International Journal of Geo-Information

MDPI

Article

A Hybrid Population Distribution Prediction Approach Integrating LSTM and CA Models with Micro-Spatiotemporal Granularity: A Case Study of Chongming District, Shanghai

Pengyuan Wang [1], Xiao Huang [2], Joseph Mango [1,3], Di Zhang [1], Dong Xu [1] and Xiang Li [1,*]

[1] Key Laboratory of Geographical Information Science (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China; 51173901039@stu.ecnu.edu.cn (P.W.); jsmmango@stu.ecnu.edu.cn (J.M.); zhangdee@stu.ecnu.edu.cn (D.Z.); dongxu_01@stu.ecnu.edu.cn (D.X.)
[2] Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA; xh010@uark.edu
[3] Department of Transportation and Geotechnical Engineering, University of Dar es Salaam, Dar es Salaam 35131, Tanzania
* Correspondence: xli@geo.ecnu.edu.cn

Abstract: Studying population prediction under micro-spatiotemporal granularity is of great significance for modern and refined urban traffic management and emergency response to disasters. Existing population studies are mostly based on census and statistical yearbook data due to the limitation of data collecting methods. However, with the advent of techniques in this information age, new emerging data sources with fine granularity and large sample sizes have provided rich materials and unique venues for population research. This article presents a new population prediction model with micro-spatiotemporal granularity based on the long short-term memory (LSTM) and cellular automata (CA) models. We aim at designing a hybrid data-driven model with good adaptability and scalability, which can be used in more refined population prediction. We not only try to integrate these two models, aiming to fully mine the spatiotemporal characteristics, but also propose a method that fuses multi-source geographic data. We tested its functionality using the data from Chongming District, Shanghai, China. The results demonstrated that, among all scenarios, the model trained by three consecutive days (ordinary dates), with the granularity of one hour, incorporated with road networks, achieves the best performance (0.905 as the mean absolute error) and generalization capability.

Keywords: population distribution prediction; micro-spatiotemporal granularity; long short-term memory (LSTM); cellular automata (CA); integrated model

Citation: Wang, P.; Huang, X.; Mango, J.; Zhang, D.; Xu, D.; Li, X. A Hybrid Population Distribution Prediction Approach Integrating LSTM and CA Models with Micro-Spatiotemporal Granularity: A Case Study of Chongming District, Shanghai. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 544. https://doi.org/10.3390/ijgi10080544

Academic Editor: Wolfgang Kainz

Received: 3 June 2021
Accepted: 10 August 2021
Published: 13 August 2021

1. Introduction

Population distribution prediction refers to estimating the population of a specific geographic unit, taking into account the impact of natural geography and the socioeconomic environment and applying scientific methods (predictive models) to estimate population development in another temporal period [1]. However, traditional population prediction generally has extremely coarse spatiotemporal granularity. With the advent of information handling techniques, new data collection methods, data processing algorithms, and improvements in computing power have made population prediction with micro-spatiotemporal granularity possible. Since ancient times, population estimation and prediction have been essential in human society for policymaking and socioeconomic planning, such as urban and health care planning at different administration levels. Studies on population prediction can be traced back to the United Kingdom in 1695 [2]. After years of exploration and efforts by mathematicians, statisticians, demographers, geographers, and other scholars, a series of models for population prediction is proposed.

In general, population prediction models can be grouped into four categories, given the approaches they adopt: (1) mathematical models, (2) statistical models, (3) demographic

models, and (4) machine learning models. In the past, many mathematical models were proposed [3,4]. Due to insufficient assumptions, however, data utilization is limited in these mathematical models, and it is difficult to determine parameters in an objective manner. The above constraints led to the development and application of statistical models in population prediction. Statistical models in population prediction mainly include regression, such as linear regression and non-linear regression models (e.g., Gompertz models), and time series models such as the autoregressive integrated moving average model (ARIMA) that combines differential, moving average, and autoregressive analysis. For example, Zakria and Muhammad [5] applied this model to simulate the population dynamics in Pakistan from 1951 to 2007, and they predicted that the country's population would reach 229 million by 2025. Despite more data and parameters that ultimately improve the prediction results, statistical models often fail to consider other factors that play significant roles in population growth, such as factors regarding societal settings and economic influences. In comparison, demographic models such as the cohort-component models consider not only the information of birth-rate, mortality-rate, and migration-rate but also age-specific information, providing detailed descriptions on the population spectrum, thus rendering more accurate results compared to pure statistical methods [6]. A study by Stoto found that the distribution of error statistics in the US Census Bureau and the UN's forecast samples was relatively stable, and such information could be utilized to predict the total population of the United States in 2000 [7]. Lee and Carter [8] used the time series method to predict age-specific mortality in the United States between 1990 and 2065. In 1996, based on an expert's judgment on the trends and uncertainties of future birth rate, mortality, and migration rates in global regions, Lutz et al. [9] predicted the demographics at global and selected regions by 2100.

Based on this study's perspectives, all methods mentioned above can be classified as the traditional methods because they largely rely on structured census and statistical data from yearbooks. Moreover, their predictions tend to have coarse spatiotemporal granularity, with yearly temporal units and coarse spatial scales (e.g., national or global). Recently, the improvement of data collecting methods and the arrival of the Big Data era, the development of Big Data analytics, and the increasing maturity of machine learning have fostered new ideas in population prediction studies. For example, Reichstein et al. [10] argued that the next-generation researches in geographic sciences should involve hybrid models that integrate physical and machine learning models. Robinson et al. [11] trained a Convolutional Neural Network (CNN) [12] with one-year synthetic remote sensing images of Landsat with a grid resolution of 0.01×0.01 (unit: degree) to predict the US population. Griffith et al. [13] observed the increasing popularity of fine-scale population studies and indicated that the future prediction methods could start at the family or even the individual level, thanks to the development of new modeling algorithms, such as agent models and artificial neural networks (ANN).

LSTM is a variation of recurrent neural networks (RNN) first proposed by Hochreiter and Schmidhuber in 1997 [14]. Since RNNs often experience problems of gradient vanishing during training, such networks can only retain short-term memory. LSTM relies on a unique set of internal mechanisms (e.g., gates and memory cells) to grant the networks the capability of long-term memory. It has been widely adopted in handwriting [15] and speech recognition [16]. However, we have found that this feature is particularly suitable for expressing the irregularities within the population change. The cellular automata (CA) with discrete states and local spatiotemporal interactions were proposed by American mathematician Stanisław Marcin Ulam in the 1940s and were used by Von Neumann to study the logical nature of self-replicating systems (White and Engelen, 1993) [17]. At present, CA models are mainly used for modeling complex systems. Compared with mathematical models, CA can simulate the evolution of natural phenomena in a more accurate manner [18], and therefore it has been adopted in simulation studies such as human migration, urban development, and land-use dynamics [19,20].

