution of population; difference; consistency

1. Introduction

Population growth has placed certain pressures on society, resources, and the ecological environment, and even affected ecosystem functions [1,2]. The critical role of population data in the study of social economy, resource utilization, and ecosystem change has been widely recognized [3]. In particular, population density data can be broadly applied in quantifying the intensity of human activities, depicting the spatial patterns of eco-environmental quality, simulating the spatial distribution of pollutant emissions, and evaluating ecological problems brought about by urbanization [4–7], as well as in other ecological research. With the development of remote sensing technology, population data

based on administrative units has become a bottleneck restricting the integrated analysis of social and natural systems [8,9]. The spatialization of population data is based on distribution rules redistributing the data from the administrative unit scale to a specified grid size, in order to estimate and simulate real population spatial distribution. The establishment of such distribution rules often takes demographic data as input, and social and economic data, administrative divisions, transportation, terrain and other elements as references [10].

At present, spatial population datasets shared at the global and regional scales include Gridded Population of the World (GPW) [11], Global Human Settlement Layer (GHSL) [12], History Database of the Global Environment (HYDE) [13–15], WorldPop [16,17], Global Urban Footprint (GUF), High-Resolution Settlement Layer (HRSL), and so on. These data have been widely used in disaster assessment and risk management [18–22], land use change modeling [23–26], public health services [27–29], and ecological environment change [30–34] and socioeconomic analysis [35] as important references for developing new population spatial datasets [36–38]. Goldewijk et al. used HYDE [13] to estimate land use change in historical periods. Gleeson et al. used GPWv3 data from 2000 [22] to study the sustainability of groundwater, finding that about 1.7 billion people lived in areas affected by groundwater pressure, and more than half of the affected population lived in China and India. Based on the malaria stability index and WorldPop data [19], Kibret et al. measured the infection rate of *Plasmodium falciparum* in areas of relevant reservoirs and found that 723 of 1268 dams were located in diseased areas (about 15 million people). Melchiorri et al. used GHSL [24] to study the evolution of global urbanization from 1990 to 2015 and the current situation, and clarified the key role of urban areas in the development and mode of global urban development.

The datasets construct spatial population data by various methods, which naturally leads to different results in the same research with other datasets. For example, when estimating how much of the population suffers from flood risk in Mexico, Haiti, and 18 other countries, the estimations using WorldPop and LandScan were 20.79% and 32.67%, respectively, higher than that obtained by HRSL [17]. In order to select the appropriate data, research about the accuracy of comparisons or validations of datasets is gradually carried out in various case areas. The research results of Bai [39] in China showed that WorldPop had the highest and GPW the lowest estimation accuracy, but the estimation accuracy of GPW in plain and basin areas was slightly higher than in other regions. The results of a validation study [40] on the GHSL datasets in urban and rural areas of the United States showed that the data were very accurate in areas with a high development level, while in rural areas, the accuracy may be low due to sparse built-up areas and a lack of reference data. Ye [37], Yang [41] and Sliuzas [42] reached different conclusions on the accuracy of datasets—Ye thought that the WorldPop dataset permitted low estimation of urban populations and high estimation of rural populations, and Yang found that there were more errors in the WorldPop data in areas with high or very low population density. The research results of Sliuzas showed that GHSL could only describe the main forms of cities, but there were quite a lot of misclassifications at the pixel level, so the accuracy was not high.

The selection and application of global datasets is a difficult for all kinds of data, and the reliability requires sufficient verification. However, validating spatial population data is far more difficult than validating a global land cover dataset, which can be verified by high-resolution remote sensing data, and a global ecosystem productivity dataset, which can be verified by long-term data collected from located observations. Therefore, mastering the characteristics of spatial and temporal layout, and the advantages of each set of data to select that is appropriate to the use of data in the research process, will better reduce the uncertainty of research. Therefore, we chose four datasets, HYDE, GPWv4, GHSL and WorldPop, which are widely used and have different spatiotemporal resolution, to conduct a comparative study in terms of the reliability of the population and the difference of spatial distribution. In order to reflect the difficulty of collecting demographic data and the influence of population density distribution on spatial data, we selected the United Kingdom, Argentina, Sri Lanka and Tibet Autonomous Region of China as the case areas. We analyzed the differences between the four datasets and the reasons for these differences from the aspects of the data production process,

estimation deviation [43], consistency of spatial population distribution [44], and population density level distribution at the administrative unit and pixel scale, so as to provide a reference for the selection of population datasets in socioeconomic or ecological environment research [41,45].

2. Materials and Methods

2.1. Case Area Selection

In order to evaluate the performance of the population datasets in areas with different topographical and urban–rural distribution characteristics, and to discover the relationship between the accuracy of datasets with geographical factors, we selected 4 case areas with different characteristics: the United Kingdom, with a high population density of 274.7 persons/km^2 and an urbanization rate of 83.4%; Argentina, with a high proportion of urban population of 91.9%; Sri Lanka, with flat terrain below 200 m; and Tibet Autonomous Region of China, with an altitude higher than 4000 m and sparse population less than 3 persons/km^2 (Table 1). The characteristics of the population distribution in the UK are as follows: overall population density is high and the urbanization rate is as high as 83.4%, forming a pattern of outward divergence with the population concentration centers in Greater London, Manchester, Birmingham, and other counties within the jurisdiction of England, and Glasgow and Edinburgh within the jurisdiction of Scotland. The overall population density of Argentina is 16.3 persons/km^2, but much of the rural population has poured into the cities due to the backward progress of economic development, resulting in a large population concentration and high proportions of urban population in Buenos Aires, Cordoba, Mendoza, and other large cities in the north, while the population density of small cities in the south is mostly less than 5 persons/km^2. Sri Lanka is relatively flat, with altitude less than 200 m, and most of the terrain is plain; the overall population density is as high as 345.6 persons/km^2. The characteristics of population distribution are as follows: population density in the west is greater than 500 persons/km^2, and in the east is mostly less than 100 persons/km^2; it decreases in all directions, with Colombo and Kandy as the areas with the highest population concentration; there is also high population distribution density in the ports, such as Jaffna in the north, with more than 2000 persons/km^2. The average altitude of Tibet Autonomous Region of China is more than 4000 m [45], and the overall population density is less than 3 persons/km^2. The population is predominantly distributed in Lhasa, Xigaze, and several agricultural counties, while high-altitude areas such as Ali and the north of Naqu are very sparsely populated.

Table 1. Characteristics of each case area.

