Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 7, No. 2, 2023

Population Spatialization Based on Convolutional Neural Network and Multi-Source Geospatial Big Data

Xusong Zhang[1, *], Maria Rosario Rodavia[2]

[1] Graduate School, Angeles University Foundation, Angeles 2009, Philippines
[2] Graduate School, Angeles University Foundation, Angeles 2009, Philippines
[*] Corresponding author: Xusong Zhang (Email: zhang.xusong@auf.edu.ph)

Abstract: This article adopts a more refined method to create a population spatial distribution model for Guizhou Province in 2021. Firstly, we divide the population data of Guizhou Province into different characteristic regions. Then, we use these regions as independent variables and population as dependent variables for analysis using convolutional neural network (CNN). In addition, we also introduced multi-source spatiotemporal data such as Points of Interest (POI), nighttime lighting, and Digital Elevation Model (DEM) for modeling to further enhance the accuracy and accuracy of the model. Through this method, we successfully generated the population spatial data of Guizhou Province in 2021 with a resolution of 1km. These data intuitively display the spatial distribution of the population in Guizhou Province, and have important reference value for policy makers, researchers, and planners.In addition, in order to verify the accuracy of our model, we evaluated and compared the accuracy of the model results. The results show that our model has high accuracy in expressing the spatial distribution of population, andcan accurately reflect the spatial distribution of population at different scales.In summary, this article successfully created a population spatial distribution model for Guizhou Province in 2021 using multiple spatiotemporal big data and multiple linear statistical regression methods, and evaluated and compared the accuracy of the model results. These results have important practical significance for understanding the population distribution, formulating land use planning and public policies in Guizhou Province.

Keywords: Population distribution, spatialization, Convolutional neural network (CNN), Guizhou Province, POI.

1. Introduction

The spatialization of demographic data can break regional restrictions and achieve the transformation of population data from administrative regions to regular grid forms, thus simulating the actual population distribution and reproducing the actual population distribution. It is of great significance for solving the problem of coupling natural resources, environment, and population, and formulating national macroeconomic decisions. Based on different population distribution influencing factors and various auxiliary data, various spatialization methods for demographic data have been developed, mainly including spatial interpolation, multi-source data fusion, remote sensing estimation, and land use modeling methods. Among them, the spatial pattern of land use/coverage is closely related to the spatial distribution of population. Based on the relationship between land use types and population distribution, the method of establishing convolutional neural network models for simulating population spatial distribution is the most widely used. On this basis, in order to reflect the differences in population spatial distribution within the same land use type, some scholars analyze the differences in population distribution characteristics of the same land use type in different geographical locations and classify land use data to improve the accuracy of the original model; Some scholars use nighttime lighting data or other auxiliary data to reclassify or extract features from land use data. The spatialization of population statistics data can break regional restrictions and achieve the transformation of population data from administrative regions to regular grid forms of population spatial data. This simulates the actual population distribution and reproduces the actual population distribution, which is beneficial for solving the coupling problem between natural resources, environment, and population The formulation of national macroeconomic decisions is of great significance. Based on different population distribution influencing factors and various auxiliary data, various spatialization methods for demographic data have been developed.

2. Research on Population Simulation Based on Convolutional Neural Network

2.1. Convolutional Neural Network

Convolutional neural network, abbreviated as CNN in English, is a way of simulating brain networks to process information through multiple neurons and convolutional kernels. It is specifically designed to process large-scale images or sensory data in a multi array format by considering local and global stationary characteristics. The structure of convolutional neural networks can be roughly divided into three layers[2]. Input layer, hidden layer, output layer. The following focuses on the hidden layer structure:

(1) Convolutional layer

In a convolutional layer, each neuron in that layer performs convolution operations with a weighted and biased convolutional kernel. Due to the use of convolutional extraction, the convolutional kernel has the function of local extraction, which can fully extract each local feature in image and text data, as well as the spatial position relationship between local features. In addition, all neurons belonging to a certain feature in the convolutional layer share the same weighted link. The biggest advantage of convolutional neural networks is that the weights of convolutional kernels are

shared during local feature extraction, which reduces the complexity of the computational process. Assuming Xij represents the overall value in the grid cells (i, j) of the input image[3]., the k-th output feature map can be calculated using the following formula:

$$y_{ijk}^{l} = w_{k}^{l} * x_{ij}^{l} + b_{k}^{l} \quad (1)$$

Where l represents the current layer and k represents the depth of the convolutional kernel. w_k^l and b_k^l represent the weights and deviations of the k-th feature in the current layer.