This study uses mobile phone signaling data, an emerging type of population data different from population data collected by traditional methods. Mobile phone data enables population prediction with micro-spatiotemporal granularity, characterized by fine-grained spatiotemporal granularity compared with traditional methods. The temporal resolution can be as fine as hours or even minutes, and the spatial resolution can be at a sub−1 km scale. Mobile phone signaling data have extensive spatial coverages, high data collection frequencies, long temporal spans, and spatial grid as a regular statistical unit. Based on the literature, extremely few methods can handle population datasets with such spatiotemporal granularity. LSTM has the ability to solve long-term dependency problems, given its unique design structure. Its design makes it suitable for processing and predicting important events with long intervals and delays in time series, ideal for investigating population dynamics. CA is a dynamic grid model with discrete time, space, and state. Since its spatial interactions and time causality are local-focused, CA is often adopted to simulate the process of the spatiotemporal evolution of complex systems (e.g., geographical environment). In addition, the concept of cells in CA is suitable to express the spatial distribution of the population.

Although there exist some efforts that integrate LSTM and CA models [21,22], their models lack the ability to integrate multi-source geographic information and, to our best knowledge, no effort has been made to investigate such integration in predicting the spatial distribution of population. In light of the background of population prediction with micro-spatiotemporal granularity and the unique characteristics of such data, we attempt to build a data-driven population prediction framework by integrating two basic models: LSTM and CA.

2. Study Area and Data

2.1. Overview of the Study Area

This study uses data from Chongming District (extending 121°09′30″ E to 121°54′00″ E and 31°27′00″ N to 31°51′15″ N), which is a municipal district in Shanghai, China. It borders the Yangtze River to the west and the East China Sea to the east. On the northern side, it borders Haimen City and Qidong City. Towards the south, the Chongming District is surrounded by three other districts of Pudong, Jiading, and Baoshan (Figure 1). With a total area of 1413 square kilometers, the Chongming District is mainly composed of islands that include Chongming Island, Hengsha Island, and Changxing Island. Among these, Chongming Island is the third-largest Chinese island and the world's largest estuary alluvial island [23].

Located at the midpoint of the Chinese coastline, Chongming District is the estuary of the Yangtze River, the longest river in China. The terrain of Chongming Island is largely flat, and the elevation of more than 90% of the area is between 3.21 m and 4.20 m (Wujing Elevation). According to the 2018 Shanghai Yearbook Statistics, by the end of 2017, the registered population in Chongming District was 675,900, including 694,600 permanent residents and 141,900 foreign residents. In the whole district, the density of population is 586 people per square kilometer. The natural population growth rate was −0.51%. In 2017, according to the Statistical Communique of National Economic and Social Development of Chongming District, the per capita disposable income of all residents in the region was 5654 US dollars, and the per capita disposable income of rural residents was 3930 US dollars. The annual GDP was 5.14 billion US dollars, of which the tertiary industries grew up more by 12.8% over the previous year, thus making a total of 2.66 billion US dollars, equivalent to 51.8% of the region's GDP.

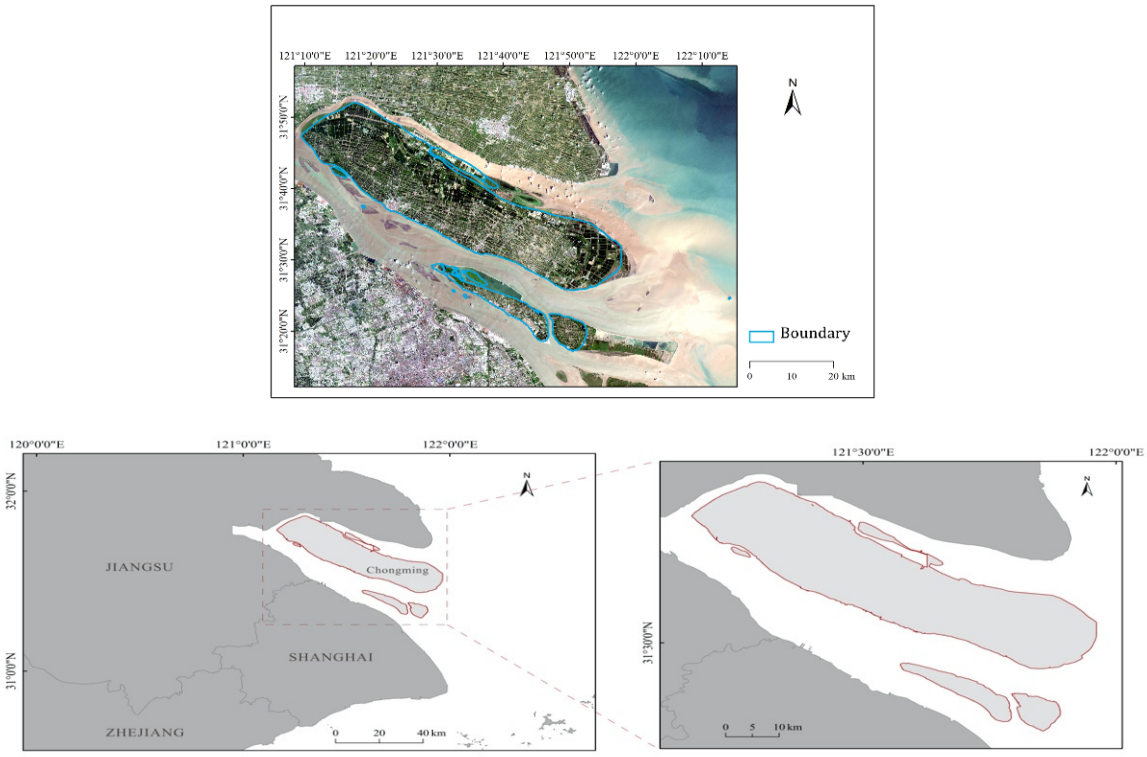

Figure 1. Location of the study area: Chongming District.

2.2. Data Characteristics

The data used in this study are mobile phone signaling data collected by China Unicom, covering the entire Chongming District, Shanghai, with a temporal span from 27 September 2018, to 10 October 2018. As mobile phone signaling data involve sensitive and private information, the actual data has been aggregated to 10 min as the temporal unit and to regular grids of 250 m by 250 m. We are not involved in the mobile phone signaling data collection and aggregation process, but we can confirm that the data we retrieve has been anonymized. This dataset collects mobile phone signaling data, precisely reflecting the number of mobile services, thus serving as an ideal proxy to population distribution. The file size is about 871 MB (MByte), with a total of 21,833,490 lines of records. The original data is in text format, with each line storing one record and fields separated by "|" (Figure 2). There are four attribute fields. Table 1 presents detailed descriptions of each field. The longitude and latitude is the central point of the aggregated population data, and central point value suggests the number of people in each grid of 250 m by 250 m.

```
1  2018/10/10 15:40|121.353637|31.808907|7
2  2018/10/08 14:40|121.576139|31.684627|18
3  2018/10/02 20:50|121.353637|31.808907|7
4  2018/10/01 05:40|121.51528|31.66222|55
5  2018/10/03 21:50|121.353637|31.808907|7
6  2018/10/05 07:30|121.462387|31.59699|66
7  2018/09/29 06:50|121.374886|31.626344|36
8  2018/10/03 11:10|121.374886|31.626344|26
9  2018/10/06 14:10|121.374886|31.626344|32
10 2018/10/01 16:50|121.353637|31.808907|7
```

Figure 2. Raw format of mobile phone signaling data.