Case Area	Characteristics	Proportion of Urban Population in 2015	Factors Assessed
United Kingdom	High population density, high proportion of urbanized population	82.63% ①	The impact of urbanization rate
Argentina	Low overall population density, population concentrated in big cities, a big gap between urban and rural areas	91.50% ①	The impact of population concentration
Sri Lanka	Flat terrain, high population density	18.26% ①	The influence of terrain flatness
Tibet Autonomous Region of China	High altitude, large-scale sparse population, difficulty of data acquisition	27.74% ②	The influence of high-altitude areas lacking data

Data sources: ① From the world bank website, https://data.worldbank.org.cn/. ② From Tibet Statistical Yearbook (2016).

2.2. Data Source

The main data used in this paper include four spatial population datasets from 2015, administrative division data, and demographic data. The four spatial population datasets are: (1) popd of History Database of the Global Environment (HYDE), version 3.2.000 [13–15], produced by PBL Netherlands Environmental Assessment Agency; (2) Population Density of Gridded Population of the World, version 4 (GPWv4) [11], produced by Center for International Earth Science Information Network (CIESIN), Columbia University; (3) GHS-POP of Global Human Settlement Layer (GHSL) [12], produced by the European Commission; and (4) WorldPop Population Counts [16] using unconstrained top-down methods completed by multiple organizations and institutions. The data of administrative divisions are derived from the Global Administrative Areas (GADM) version 3.6 produced by the Center for Spatial Sciences at the University of California and administrative divisions (Due to the lack of WorldPop data in the southeast part of the Tibet Autonomous Region, data of Longzi County, Cuona County, and Linzhi city are excluded.) of the Tibet Autonomous Region provided by the Qinghai Tibet Plateau scientific data center (http://www.tpedatabase.cn). The definition of population density according to HYDE and GPWv4 data is the number of people per square kilometer, and according to GHSL and WorldPop it is the number of people in each grid. In order to reduce the impact of data processing on spatial statistics of datasets and facilitate comparison, this paper unifies the measurement unit of the four datasets as population per square kilometer without changing their resolution. The population statistics came from the website of the World Bank and the Statistical Yearbook of the Tibet Autonomous Region of China from 2015.

There are nearly 20 years between the development of the earliest GPWv4 dataset and the latest WorldPop data entry. The data sources experienced a transformation from single demographic data to the integration of digital elevation map (DEM), land cover data and transportation network, then to remote sensing data, mobile communication and other new data sources (Table 2). The GPWv4 dataset, with a spatial resolution of 1km, is based on the 2010 official census and estimated population estimated data, supplemented by administrative boundaries and the United Nation's World Population Prospects, 2015 Revision. The data source of HYDE 3.2 is the United Nation's World Population Prospects and historical estimations from the literature [46–48], supplemented with data from the sub-national population statistics of Populstat and other sources. HYDE constructed a continuous population time series with spatial resolution of 10km for each country's province or state [11]. Using remote sensing satellite data and volunteered geographic information, GHSL generates fine built-up areas and decomposes the GPWv4 produced by CIESIN to generate population distribution maps with higher spatial resolution (250m) and more detailed spatial expression. There are many input data for WorldPop, including elevation, slope, land cover, infrastructure, satellite data, and mobile phone communication data, in addition to the 2010 national census and official population estimation data. At present, year-on-year time series data with a spatial resolution of 100 m from 2000 to 2020 have been developed.

In terms of the production method, GPWv4, based on the area weighting method, is the only dataset of the 4 that is not spatialized by modeling. The production mode is simple and ensures the accuracy of the total population within the administrative unit. However, the disadvantage is that it is based on the assumption that humans are evenly distributed in space. The HYDE3.2 dataset generates a combined weight layer based on soil suitability, road accessibility, distance from water body, night light and other indicators to spatialize population data. This model is applicable globally, but does not take into account additional uncertainties in the region. GHSL uses remote sensing satellite data and volunteered geographic information to generate built-up areas with a spatial resolution of 38 m, and according to the proportion of built-up area in each grid, decomposes GPW again based on a linear regression method. The modeling method is simple, considering that the population is mainly distributed in built-up areas, but ignores administrative boundaries. With the development and application of machine learning and other algorithms, WorldPop uses a random forest model to quantify the relationship between model factors such as land cover, satellite data, mobile phone communication

and micro-census so as to generate a weight layer and reallocate census data. Among the four datasets, the geographic information data source of WorldPop is more sufficient, and its random forest model is superior to classification and variable importance ranking [49], which presents the development direction of the spatialization method in the future.

Table 2. Basic information of spatial population datasets.

Dataset	GPWv4	HYDE	GHSL	WorldPop
Resolution	1 km	10 km	250 m, 1 km	100 m
Period	2000, 2005, 2010, 2015, 2020	10000BC–2017AD	1975, 1990, 2000, 2015	2000–2020
Unit coverage	People per km^2 global	People per km^2 global	People per grid global	People per grid global
Source data and auxiliary data	2010 official census, administrative boundaries, World Population Prospects, 2015 Revision.	World Population Prospects, historical estimation in the literature [49–51], sub-national population statistics of Populstat and other sources [16]	GPWv4, remote sensing satellite data, volunteer geographic information	2010 official census, elevation, slope, land cover, infrastructure, satellite data, mobile phone communication data
Spatialization method	According to the proportion of land area relative to the overall area of each pixel, population is distributed proportionally (evenly distributed)	According to the weight layer based on soil suitability, road accessibility, slope, distance from water body, night light and other indicators to spatialize population	According to the proportion of built-up area based on remote sensing image in each grid, the residential population data are allocated to each grid	According to the weight layer generated by the prediction model of random forest to redistribute census data
Year of production	1995	2001	2011	October, 2013
Limitations	The assumption that populations are evenly distributed in space is unreasonable.	The model is globally applicable, without considering other uncertainties in regions	The administrative boundary is ignored	
Production organization or unit	CIESIN, Columbia University	Netherlands Environmental Assessment Agency	European Commission	University of Southampton and other organizations

2.3. *Analysis Method of Spatial Distribution Consistency*

The spatial distribution consistency analysis measured the consistency of population spatial distribution in the 4 datasets by comparing them pixel by pixel. The process of consistency analysis is as follows: (1) The units of the 4 datasets are converted and unified into people per km^2. (2) According to the different population density characteristics of each case area (Table 1), population density is reclassified to 9 levels based on the natural breakpoint method [50,51] (Table 3). (3) Raster calculation is performed on the 4 datasets after reclassification to obtain the grid data reflecting data consistency. The grid data include instances where 4 datasets are consistent, 3 datasets are consistent, 2 datasets are consistent, and each dataset is inconsistent with the others, respectively defined as completely consistent, highly consistent, lowly consistent, and completely inconsistent [50]. (4) Datasets are compared pairwise and analyzed to determine whether they are consistent and the proportion of consistency. (5) Statistical analysis is conducted by zonal statistical tools in Arcgis, which refers to the distribution of population density levels in 2 types of consistent regions (Figure 1).