(2) Activation function

The activation function is a function with a nonlinear relationship, typically including the sigmoid function, Tanh function, and ReLU function. Due to the fact that the characteristics of the sample may be linear or nonlinear, the output in the convolutional layer is the result of matrix multiplication, which has a strong linear relationship. Therefore[4]., it is necessary to perform nonlinear operations on the neurons in the matrix, so that the output neurons have nonlinear relationships and the neural network has the ability to describe nonlinear relationships. The ReLU function is as follows:

$$f(x) = \max(0, x) \quad (2)$$

(3) Pooling layer

In the pooling layer, the maximum value of adjacent grid cell groups in the same block in the convolutional layer is extracted, and the output is downsampled to extract features. This method, known as maximum pooling technology, is widely used in convolutional and neural networks because of its superior feature extraction advantages. The purpose of the pooling layer is to reduce the spatial size of the input overall feature map and avoid overfitting problems caused by over sampling of the model.

$$f(x^{l}) = \max(x_{i}^{l}) \quad (3)$$

$f(x^l)$ represents the maximum pooling result, and $\max(x_i^l)$ is the maximum value of the neurons in the I-th unit of the current layer.

(4) Fully connected layer

The fully connected layer is composed of convolutional and pooling layers in the form of a connected weight matrix to extract features and obtain multiple one-dimensional feature points. Its function is to integrate the previously extracted local feature information, transform one-dimensional feature vectors, and use classification functions for classification or regression functions for fitting.

2.2. Overview of the study area

Guizhou is a province in China, abbreviated as "Qian" or "Gui", located in the hinterland of the southwestern inland region of the People's Republic of China. The terrain of Guizhou is high in the west and low in the east, tilting from the middle to the north, east, and south, known as the "eight mountains, one water, and one field". Guizhou has a subtropical monsoon climate, covering an area of 176200 square kilometers. The provincial capital is Guiyang City. At the end of 2022, the permanent population of Guizhou Province was 38.56 million. Guizhou Province has a total of 50 counties, 11 autonomous counties, 10 county-level cities, 16 municipal districts, 1 special zone, and a total of 88 county-level administrative regions. The province has 4 prefecture level cities, 3 autonomous prefectures, 2 regions, and a total of 9 prefecture level administrative regions. In addition, there are 122 townships, 192 ethnic townships, 831 towns, and 364 streets, totaling 1509 township level administrative regions. Guizhou is a transportation hub in the southwest region and an important component of the Yangtze River Economic Belt. Guizhou is the first national level big data comprehensive experimental zone in China, a world-renowned mountain tourism destination and major mountain tourism province, as well as a national ecological civilization experimental zone and an inland open economy experimental zone. In addition, the historical representative culture of Guizhou is the "Guizhou Guizhou Culture", which is also the birthplace of ancient Chinese humanity and one of the birthplaces of ancient Chinese culture. In 2022, the regional GDP of Guizhou Province was 2016.458 billion yuan, an increase of 1.2% compared to the same period last year at constant prices. Among them, the added value of the primary industry was 286.118 billion yuan, an increase of 3.6%; The added value of the secondary industry is 705.703 billion yuan. (Figure 1.)