Table 1. Field descriptions of mobile phone signaling data.

Field Meaning	Format	Remark
Time	Counting time year/month/day/hour: minute	Accurate to 10 min
Longitude	Floating point number retained to six decimal places	WGS-84 geographic coordinate system
Latitude	Floating point number retained to six decimal places	WGS-84 geographic coordinate system
Statistics	Integer	Mobile users in the grid

The original data contain collected time, latitude, and longitude information. Since the integrated model requires serialization characteristics for the input data, we re-ranked the data according to the timestamp and divided them into individual files by day. Finally, 14 text files are obtained, ranging from 27 September 2018 to 10 October 2018, sorted by timestamp. Thus, data of different longitudes and latitudes at the same time are gathered together. It should be noted that all data are referenced on the World Geographic Coordinate System (WGS-84), and their spatial sampling accuracy (250 m × 250 m) is determined during collection. Since it is difficult to determine spatial characteristics of data in text format, we used the Geospatial Data Abstraction Library (GDAL) to convert these text files into raster format to facilitate visual representation. Figure 3 presents the population distribution at 12:00 p.m. on 27 September 2018 in Chongming District in a raster format. The cell attribute indicates the number of people in each grid.

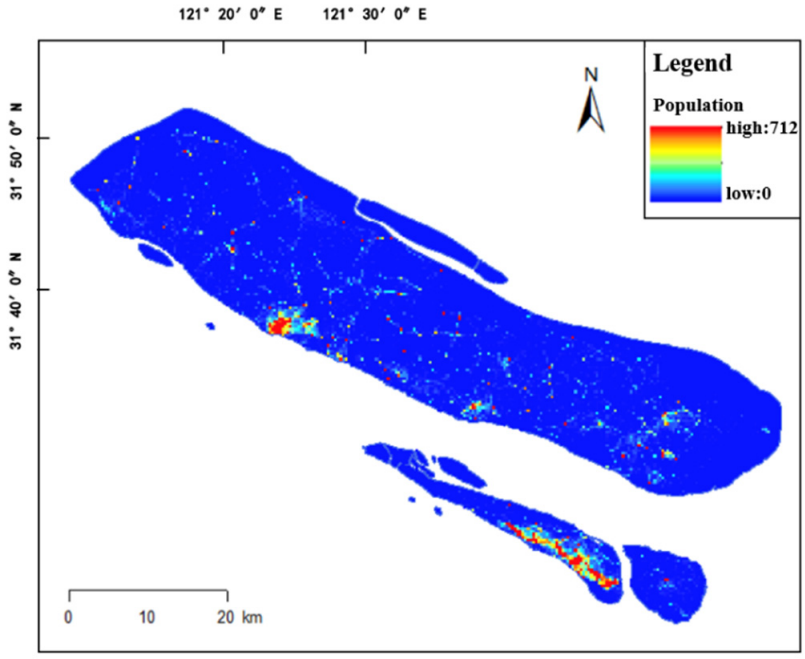

Figure 3. Population distribution in Chongming District in a raster format at 12:00 p.m. on 27 September 2018.

3. Method of Model Integration

3.1. Overview of the LSTM and CA Models

In this study, the temporal period for data collection is considerably long. Compared with classic time series models, LSTM can benefit the extraction of long-term features in the temporal dimension and the extraction of those features in an unsupervised manner, largely increasing its utility and generalizability. The internal structure of the neural networks is presented in Figure 4. X_t denotes the original input in the current time step. C_{t-1} and C_t represent the memory units of the previous time step and the current time step, respectively. H_{t-1} and H_t represent the hidden state of the previous time step and the current time step, respectively.

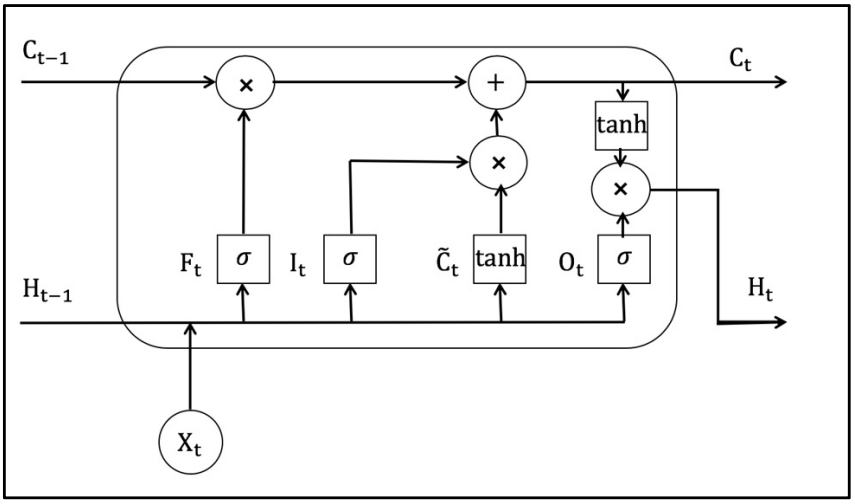

Figure 4. The internal structure of long short-term memory (LSTM) neural networks.

As mentioned above, to mitigate the problem of short-term memory, LSTM constructs a special structure inside the networks with memory cells, which can pass the information along the sequence, and the structure can be understood as the "memory" of the networks. It includes three gates: forget gate, input gate, and output gate.

Given time step t, the calculating process of the forget gate, input gate, and output gate can be presented as follows:

$$f_t = \sigma\left(x_t W_{xf} + h_{t-1} W_{hf} + b_f\right) \tag{1}$$

$$i_t = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \tag{2}$$

$$o_t = \sigma(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \tag{3}$$

where f_t, i_t, $o_t \in \mathbb{R}^{n*h}$ (n is the number of samples and h is the number of hidden units) are forget gate, input gate, and output gate, respectively. The current time step $x_t \in \mathbb{R}^{n*d}$ (d is the number of features) is the input, and the hidden state of the previous time step is $h_{t-1} \in \mathbb{R}^{n*h}$. W_{xf}, W_{xi}, $W_{xo} \in \mathbb{R}^{d*h}$ and W_{hf}, W_{hi}, $W_{ho} \in \mathbb{R}^{h*h}$ are the weights. b_f, b_i, $b_o \in \mathbb{R}^{1*h}$ denote the bias parameters.