Table 3. Population density classification based on natural breakpoint method.

Case Area	Population Density at Different Levels (People per km^2)								
	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**	**Level 7**	**Level 8**	**Level 9**
United Kingdom	≤1	2–25	26–50	51–100	101–200	201–400	401–500	501–1000	>1000
Argentina	≤1	2–5	6–10	11–25	26–50	51–100	101–200	201–500	>500
Sri Lanka	≤10	11–50	51–100	101–200	201–500	501–1000	1001–2000	2001–3000	>3000
Tibet Autonomous Region of China	≤1	2–5	6–10	11–25	26–50	51–100	101–200	201–500	>500

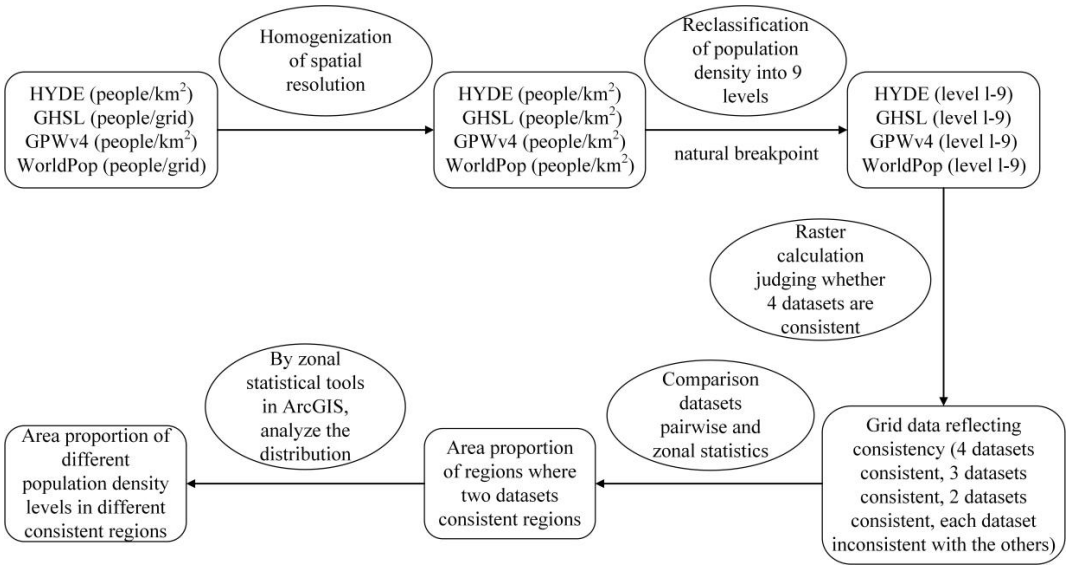

Figure 1. Process of consistency analysis.

3. Results

3.1. Accuracy of Population Estimation

Taking the World Bank and Statistical Yearbook data of 2015 as the reference for the total population, this paper compares the accuracy of the population estimated by four spatial datasets (Table 4). The total populations of GPWv4 and GHSL are the closest to the statistical data, and the absolute value of relative error is within 3%. The reason for this may be that GPWv4 is based on the 2010 census data and allocates the population within each administrative unit so as to keep the population in each unit unchanged, while GHSL is a refined spatial allocation based on GPWv4. The relative error of WorldPop is negative, and the absolute value of relative error is the largest in Argentina, Sri Lanka and Tibet, which may be explained by a small amount of regional data in each case area. Although the data sources of WorldPop and GPWv4 are based on the 2010 census data, there is a great gap between them in total population. The relative error of WorldPop in Argentina, Sri Lanka and the Tibet Autonomous Region of China is as high as 20%, which shows that differences in population density and distribution patterns simulated by different spatialization methods make the total amount unequal. HYDE shows that the accuracy varies in different regions. In the UK, the relative error is only −3.71%, and the absolute value in Argentina and Sri Lanka is less than 10%, while the relative error in the Tibet Autonomous Region is −15.03%.

Table 4. Total population and estimation deviation of datasets.

		WorldBank/Statistical Yearbook	HYDE	GPWv4	GHSL	WorldPop
United Kingdom	Total population (unit: 10,000 persons)	6512.89	6271.11	6503.65	6455.65	6404.95
	relative error (%)		−3.71%	−0.14%	−0.88%	−1.66%
Argentina	Total population (unit: 10,000 persons)	4313.20	4095.07	4270.79	4341.77	3337.86
	relative error (%)		−5.06%	−0.98%	0.66%	−22.61%
Sri Lanka	Total population (unit: 10,000 persons)	2097.00	2295.84	2049.89	2046.46	1697.35
	relative error (%)		9.48%	−2.25%	−2.41%	−19.06%
Tibet Autonomous Region of China	Total population (unit: 10,000 persons)	296.07	251.57	302.11	292.66	256.40
	relative error (%)		−15.03%	2.04%	−1.15	−20.62%

The application of remote sensing, mobile phone communication, and other geospatial data will make the data in areas with abundant information more accurate. In the UK, the absolute value of the relative error of the four datasets is lower than 4%, and that of GPWv4 and GHSL is less than 1%. In Argentina, Sri Lanka and the Tibet Autonomous Region of China, the results of GPWv4 and GHSL are similar. The relative error of GHSL in Argentina, Sri Lanka and the Tibet Autonomous Region of China is 0.66%, −2.41% and −1.15%, respectively; however, the absolute deviation between HYDE and WorldPop is between 5% and 25%.

Compared with the other three regions, the relative error of the Tibet Autonomous Region in China is generally larger, especially with HYDE using the literature's historical data and WorldPop using multi-source geographic information data such as communication data. The accuracy of the estimation of the Tibet Autonomous Region is far lower than that of other regions, with a deviation of about 20%. There may be two reasons for this: in terms of massive sparsely populated areas at high altitude, the scale of population statistics is not precise enough [8], and/or it is difficult to obtain new auxiliary data such as household survey and mobile phone communication data, which makes the error of the spatial population dataset larger.

3.2. Consistency Analysis of Population Spatial Distribution

Contrary to the population accuracy, due to the lack of geographical information data, the spatial distribution characteristics of the four datasets are basically the same and the datasets are the most consistent in the Tibet Autonomous Region of China (Figure 2). Thus, the proportion of completely/highly consistent regions is as high as 97.01%. In the UK, the proportion of completely/highly consistent regions is the lowest, at only 66.75% (Table 5). The proportion of completely/highly consistent regions is slightly higher in Argentina than in Sri Lanka, at 82.06% and 81.80%, respectively (Table 5), although Sri Lanka's urbanization rate is lower (the proportion of urban population in 2015 was 18.3%), and Argentina's urbanization rate is high (91.5% in 2015). However, Sri Lanka's population differentiation is more complicated, and the overall population density is higher compared to Argentina, which indicates that the spatial distribution pattern of datasets is quite different in areas with high population density and complex variation.