At time step t, the candidate memory cell $\widetilde{C}_t \in \mathbb{R}^{n*h}$ is calculated as:

$$\widetilde{C}_t = tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \tag{4}$$

where $W_{xc} \in \mathbb{R}^{d*h}$ and $W_{hc} \in \mathbb{R}^{h*h}$ are the weights of the parameters, and $b_c \in \mathbb{R}^{1*h}$ is the bias parameter. The update of the memory cells is based on the memory cells passed in the previous time step. The combination of the forget gate, and the input gate is able to control the flow of internal information:

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t \tag{5}$$

where the forget gate f_t determines how much information is retained in the memory cell $C_{t-1} \in \mathbb{R}^{n*h}$, and the input gate i_t determines how much information is added in the candidate memory cell $\widetilde{C}_t$.

The hidden state is realized by controlling the flow of internal information of the memory cell through the output gate. The calculating process of the hidden state is as follows:

$$h_t = o_t * tanh(C_t) \tag{6}$$

With the gates mentioned above, LSTM can effectively extract long-term temporal patterns in mobile phone signaling data.

A CA system contains four basic elements: cells, states, neighborhoods, and conversion rules. Cells are the basic objects (or units) of a CA system. Space formed by all cells is called cell space. States of cells (generally discrete) have different choices based on different scenarios. However, the population data in this study are continuous. Conversion rules act on local scopes, and neighboring cells need to be defined to form neighborhoods. The conversion rule is a function that determines the state of the next moment according to the current state of the cell and the states of the neighboring cells. Therefore, this study aims to determine the neighborhood and extract the conversion rules effectively.

3.2. Model Integration

The integrated model we proposed is a data-driven model that can adapt to the spatiotemporal characteristics of mobile phone signaling data. Due to the high frequency and long acquisition period, classic time series methods such as ARIMA may not be feasible, as they cannot analyze multi-dimensional data and require stationary data or stationary after differentiating. LSTM benefits the extraction of both short-term and long-term temporal features. In addition, the population data is collected at the grid level, similar to the concept of cells in CA. Moreover, the neighborhood and extraction algorithms of conversion rules in CA have good scalability. Therefore, we integrate these two models to build a population prediction model with a micro-spatiotemporal granularity that can fully adapt to the characteristics of mobile phone signaling data. In the process of integrating LSTM and CA models, we consider two main aspects. First, we need to take advantage of the respective advantages of LSTM in sequence data and CA in spatial simulation. Second, we need to ensure the generalizability of the model.

From a macro perspective, a CA model is the upper-level framework of the integrated model, which defines the interaction between the cells in the simulation space through varying definitions of neighborhoods. The LSTM can be seen as an underlying algorithm for automatically extracting conversion rules in CA. In this study, LSTM is able to handle sequential data with the input of a three-dimensional tensor that includes samples (corresponding to the grids), timesteps, and features. Relying on the upper-level CA model, features can incorporate geographical information from various neighboring scenarios. The output of the CA-LSTM-integrated model is a two-dimensional tensor that includes samples (grids) and the corresponding population counts (in this study, mobile user counts). In this integration framework, CA and LSTM, respectively, handle information from spatial and temporal perspectives. Their integration benefits efficient spatiotemporal data fusion that leads to accurate population prediction. Specifically, there are two considerations: one is to use the cells in the CA to represent the grid. Different definitions of the neighborhood can incorporate more spatial information into the model. The second is to take advantage of the unique mechanism of LSTM for extracting complex time series features in population data. Figure 5 presents the schematic diagram of the integration between LSTM and CA.

In Figure 5, X denotes the original input while X' denotes the model input after temporal selection and neighborhood construction. Different uses of t represent different timestamps. C_{t-1}, C_t, H_{t-1}, and H_t are the intermediate input and output of the LSTM unit during the training process, and y represents the model output.

In general, the integrated model first selects the time series of the original data, then builds features in the attribute dimension according to the definition of the neighborhood in CA. Further, structurally transformed data are fed into the LSTM for training. The calculation at each time step generates new memory cells and hidden states. At the end of the entire training process, LSTM learns a set of parameters inside the model that can be used for later predictions.

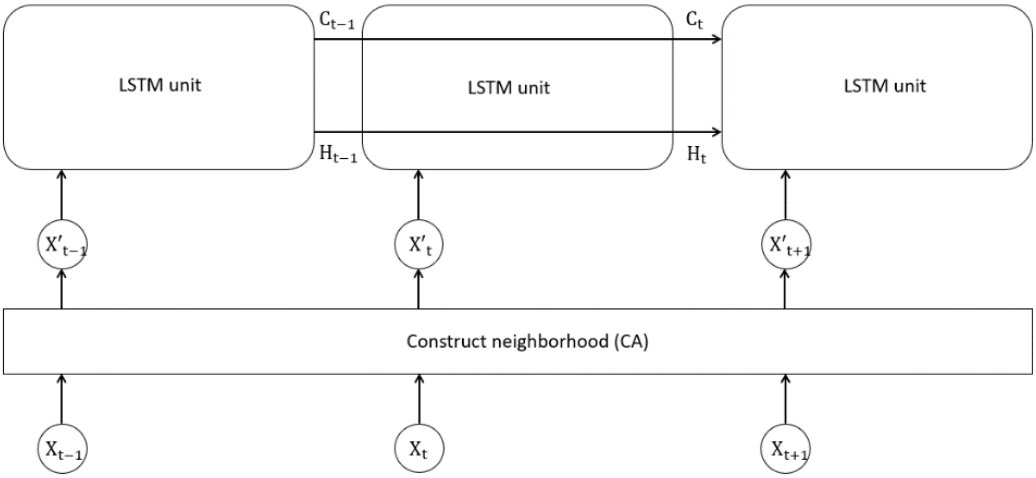

Figure 5. Schematic diagram of integrated LSTM and CA model.

Another consideration of this model integration is whether the model itself is with sufficient scalability and universality. The structural requirements of the integrated model for the input data are three-dimensional tensors with the form of sample, timestep, and feature, where the timestep corresponds to the information on the time series, and the feature corresponds to the definition of the neighborhood in the CA. The training of the integrated model is data-driven, and the information in the temporal and spatial dimensions is implicit in the input data. As no additional controls are required besides the adjustment of input data, the integrated model is with strong scalability and generalizability. In addition, given that internal parameters are with acceptable initialization settings, manual adjustments are not required, which demonstrates the robustness and automaticity of the proposed model.

3.3. Spatial Extension

The spatial extension of the model in this paper is achieved via the perspective of neighborhood construction. In the experiment, we select three different neighborhoods: the special neighborhood of the central cell, the Moore neighborhood (R = 1), and the extended Moore neighborhood (R = 1) by extracting cell weights based on road networks.

The first neighborhood construction is relatively simple, that is, the choice of the central cell while ignoring its neighborhoods. The Moore neighborhood (R = 1) is a classic method of defining neighborhood, as shown in Figure 6.

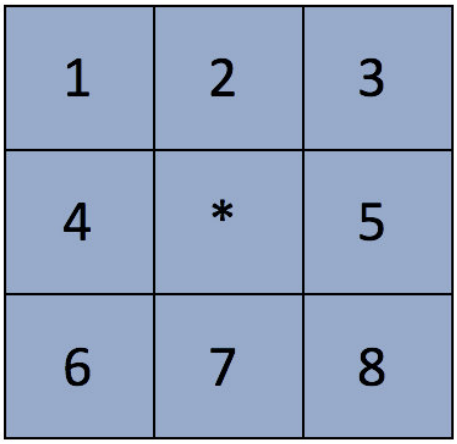

Figure 6. Moore neighborhood (R = 1).

The combination of the Moore neighborhood (R = 1) and algorithm for extracting cell weights based on road networks is another way to build neighborhoods. This paper draws on the idea of "decay neighborhood with distance" [24–26] from the previous studies and uses data of road networks to ensure that proper spatial information is considered. Figure 7 presents the road networks in Chongming District.

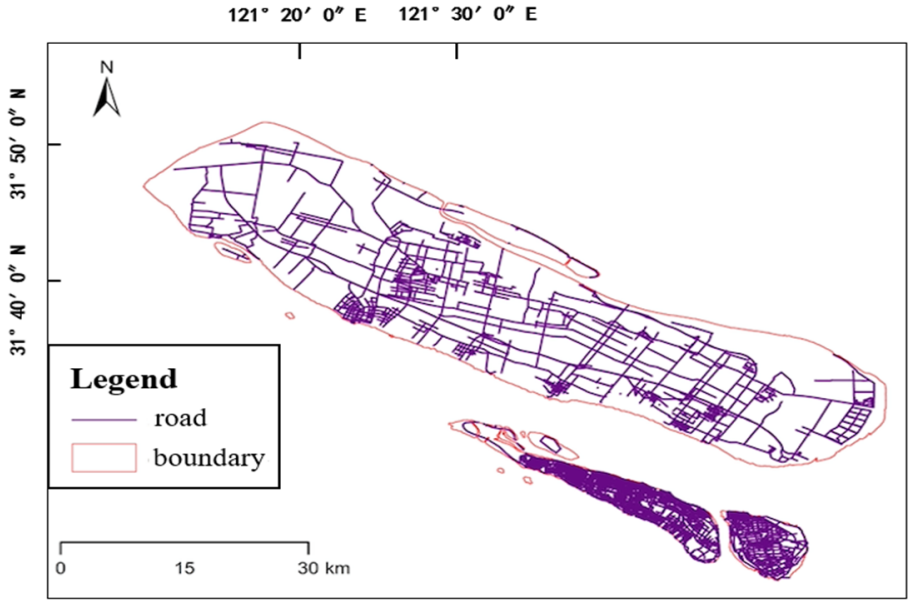

Figure 7. Road networks in Chongming District.

The method of constructing neighborhoods from road networks is based on the assumption that population movements are affected by surrounding areas of interest (such as roads, tourist areas, etc.). With the consideration of road elements, people are with a tendency to move towards the road in a random state, as the cells closer to the road have a greater impact on the central cell. To ensure that the total impact weight of all cells is 1, the weights are normalized via the following formula:

$$W_i = \frac{e^{-D_i}}{e^{\sum_{j=1}^{9} -D_j}} \quad (7)$$

where W_i represents the weight of the i-th cell, and D_i represents the distance of the i-th cell from the nearest road. Given the distance to the nearest road, the weights are first calculated considering an exponential decaying effect, then normalized after the summation.

With the use of the Moore neighborhood (R = 1), road data are integrated into the model, as cell weights of the neighborhood are calculated based on the road network. Such information affects the extraction of spatial features during model training.

4. Experiments and Results

In the first experiment, we selected the central cell without considering its neighborhood. Since the period of the data covers the Chinese National Day (1 October) and ordinary dates, the training was conducted separately to investigate the heterogeneity in the temporal pattern of population changes. In addition, the minimum time granularity of the data (10 min) and different time granularities (i.e., 1 h) were used to show the impact on the generalization effect of the different models. We also considered combining data from multiple days to reveal the longer-term characteristics in the time series. An additional experiment was conducted by applying extended neighborhoods that aim to include spatial information in the model. Figure 8 presents the specific training method.

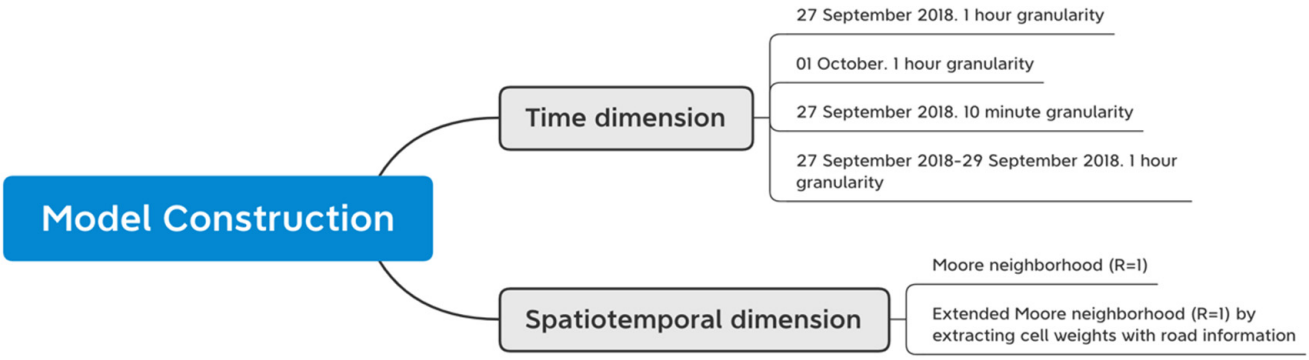

Figure 8. Roadmap of experiments.

In total, we trained six models, in which the first four models were constructed based on a single temporal dimension, and the other two were composite models that integrated spatiotemporal perspectives. For purposes of evaluating the generalization effect of the model, we used two evaluation methods: an absolute evaluation and a relative evaluation. The absolute evaluation applies the mean absolute error (MAE) as the evaluation metric:

$$J(y_i, f(x_i;\theta)) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i;\theta)| \quad (8)$$

where y_i is the true value, θ is the training weights, and $f(x_i;\theta)$ is the model output.

Relative evaluation is a baseline assessment. Before training the model, an ideal model based on common sense as a reasonable reference point was established. After comparing the trained model with the baseline model, the relative quality of the model can be judged. Since our research problem targets time series population prediction, the simplest model assumption is that the predicted results at the current time step in the test set are the same as the value at the same time in the training set. In this study, the division of training set and testing set is unique, given the fact that our model has spatial and temporal dependencies. Therefore, unlike the traditional method of splitting the dataset in a fixed ratio, we train with data for a period of time and predict with data for another period of time. All implementations in this study are completed via Python programming language, specifically using TensorFlow and Keras libraries. In this study, the hyperparameters for all models are the same. We use RELU as the activation function with the Adam algorithm as the optimizer (default setting).