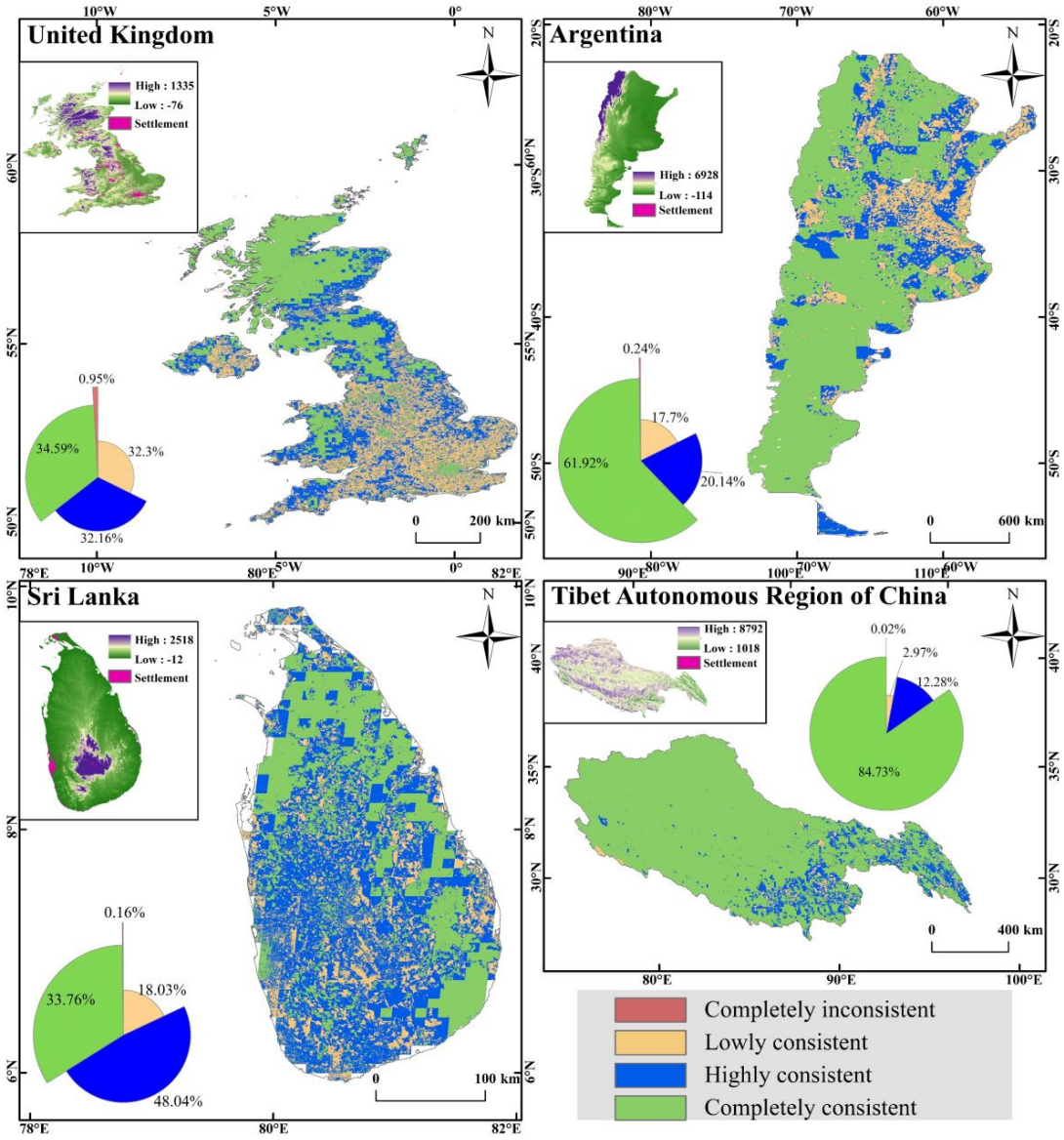

Figure 2. Spatial distribution consistency of four datasets in each case area (completely consistent: four datasets are consistent; highly consistent: three datasets are consistent; lowly consistent: two datasets are consistent; completely inconsistent: each dataset is inconsistent with the others).

Table 5. Statistics of spatial distribution consistency of four sets of data.

	Completely Consistent	Highly Consistent	Lowly Consistent	Cmpletely Inconsistent
United Kingdom	34.59%	32.16%	32.30%	0.95%
Argentina	61.92%	20.14%	17.70%	0.24%
Sri Lanka	33.76%	48.04%	18.03%	0.16%
Tibet Autonomous Region of China	84.73%	12.28%	2.97%	0.02%

Pairwise comparison and analysis of the data show that the highest consistency exists between WorldPop and other data, which may be related to its abundant data sources and auxiliary data and reasonable redistribution rules. In the UK and Argentina, the consistency between WorldPop and GHSL is the highest, and is 5–15% higher than that between WorldPop and GPWv4. In Sri Lanka and the Tibet Autonomous Region of China, the consistency between WorldPop and GPWv4 is the highest, and the

consistency between WorldPop and GHSL is 4–30% lower than that between WorldPop and GPWv4. This indicates that the portrayal of characteristics of population distribution varies depending on the spatialization methods in different case areas, and GHSL, which is integrated with built-up areas extracted from remote sensing, is more advantageous in areas with a high urbanization level (Table 6).

Table 6. Comparative analysis of spatial consistency of data sets.

Case areas	Consistency (%)	HYDE	GPWv4	GHSL	WorldPop
United Kingdom	HYDE	100.00			
	GPWv4	65.27	100.00		
	GHSL	40.65	61.94	100.00	
	WorldPop	52.14	79.72	85.63	100.00
Argentina	HYDE	100.00			
	GPWv4	81.53	100.00		
	GHSL	83.31	66.26	100.00	
	WorldPop	90.59	76.42	91.23	100.00
Sri Lanka	HYDE	100.00			
	GPWv4	79.22	100.00		
	GHSL	44.91	55.18	100.00	
	WorldPop	71.61	90.83	61.13	100.00
Tibet Autonomous Region of China	HYDE	100.00			
	GPWv4	92.21	100.00		
	GHSL	90.60	93.15	100.00	
	WorldPop	93.55	97.68	93.90	100.00

3.3. Consistency of Datasets in Different Population Density Levels

In order to explore the spatial relationship between population density and consistency, we conducted a statistical analysis of the distribution of population density levels in consistent or inconsistent regions. Judging from the distribution of population density levels in completely/highly consistent regions (Figure 3), in the four case areas, each dataset is dominated by low-density population distribution of level 1–3, with an area proportion of more than 45%, which indicates that the data consistency is great among extremely low population density areas. Especially in Tibet and Argentina, the proportion of areas wherein population density is level 1 and 2 is as high as 77%, and in the UK it is 51%. In Sri Lanka, where the population density is high and the spatial distribution is relatively uniform, there are always highly consistent areas for each population density level. In the UK, with a high level of urbanization, 40–82% of the high-density population areas (level 7–9) are highly consistent (WorldPop, 81.72%; GHSL, 66.94%; GPWv4, 58.90%; HYDE, 40.47%). Since HYDE contains historical data for long time series, its spatial resolution is far lower than that of the others. Therefore, in densely populated and highly heterogeneous areas, spatial accuracy will be reduced due to the influence of mixed pixels and the precision of original data, which is reflected in the UK and Sri Lanka (Figure 3.). Lowly consistent/completely inconsistent regions are mainly distributed in the medium population areas with a high urbanization level. Among the medium density population areas in the UK and Argentina, in 62–93% of the regions the four datasets are completely inconsistent, or only two are consistent (Figure 4, Table 7).

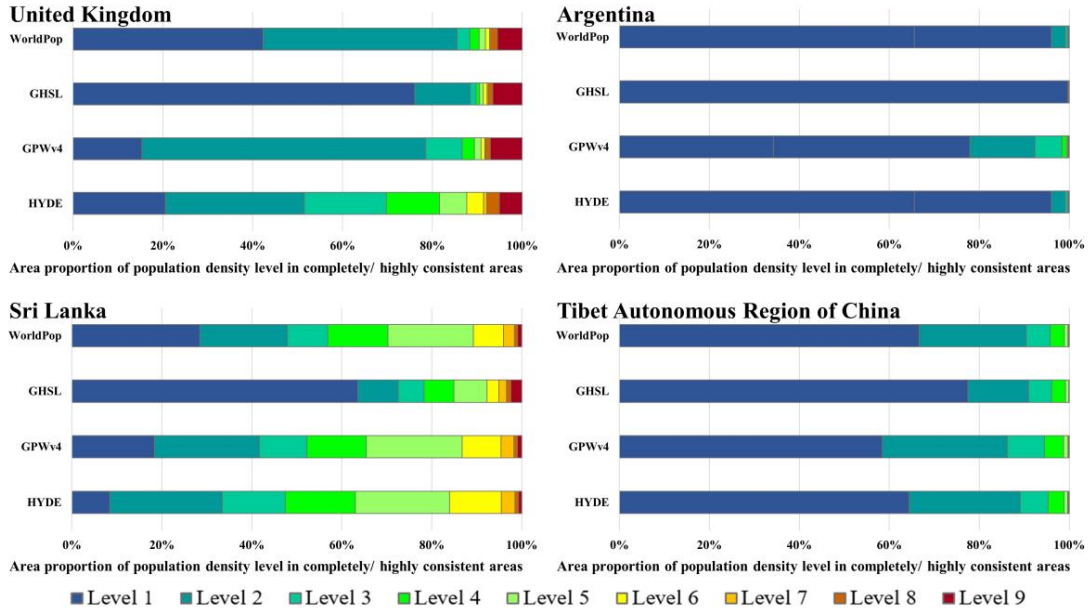

Figure 3. Area proportion of population density level in completely/highly consistent areas.

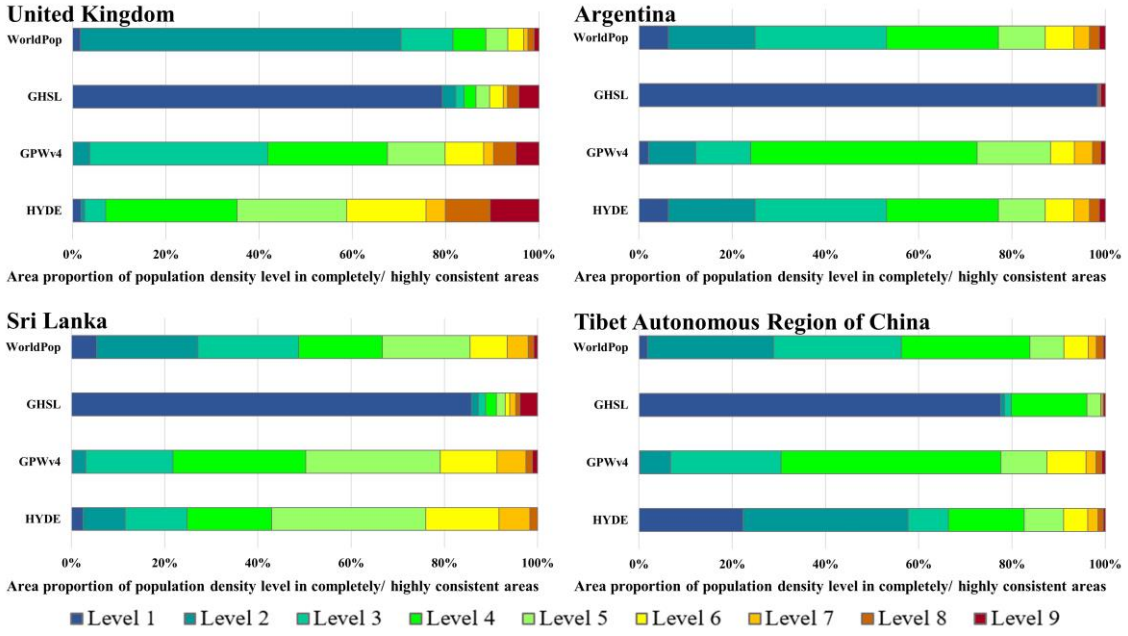

Figure 4. Area proportion of population density level in lowly consistent/completely inconsistent areas.

Table 7. Area proportion of lowly consistent/completely inconsistent areas in medium population density areas.

	HYDE	GPWv4	GHSL	WorldPop
United Kingdom	62.49%	81.77%	63.85%	63.13%
Argentina	93.92%	67.85%	72.03%	93.92%
Sri Lanka	22.19%	25.54%	6.47%	20.31%
Tibet Autonomous Region of China	16.70%	27.31%	13.28%	23.05%

4. Discussion

It can be seen that the spatial patterns of the spatial population datasets produced by different methods and data sources are very similar in Tibet, where data are scarce and the population is sparse.