4.1. Time Dimension

Neural network training requires a balance between optimization and generalization. One of the focuses of this study is to improve the generalizability of the model, as the application of population prediction models requires a certain degree of robustness. Since neural networks have a strong ability in data fitting, the problem of overfitting often occurs during training. We added an L2 regular term to the loss function when designing the model to solve this problem. Another possible solution is to perform manual adjustment by regulating the number of learning parameters. The number of learning parameters determines the number of features that a model can learn, and the larger the number, the more complex the model. A small number of parameters can lead to underfitting, while a high number can lead to overfitting, which greatly affects the model's generalization ability. Therefore, the adjustment of hyperparameters in studies should primarily target this parameter. Despite that, there are many hyperparameters in the deep learning framework (e.g., batch size and epoch); only the number of learning parameters remained adjustable in this experiment with other hyperparameters fixed to ensure comparability in this study.

Four models were trained, respectively, using 1 h granularity data on 27 September 2018 (Model 1), 1 h granularity data on 1 October 2018 (Model 2), 10 min granularity

data on 27 September 2018 (Model 3), and 1 h granularity data on 27–29 September 2018 (Model 4). The numbers of best learning parameters are, respectively, set as 10, 10, 18, and 22. The baseline errors are 1.245, 1.636, 1.245, and 1.871, respectively. The generalization performance on the test set is shown in Tables 2–5. Figure 9 presents the generalization error plot of the four models on the test set. Baseline error is calculated from baseline assessment, and the test error is derived from the prediction of the trained model.

Table 2. Errors on the entire test set of the 10-parameter model by 1 h granularity on 27 September 2018 (Model 1).

Date	Baseline Error	Test Error
28 September 2018	1.245	1.161
29 September 2018	1.459	1.122
30 September 2018	2.31	1.161
1 October 2018	6.141	1.174
2 October 2018	6.587	1.229
3 October 2018	6.73	1.27
4 October 2018	6.626	1.314
5 October 2018	6.022	1.131
6 October 2018	5.478	1.187
7 October 2018	5.289	1.371

Table 3. Errors on the entire test set of the 10-parameter model by 1 h granularity on 1 October 2018 (Model 2).

Date	Baseline Error	Test Error
27 September 2018	6.141	1.216
28 September 2018	6.104	1.202
29 September 2018	5.901	1.168
30 September 2018	5.177	1.176
2 October 2018	1.636	1.286
3 October 2018	1.949	1.334
4 October 2018	2.076	1.369
5 October 2018	1.97	1.138
6 October 2018	2.393	1.102
7 October 2018	3.107	1.21

Table 4. Errors on the entire test set of the 18-parameter model by 10 min granularity 27 September 2018 (Model 3).

Date	Baseline Error	Test Error
28 September 2018	1.245	0.902
29 September 2018	1.459	0.997
30 September 2018	2.31	1.214
1 October 2018	6.141	2.034
2 October 2018	6.587	2.215
3 October 2018	6.73	2.42
4 October 2018	6.626	2.359
5 October 2018	6.022	2.136
6 October 2018	5.478	2.133
7 October 2018	5.289	1.92

Table 5. Errors on the entire test set of the 22-parameter model by 1 h granularity during 29 September 2018–27 September 2019 (Model 4).

Date	Baseline Error	Test Error
30 September 2018	1.871	0.963
1 October 2018	5.901	0.956
2 October 2018	6.325	0.992
3 October 2018	6.439	1.04
4 October 2018	6.351	1.134
5 October 2018	5.749	0.982
6 October 2018	5.273	0.919
7 October 2018	5.166	0.961

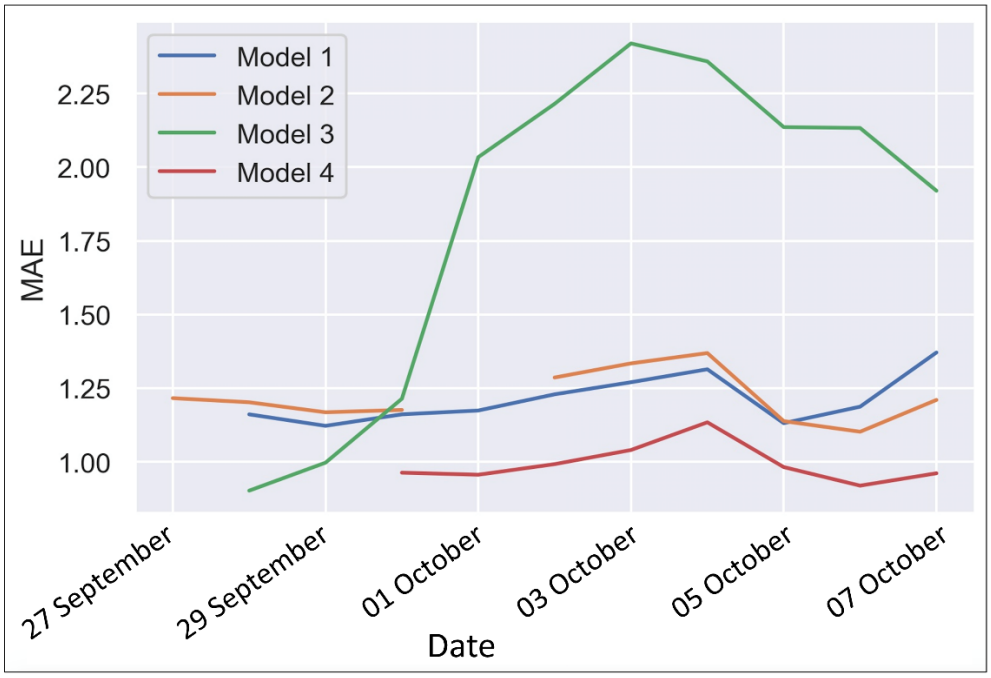

Figure 9. The plot of daily test errors of model-1, 2, 3, 4.

According to Figure 9, it is observed that with an increase in the amount of data and the decrease in time granularity, the number of optimal learning parameters of the model increases, suggesting that the model needs more parameters to express such characteristics in the time dimension. When comparing Model 1 and Model 2, we found that there exists a certain periodic feature in the single-day population data, and the expression of this feature on the normal date and the National Day is different. Despite such discrepancies, the similarity of changes in population dynamics during one day is notable. We also observed that this rule performs better on data of ordinary dates. When comparing Model 1 and Model 3, we found that a small time granularity does not necessarily perform better. The generalization error of the data at 1 h granularity is lower than that of the data at 10 min. As in certain occasions, specific characteristics exist under a specific granularity. Sometimes, data with a fine time granularity may lose key features. In addition, even random data can express certain correlations, as some scholars have found that some correlations can be found even for data with complete noises [21]. When comparing Model 1 and Model 4, we found that the multi-day data generally contain richer information and have hidden features on the long-term sequence. Given its great performance, the subsequent model training is based on Model 4.