In the data selection of such regions, the accuracy of population estimation and the time scale needed for research are the main considerations. For regions with high levels of urbanization, we should not only consider spatiotemporal resolution and accurate quantity, but also pay more attention to the uncertainty of data in areas with medium population density. Based on the results, a table is summarized to show the applicability of datasets in different population density areas (Table 8). This study serves as a basis for not only the selection of population data, but also the future development of population spatialization. In areas where data are lacking, improving the accuracy of spatial population datasets depends more on continuously refining demographic data [52–56] and abundant data sources [57]. The difficulty in obtaining data in areas at high altitude and with poor data quality may be the reason for the large relative error in the Tibet Autonomous Region of China [58]. Remote sensing, mobile communication, and other big data will play important roles in improving the accuracy of spatial population data in areas with deficient data. For areas with medium population density, with the development of spatialization methods, from simple interpolation to machine algorithms based on intelligent models such as neural networks, decision trees, genetic algorithms and random forest [9,48,59,60], strengthening the experimental research and verifying such areas will improve the reliability and consistency between datasets. Verifying the accuracy of spatial population datasets is a massive problem in the research. According to the comparison between the population of spatial datasets and census data in this study, not only are there differences in spatial layout, but there is also about 20% deviation in the population. Therefore, in areas with different geographical characteristics and with more detailed statistical units, even at grid scale, it is also a necessary development direction of population spatialization to develop standard experimental areas, and to provide verification data for the accurate quantity and spatiotemporal layout of spatial data designed by various applications. Besides, urban/rural populations are two concepts of population geography corresponding to urban and rural areas. When it comes to urban population in most countries, the population of small cities generally is included, while in China, it usually refers to the population of towns [61]. Although the population scale for towns in China is equivalent to that for small cities of other countries, the difference in definition for urban/rural population may have a slight effect.

Table 8. Applicability of datasets in different population density areas.

Dataset	Accuracy of Population Estimation	Consistency with Other Datasets	Consistency of Datasets in Different Population Density Levels		
			Low-Density Area	Medium-Density Area	High-Density Area
HYDE	★★★	★★	★★★	★	★★
GPWv4	★★★★	★★★	★★	★	★★
GHSL	★★★★	★★	★★★★	★★	★★★★
WorldPop	★★	★★★★	★★★	★	★★★

Notes: The number of stars indicates the degree of applicability of datasets.

5. Conclusions

In order to understand differences in the number and spatial distribution of the main spatial population datasets in the world, four datasets with different spatiotemporal resolutions (HYDE, GPWv4, GHSL and WorldPop), developed based on multiple data sources and spatialization methods, were selected, and Sri Lanka, the UK, Argentina and the Tibet Autonomous Region of China were taken as the case areas. This paper conducted research from the aspects of relative error of population, consistency of population spatial distribution, and the characteristics of population density distribution within consistent and inconsistent regions. Furthermore, this paper analyzed the causes of the differences by combining the data production process and the difficulty of data acquisition, urbanization level and the characteristics of population distribution for the case areas. The results show the following:

(1) The differences in source data and spatialization methods between datasets affect their accuracy. The development of remote sensing and deep learning technology promotes the progress of data collection and spatialization methods. Therefore, the accuracy of each dataset in the study is very different. Because GPWv4 is based on 2010 census data for allocation according to the principle that the population in each administrative unit is unchanged, and GHSL is based on GPWv4 for secondary spatialization, their absolute value for the relative error of total population is the smallest, both of which being within 3%. Although WorldPop uses the same data source as GPWv4, the relative error of the former is as high as 20% in Argentina, Sri Lanka and the Tibet Autonomous Region of China, due to different spatialization methods. HYDE, for the purpose of producing long time series historical data, has medium accuracy for estimating the population of the UK, Argentina and Sri Lanka;

(2) The application of geospatial data makes the datasets more accurate in the UK with abundant information, where the absolute value of the relative error of the four datasets is less than 4%. In other case areas, the absolute value of the relative error of GPWv4 and GHSL is less than 3%, and that of HYDE and WorldPop is between 5% and 25%. Affected by the imprecision of statistical data and the difficulty in obtaining new auxiliary data, the relative error of datasets in the Tibet Autonomous Region of China is relatively large, especially with HYDE using historical literature data and WorldPop using multi-source geographic information data. With regard to the ability to describe spatial distribution, the pairwise consistency between WorldPop and the other three datasets is the highest due to the fusion of multiple data sources, and GHSL, which mixes built-up area distribution information extracted from remote sensing, has more advantages in terms of spatial consistency in areas with a high urbanization level. It is difficult to spatialize population distribution in areas with complex variation, characterized by reduced consistency in spatial distribution. The consistency of population spatial distribution for the four datasets is the highest in the Tibet Autonomous Region of China, where the total proportion of four and three datasets being consistent is as high as 97.01%. On the other hand, in the UK, where the population spatial distribution is complex, only 66.75% of the regions are completely or highly consistent;

(3) Areas where the four datasets are completely/highly consistent are mainly distributed in low population density areas. In Tibet, Argentina and the UK, the proportions of level 1 and 2 in completely/highly consistent areas are as high as 89%, 76% and 92,% respectively, indicating that data consistency is great in low-density areas. In addition, in highly urbanized and densely populated areas, the spatial distribution of each dataset is also highly consistent, and 62% of high-density population areas in the UK are completely/highly consistent areas. The lowly consistent/completely inconsistent regions are mainly distributed in the middle density areas with a high urbanization rate, and 62–93% of middle density population areas in the UK and Argentina are lowly consistent/completely inconsistent regions.

Author Contributions: Conceptualization, Huimin Yan and Yanzhao Yang; methodology, Ruxia Chen; data curation, Ruxia Chen and Wenpeng Du; writing—original draft preparation, Ruxia Chen; writing—review and editing, Huimin Yan, Fang Liu and Wenpeng Du; visualization, Ruxia Chen; supervision, Yanzhao Yang; funding acquisition, Huimin Yan. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA19040301) and the second Tibetan Plateau Scientific Expedition and research (Grant No. 2019QZKK1006).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, W.H.; Niu, S.W. A comparative study on the impact of population growth and consumption increase on China's resource-environment. *Chin. J. Popul. Sci.* **2009**, *2*, 66–73, 112.
2. Du, W.P.; Yan, H.M.; Yang, Y.Z.; Liu, F. Evaluation methods and research trends for ecological carrying capacity. *J. Resour. Ecol.* **2018**, *9*, 115–124.
3. Clarke, J.I.; Rhind, D.W.; Becket, C.; Wilkes, A.; Sadler, G.; Short, J. *Population Data and Global Environmental Change*; ISSC: Paris, France, 1992.

4. Wei, J.B.; Xiao, D.N.; Xie, F.J. Evaluation and regulation principles for the effects of human activities on ecology and environment. *Prog. Geogr.* **2006**, *2*, 36–45.
5. Chen, W.X.; Li, J.F.; Zeng, J.; Ran, D.; Yang, B. Spatial heterogeneity and formation mechanism of eco-environmental effect of land use change in China. *Geogr. Res.* **2019**, *38*, 2173–2187.
6. Wilson, S.J.; Steenhuisen, F.; Pacyna, J.M.; Pacyna, E.G. Mapping the spatial distribution of global anthropogenic mercury atmospheric emission inventories. *Atmos. Environ.* **2006**, *40*, 4621–4632. [CrossRef]
7. Li, B. The Research on Urban Heat Island Effect of the Transboundary Area in the Tumen River from 2003 to 2016. Master's Thesis, Yanbian University, Yanbian, China, 2019.
8. Fu, H.Y.; Li, M.C.; Zhao, J.; Liu, Y.X. Summary of grid transformation models of population data. *Hum. Geogr.* **2006**, *21*, 115–119.
9. Jin, J.; Li, C.M.; Yin, J.; Lin, Z.J. Investigation on the model for spatial distribution of population data. *Acta Geod. Et Cartogr. Sin.* **2003**, *3*, 278–282.
10. Bai, Z.Q.; Wang, J.L.; Yang, F. Research progress in spatialization of population data. *Prog. Geogr.* **2013**, *32*, 1692–1702.
11. Center for International Earth Science Information Network-CIESIN-Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11[DB/OL]*; NASA Socioeconomic Data and Applications Center (SEDAC): Palisades, NY, USA, 2018.
12. Schiavina, M.; Freire, S.; MacManus, K. *GHS Population Grid, Derived from GPW4, Multitemporal (1975, 1990, 2000, 2015) [DB/OL]*; European Commission, Joint Research Centre, JRC Data Catalogue: Ispra, Italy, 2015.
13. Goldewijk, K.K.; Beusen, A.; Drecht, G.V.; Martine, D.V. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Glob. Ecol. Biogeogr.* **2010**, *20*, 73–86. [CrossRef]
14. Goldewijk, K.K.; Beusen, A.; Janssen, P. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *Holocene* **2010**, *20*, 565–573. [CrossRef]
15. Goldewijk, K.K.; Beusen, A.; Doelman, J.; Stehfest, E. Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth Syst. Sci. Data* **2017**, *9*, 927–953. [CrossRef]
16. WorldPop (www.worldpop.org-School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur); Center for International Earth Science Information Network (CIESIN), Columbia University. *WorldPop[DB/OL]*; University of Southampton: Southampton, UK, 2018.
17. Smith, A.; Bates, P.D.; Wing, O.; Christopher, S.; Quinn, N.; Neal, J. New estimates of flood exposure in developing countries using high-resolution population data. *Nat. Commun.* **2019**, *10*, 1814. [CrossRef] [PubMed]
18. Pesaresi, M.; Ehrlich, D.; Florczyk, A.J.; Freire, S. The global human settlement layer from landsat imagery. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10 July 2016; pp. 7276–7279.
19. Kibret, S.; Lautze, J.; McCartney, M.; Wilson, G.G.; Nhamo, L. Malaria impact of large dams in sub-Saharan Africa: Maps, estimates and predictions. *Malar. J.* **2015**, *14*, 339. [CrossRef]
20. World Resources Institute. World Resources Report 2010-2011: Decision Making in a Changing Climate. *Sustainability* **2011**, *4*, 305.
21. Smith, K. We are seven billion. *Nat. Clim. Chang.* **2011**, *1*, 331–335. [CrossRef]
22. Gleeson, T.; Wada, Y.; Bierkens, M.F.P.; van Beek, L.P.H. Water balance of global aquifers revealed by groundwater footprint. *Nature* **2012**, *488*, 197–200. [CrossRef]
23. Balk, D.L.; Nghiem, S.V.; Jones, B.R.; Liu, Z. Up and out: A multifaceted approach to characterizing urbanization in Greater Saigon, 2000-2009. *Landsc. Urban Plan.* **2019**, *187*, 199–209. [CrossRef]
24. Melchiorri, M.; Florczyk, A.J.; Freire, S.; Schiavina, M.; Pesaresi, M.; Kemper, T. Unveiling 25 years of planetary urbanization with remote sensing: Perspectives from the Global Human Settlement Layer. *Remote Sens.* **2018**, *10*, 768. [CrossRef]
25. Goldewijk, K.K. Three centuries of global population growth: A spatial referenced population (density) database for 1700–2000. *Popul. Environ.* **2005**, *26*, 343–367. [CrossRef]
26. Seto, K.C.; Guneralp, B.; Hutyra, L.R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16083–16088. [CrossRef]