4.2. Spatiotemporal Dimension

The above four models only consider a single time dimension and do not adopt the concept of neighborhood. In this experiment, we incorporate spatial attribute information into the model from two perspectives: (1) the classic Moore neighborhood (R = 1) (Model 5), and (2) the extended Moore neighborhood (R = 1) combined with extracting cell weights based on road networks (Model 6). The best learning parameters are set as 30 and 35, respectively, and the baseline errors are both 1.871. The generalization errors on the test set are shown in Tables 6 and 7. Figure 10 presents the errors for all six models.

Table 6. Errors on the entire test set of the 30-parameter model by 1 h granularity and Moore neighborhood 29 September 2018–27 September 2019 (Model 5).

Date	Baseline Error	Test Error
30 September 2018	1.871	0.922
1 October 2018	5.901	0.935
2 October 2018	6.325	0.96
3 October 2018	6.439	0.978
4 October 2018	6.351	1.06
5 October 2018	5.749	0.906
6 October 2018	5.273	0.845
7 October 2018	5.166	0.931

Table 7. Errors on the entire test set of a 35-parameter model by 1 h granularity and extended neighborhood 29 September 2018–27 September 2019 (Model 6).

Date	Baseline Error	Test Error
30 September 2018	1.871	0.874
1 October 2018	5.901	0.899
2 October 2018	6.325	0.909
3 October 2018	6.439	0.947
4 October 2018	6.351	1.048
5 October 2018	5.749	0.882
6 October 2018	5.273	0.813
7 October 2018	5.166	0.867

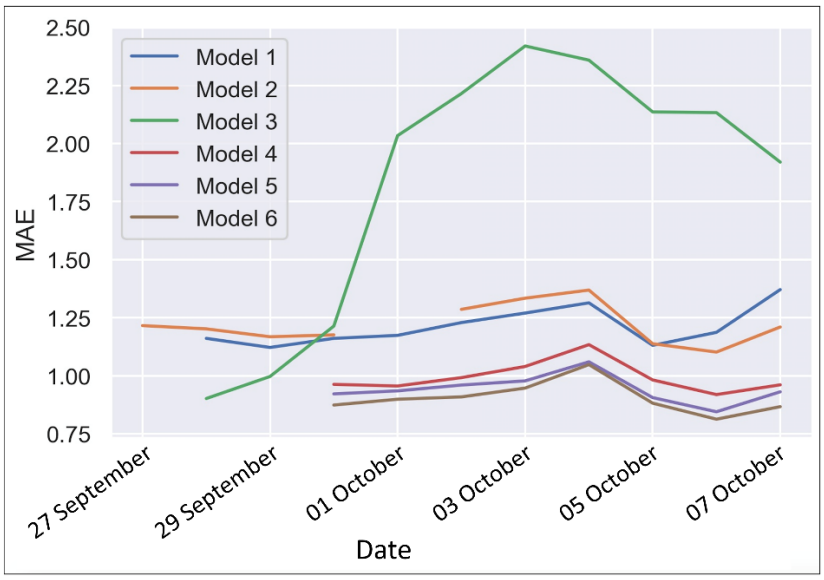

Figure 10. The plot of daily test errors of model-1, 2, 3, 4, 5, 6.

We observed that the quality of the models is further improved with the consideration of spatial information. The spatial assumption of the Moore neighborhood is spatially contiguous, which limits model performance to a certain degree. In comparison, Model 6 fuses road network information, allowing more detailed spatial information to be introduced into the model, thereby further improving the quality of the model. However, the introduction of road network information leads to increased model complexity.

5. Discussion

Population distribution prediction has always been a hot research topic in geography and other spatial science-related fields. Better population distribution knowledge is expected to benefit a wide range of domains that include smart cities [27], regional planning [28,29], transportation management [29], and disaster mitigation [30], to list a few. After reviewing traditional population distribution prediction models, we notice that traditional methods, largely relying on the structured census, tend to have coarse spatiotemporal granularity. In recent years, the advances in information handing techniques, the introduction of data processing approaches, and improved computational power led to the popularity of population distribution prediction with fine spatiotemporal granularity. Entering the Big Data era, scholars start to resort to machine learning and automated algorithms, aiming to address the challenges in fine-scale population modeling studies.

LSTM and CA are popular approaches that, respectively, address temporal and spatial problems. Given its design, LSTM is featured by its capability in predicting events with long intervals and delays in time series, making it an ideal algorithm for population distribution prediction. On the other hand, CA has the capability of simulating spatiotemporal evolutions. The unique features from LSTM and CA motivate us to think about their potential integration, with CA as the upper-level framework and LSTM as the underlying algorithm to derive the conversion rules. Taking advantage of the mobile phone signaling data collected by China Unicom, we test the proposed CA-LSTM-integrated model in predicting the population distribution in Chongming District, Shanghai by setting different time granularity, timesteps, and neighboring scenarios. The results suggested that models with finer time granularity do not guarantee better performance, as data separated by finer granularity may lose key features, posing challenges for the models when they are making the prediction. We also notice that an increase in data put benefits the accuracy of the prediction, which is expected because deep learning models rely on data to establish a robust mapping between the input and the output. As for neighboring scenarios, we notice that models that consider spatial information outperform models that do not, which proves the important role of spatial dependency in population distribution prediction. In addition, we observe that models that consider spatial information characterized by road networks outperform models that consider spatial information characterized by the Moore neighborhood (spatially contiguous). This result points to the fact that population dynamics are more intertwined with road network settings than spatial closeness. Further efforts can be made to involve other spatial scenarios in population distribution prediction, aiming to better capture the factors that drive population movement.

Unfortunately, the mobile phone signaling data we have only lasted 14 days, which, to some degree, limits our investigation. In addition, cellphone counts only capture a certain spectrum of the population, while people who do not have access to mobile devices or who do not carry devices outside are underrepresented. In our future work, we plan to explore the capability of the proposed model in handling a longer time period and include different time granularities. In addition, the performance of different granularities needs to be tested with more case studies. We also intend to integrate more geographic information to investigate the potential improvement of the model's performance.

6. Conclusions

Although there exist some efforts that integrate LSTM and CA models, their models lack the ability to integrate multi-source geographic information and, to our best knowl-

edge, no effort has been made to investigate such integration in predicting the spatial distribution of population. Taking mobile phone signaling data as input and Chongming District of China as the study case, we proposed a model that integrates LSTM and CA to extract population dynamics from spatiotemporal dimensions. The integrated model is constructed in an unsupervised manner with great scalability and the capability of fusing multi-source geographic data. We also applied different time granularities and varying methods of spatial neighborhoods.

Despite the great performance of the proposed model, challenges still remain. In this case study, we observed that the generalization abilities of trained models differ greatly, indicating that the training methods should be dependent on different research data types and scenarios. That is to say, different training data can lead to different best parameter settings. The internal parameters of the model should be adjusted appropriately in order to achieve optimal performance. The best model obtained in this paper is based on the training data of the extended Moore neighborhood (R = 1), combined with three-day ordinary dates, one-hour granularity, and extracting cell weights from road networks. The calculated testing error (MAE) for all cells is 0.905. The generality and robustness of the integrated model can be further improved by including more data sources. Benefiting from the proposed method of neighborhood construction, other data sources can be easily incorporated into our proposed model.

Author Contributions: Conceptualization, Pengyuan Wang and Xiang Li; methodology, Pengyuan Wang and Dong Xu; formal analysis, Pengyuan Wang and Di Zhang; investigation, Joseph Mango; resources, Xiao Huang; writing—original draft preparation, Pengyuan Wang; writing—review and editing, Xiang Li and Xiao Huang; funding acquisition, Xiang Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Grant/Award Number: 41771410; Ministry of Education of China, Grant/Award Number: 19JZD023.

Data Availability Statement: The dataset and codes in this study can be found at https://figshare.com/s/ea7fcdff4e977135a20c (accessed on 25 July 2021).

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

References

1. Wu, S.S.; Qiu, X.; Wang, L. Population estimation methods in GIS and remote sensing: A review. *GIScience Remote Sens.* **2005**, *42*, 80–96. [CrossRef]
2. Nusteling, H.P. The population of England, 1539–1873: An issue of demographic homeostasis. *Hist. Mes.* **1993**, *8*, 59–92. [CrossRef]
3. Glass, D.; Appleman, P. Thomas Robert Malthus: An Essay on the Principle of Population. *Popul. Stud.* **1976**, *30*, 369. [CrossRef]
4. Leslie, P.H. On the use of matrices in certain population mathematics. *Biometrika* **1945**, *33*, 183–212. [CrossRef] [PubMed]
5. Zakria, M.; Muhammad, F. Forecasting the population of Pakistan using ARIMA models. *Pak. J. Agric. Sci.* **2009**, *46*, 214–223.
6. Lutz, W.; Sanderson, W.; Scherbov, S. Doubling of world population unlikely. *Nature* **1997**, *387*, 803–805. [CrossRef]
7. Stoto, M.A. The accuracy of population projections. *J. Am. Stat. Assoc.* **1983**, *78*, 13–20. [CrossRef]
8. Carter, L.; Lee, R. Modeling and forecasting US sex differentials in mortality. *Int. J. Forecast.* **1992**, *8*, 393–411. [CrossRef]
9. Lutz, W.; Sanderson, W.; Scherbov, S. Expert-Based Probabilistic Population Projections. *Popul. Dev. Rev.* **1998**, *24*, 139. [CrossRef]
10. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Prabhat Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
11. Robinson, C.; Hohman, F.; Dilkina, B. A deep learning approach for population estimation from satellite imagery. In Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, Redondo Beach, CA, USA, 7–10 November 2017; pp. 47–54.
12. Rumelhart, D.; Hinton, G.; Williams, R. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
13. Griffith, C.S.; Swanson, D.A.; Knight, M. DOMICILE 1.0: An agent-based simulation model for population estimates at the domicile level. In *Opportunities and Challenges for Applied Demography in the 21st Century*; Springer: Dordrecht, The Netherlands, 2012; pp. 345–370.
14. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
15. Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 855–868. [CrossRef]

16. Zeyer, A.; Doetsch, P.; Voigtlaender, P.; Schlüter, R.; Ney, H. A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2462–2466.
17. White, R.; Engelen, G. Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to the Evolution of Urban Land-Use Patterns. *Environ. Plan. A Econ. Space* **1993**, *25*, 1175–1199. [CrossRef]
18. Itami, R. Simulating spatial dynamics: Cellular automata theory. *Landsc. Urban Plan.* **1994**, *30*, 27–47. [CrossRef]
19. Barlovic, R.; Santen, L.; Schadschneider, A.; Schreckenberg, M. Metastable states in cellular automata for traffic flow. *Eur. Phys. J. B* **1998**, *5*, 793–800. [CrossRef]
20. Esser, J.; Schreckenberg, M. Microscopic Simulation of Urban Traffic Based on Cellular Automata. *Int. J. Mod. Phys. C* **1997**, *8*, 1025–1036. [CrossRef]
21. Devi, N.S.S.S.N.U.; Mohan, R. Long Short-Term Memory with Cellular Automata (LSTMCA) for Stock Value Prediction. In *Data Engineering and Communication Technology*; Springer: Singapore, 2020; pp. 841–848.
22. Liu, L.; Sun, X.-K. Volcanic Ash Cloud Diffusion From Remote Sensing Image Using LSTM-CA Method. *IEEE Access* **2020**, *8*, 54681–54690. [CrossRef]
23. Liu, S.; Xue, D.; Gao, J. The Assessment of Island-Type Urban Ecosystem Based on the Environmental Carrying Capacity Model—A Case Study of Shanghai's ChongMing Island. *Adv. Mater. Res.* **2012**, *573–574*, 325–330. [CrossRef]
24. Padró-Martínez, L.T.; Patton, A.; Trull, J.B.; Zamore, W.; Brugge, D.; Durant, J.L. Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmos. Environ.* **2012**, *61*, 253–264. [CrossRef]
25. Requia, W.J.; Roig, H.L.; Adams, M.D.; Zanobetti, A.; Koutrakis, P. Mapping distance-decay of cardiorespiratory disease risk related to neighborhood environments. *Environ. Res.* **2016**, *151*, 203–215. [CrossRef] [PubMed]
26. Freedman, L.; Pee, D. Return to a Note on Screening Regression Equations. *Am. Stat.* **1989**, *43*, 279.
27. Garcia-Retuerta, D.; Chamoso, P.; Hernández, G.; Guzmán, A.S.; Yigitcanlar, T.; Corchado, J.M. An efficient management platform for developing smart cities: Solution for real-time and future crowd detection. *Electronics* **2021**, *10*, 765. [CrossRef]
28. Langford, M.; Higgs, G.; Radcliffe, J.; White, S. Urban population distribution models and service accessibility estimation. *Comput. Environ. Urban. Syst.* **2008**, *32*, 66–80. [CrossRef]
29. Martín, Y.; Li, Z.; Ge, Y.; Huang, X. Introducing Twitter DAILY estimates of residents and NON-RESIDENTS at the county level. *Soc. Sci.* **2021**, *10*, 227. [CrossRef]
30. Huang, X.; Wang, C.; Li, Z.; Ning, H. A 100 m POPULATION grid in the Conus by DISAGGREGATING census data with Open-source Microsoft building footprints. *Big Earth Data* **2020**, *5*, 112–133. [CrossRef]