27. Thomson, D.R.; Linard, C.; Vanhuysse, S.; Steele, J.E.; Shimoni, M.; Siri, J.; Caiaffa, W.T.; Rosenberg, M.; Wolff, E.; Grippa, T.; et al. Extending data for urban health decision n-making: A menu of new and potential neighborhood-level health determinants datasets in LMICs. *J. Urban Health-Bull. N. Y. Acad. Med.* **2019**, *96*, 514–536. [CrossRef] [PubMed]
28. Ouma, P.O.; Maina, J.; Thuranira, P.N.; Macharia, P. Access to emergency hospital care provided by the public sector in sub-Saharan Africa in 2015: A geocoded inventory and spatial analysis. *Lancet Glob. Health* **2018**, *6*, e342–e350. [CrossRef]
29. Sorichetta, A.; Bird, T.J.; Ruktanonchai, N.W.; Erbach-Schoenberg, E.Z.; Pezzulo, C.; Tejedor, N.; Waldock, I.C.; Sadler, J.D.; Garcia, A.J.; Sedda, L.; et al. Mapping internal connectivity through human migration in malaria endemic countries. *Sci. Data* **2016**, *3*, 160066. [CrossRef] [PubMed]
30. Goldewijk, K.K.; Ramankutty, N. Land cover change over the last three centuries due to human activities: The availability of new global data sets. *GeoJournal* **2004**, *61*, 335–344. [CrossRef]
31. Ellis, E.C.; Goldewijk, K.K.; Siebert, S.; Lightman, D. Anthropogenic transformation of the biomes, 1700 to 2000. *Glob. Ecol. Biogeogr.* **2010**, *19*, 589–606. [CrossRef]
32. Houweling, S.; van der Werf, G.R.; Goldewijk, K.K.; Röckmann, T. Early anthropogenic CH_4 emissions and the variation of CH_4 and $^{13}CH_4$ over the last millennium. *Glob. Biogeochem. Cycles* **2008**, *22*, GB10021. [CrossRef]
33. Gaston, K.J.; Blackburn, T.M.; Goldewijk, K.K. Habitat conversion and global avian biodiversity loss. *Proc. R. Soc. B-Biol. Sci.* **2003**, *270*, 1293–1300. [CrossRef]
34. Maisels, F.; Strindberg, S.; Blake, S.; Wittemyer, G. Devastating decline of forest elephants in Central Africa. *PLoS ONE* **2013**, *8*, e59469. [CrossRef]
35. MacPherson, P.; Khundi, M.; Nliwasa, M.; Choko, A.T.; Phiri, V.K.; Webb, E.L.; Dodd, P.J.; Cohen, T.; Harris, R.; Corbett, E.L. Disparities in access to diagnosis and care in Blantyre, Malawi, identified through enhanced tuberculosis surveillance and spatial analysis. *BMC Med.* **2019**, *17*, 21. [CrossRef]
36. Sun, Z.C.; Xu, R.; Du, W.J.; Wang, L.; Lu, D.S. High-resolution urban land mapping in China from Sentinel 1A/2 imagery based on Google Earth Engine. *Remote Sens.* **2019**, *11*, 752. [CrossRef]
37. Ye, T.T.; Zhao, N.Z.; Yang, X.C.; Ouyang, Z.T.; Liu, X.P.; Chen, Q.; Hu, K.J.; Yue, W.Z.; Qi, J.G.; Li, Z.S.; et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Sci. Total Environ.* **2019**, *658*, 936–946. [CrossRef]
38. Tan, M.; Lin, K.; Liu, L.; Zhu, Y.H.; Wang, D.S. Spatialization of population in the Pearl River Delta in 30 m grids using random forest model. *Prog. Geogr.* **2017**, *36*, 1304–1312.
39. Bai, Z.Q.; Wang, J.L.; Wang, M.M.; Gao, M.X.; Sun, J.L. Accuracy assessment of multi-source gridded population distribution datasets in China. *Sustainability* **2018**, *10*, 1363. [CrossRef]
40. Leyk, S.; Uhl, J.H.; Balk, D.; Jones, B. Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States. *Remote Sens. Environ.* **2018**, *204*, 898–917. [CrossRef]
41. Yang, X.H.; Wang, N.B.; Jiang, D.; Xiong, L.Y.; Liu, H.H. Regionalization of population distribution based on spatial analysis. *Acta Geogr. Sin.* **2002**, *57*, 76–81.
42. Sliuzas, R.; Kuffer, M.; Kemper, T. Assessing the quality of Global Human Settlement Layer products for Kampala, Uganda. In Proceedings of the 2017 Joint Urban Remote Sensing Event, Dubai, UAE, 6–8 March 2017.
43. Wang, X.M.; Li, X.; Ma, M.G. Pixelizing the population statistics of inland river basin in arid regions—A case study of Heihe River. *J. Arid Land Resour. Environ.* **2007**, *6*, 39–47.
44. Lai, C.X.; Yan, H.M.; Du, W.P.; Hu, Y.F. The variations and causes of grassland distribution in Kazakhstan from the global land cover datase. *J. Geo-Inf. Sci.* **2019**, *21*, 372–383.
45. Wang, X.M.; Li, X.; Ma, M.G. Advance and case analysis in population spatial distribution based on remote sensing and GIS. *Remote Sens. Technol. Appl.* **2004**, *19*, 320–327.
46. Wang, C.; Kan, A.K.; Zeng, Y.L.; Li, G.Q.; Wang, M.; Ci, R. Population distribution pattern and influencing factors in Tibet based on random forest model. *Acta Geogr. Sin.* **2019**, *74*, 664–680.
47. McEvedy, C.; Jones, R. Atlas of world population history. *Med. Hist.* **1979**, *23*, 242.
48. Zhao, Z.W. A Concise History of World Population (Fourth edition) [Book Review]. *J. Popul. Res.* **2007**, *2*, 253–254. [CrossRef]
49. Li, X.L.; Xiu, C.L.; Шендрик, A.B.; Wang, Q. Comparing spatial pattern of population density of Sino-Russian large costal metropolitans: Case study of St. Petersburg and Dalian. *Econ. Geogr.* **2018**, *38*, 78–86.
50. Wang, W.; Yan, H.M.; Yang, Y.Z.; Du, W.P. Evaluation of land resources carrying capacity of Tibetan counties based on dietary nutritional demand. *J. Nat. Resour.* **2019**, *34*, 921–933.

51. Ge, M.L.; Feng, Z.M. Population distribution of China based on GIS: Classification of population densities and curve of population gravity centers. *Acta Geogr. Sin.* **2009**, *64*, 202–210.
52. Broadberry, S. The world economy: A millennial perspective. *Bus. Hist.* **2002**, *44*, 158–160.
53. Wardrop, N.A.; Jochem, W.C.; Birdird, T.J.; Chamberlain, H.R.; Clarke, D.; Kerr, D.; Bengtsson, L.; Juran, S.; Seaman, V.; Tatem, A.J. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3529–3537. [CrossRef] [PubMed]
54. Wang, K.J.; Cai, H.Y.; Yang, X.H.; Zhang, Y. Spatialization method for census data based on reclassifying residential land use in urban areas—A case study in the middle reaches of the Yangtze River Watershed. *Remote Sens. Technol. Appl.* **2015**, *30*, 987–995.
55. Dong, N.; Yang, X.H.; Cai, H.Y. A method for demographic data spatialization based on residential space attributes. *Prog. Geogr.* **2016**, *35*, 1317–1328.
56. Ye, J.; Yang, X.H.; Jiang, D. The grid scale effect analysis on town leveled population statistical data spatialization. *J. Geo-Inf. Sci.* **2010**, *12*, 40–47. [CrossRef]
57. Gao, Z.H. Study on Spatial Distribution of Statistical Data in Regional Ecology and Environment Assessment—A Case Study of Shandong Province. Master's Thesis, Shandong Normal University, Jinan, China, 2012.
58. Leyk, S.; Gaughan, A.E.; Adamo, S.B.; Sherbinin, A.D.; Balk, D.; Freire, S.; Rose, A.; Stevens, F.R.; Blankespoor, B.; Frye, C.; et al. The spatial allocation of population: A review of large-scale Gridded population data products and their fitness for use. *Earth Syst. Sci. Data* **2019**, *11*, 1385–1409. [CrossRef]
59. Wang, K.J. Multi-Scales Spatialization Modeling for Statistical Demographic Data. Master's Thesis, East China Normal University, Shanghai, China, 2015.
60. Qiu, G. High-Resolution Population Mapping Using a Random Forest Model: A Case Study in Zhengzhou. Master's Thesis, Graduate School of Inner Mongolia Normal University, Huhhot, China, 2019.
61. Shi, Y.L. Urban and rural division and urban population statistics: A comparative study between China and foreign countries. *Urban Probl.* **1993**, *1*, 22–26.