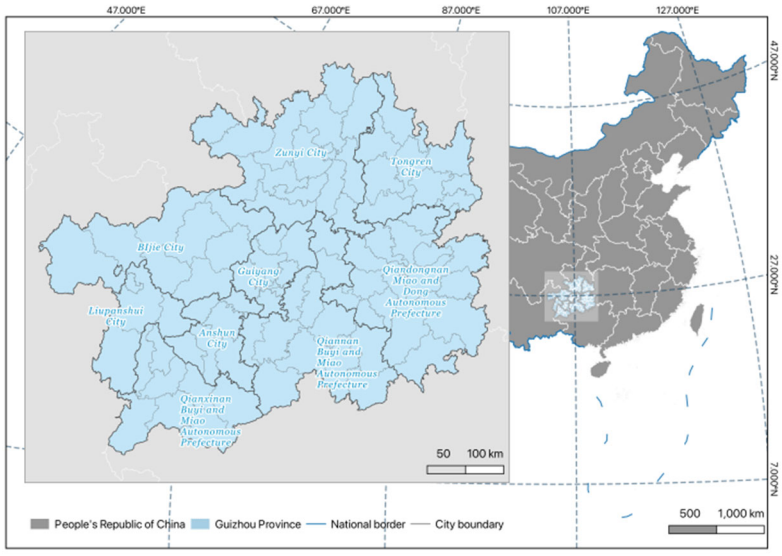

Figure 1. Guizhou Map Area

2.3. Data preprocessing

The basic data used in this article mainly includes demographic data, points of interest (POI), administrative division data, land use data, DMSP/OLS night light data, DEM and basic geographic information data such as rivers and slopes. Data processing mainly includes data integration and proofreading, statistical and spatial data matching, as well as projection transformation and resampling. The population of each prefecture level city is obtained by merging all districts under its jurisdiction. All raster data are cropped to Anhui Province and sampled to 1 km using the nearest neighbor resampling method × 1 km resolution data. This article also divides and processes the differences within urban residential areas based on nighttime lighting data. The light intensity values of DMSP/OLS nighttime light data have an indicative effect on the spatial distribution of population, and can be used as a classification standard to reclassify urban residential areas. The reclassified data is then used for spatialization of population data based on land use, and the accuracy of the results is significantly improved. The second classification of urban land with high population density . Considering the different levels of ecological environment and economic development in different regions, as well as significant differences in population distribution, using the same spatialized model for population spatial modeling may have limited accuracy. Therefore, this article conducts a regional population characteristic consistency zoning in advance before population spatial modeling.

3. Experimentation

Convolutional Neural Networks (CNN) can be used for both classification algorithms and regression algorithms. One of its advantages is that the upper and lower layers of the network share weight parameters, thereby accelerating the computational efficiency of the model. On this basis, we believe that Convolutional Neural Networks (CNN) can provide a more effective method for simulating population siz[5].e. This article uses the convolutional network idea to regress and fit the population and population factors of statistical data. The specific process is shown in the Figure 2.

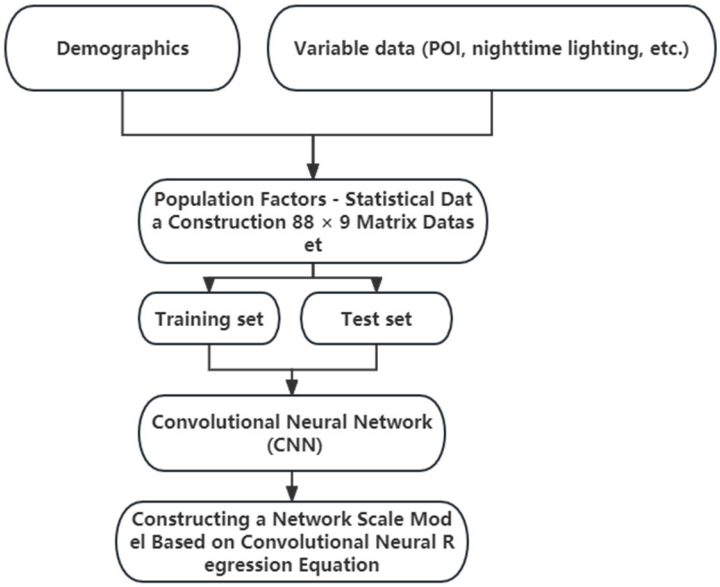

Figure 2. Population spatialization process based on Convolutional neural networks

It is divided into three steps:

(1) In the data preprocessing stage, the nighttime lighting data, social media data, terrain data, and land cover data are preprocessed to extract 1km of corresponding population distribution influencing factors × 1km grid value and average population at township level. This article takes the average value of 9 indicators at the township level as the independent variable and the population density at the township level as the dependent variable, and constructs a matrix size of 88 × Construct a population spatialization model based on convolutional neural networks using a dataset of 9. Data preprocessing calculation requires normalizing each population impact factor, that is, changing the data of each population impact factor from a positive integer to a value of 0-1. Finally, the training set and validation set are divided into input models in a 7:3 ratio.

(2) Parameter Setting and Model Training Building a Convolutional Neural Network (CNN) model by calling TensorFlow library on the Python language platform. Construct convolutional layer, pooling laye[6].r, and fully connected layer by setting reasonable parameters in modules. In order to construct the matrix reasonably, the influence factors of the original data were increased to 16 columns and cut into 4 × 4 matrix, then the input size is 4 × Matrix of 4. The first layer convolutional network adopts a matrix size of 2 × 2 neurons with 32 convolutional kernels, input to the activation function ReLU after passing through the convolutional layer, and the pooling layer is 2 × 2 matrices, with a quantity of 32. The second layer convolutional network adopts 2 × 2 neurons, 64 convolutional kernels, set ReLU as the activation function, and the pooling layer is also 2 × 2 matrices, with a quantity of 64. The final number of hidden nodes in the fully connected layer is 512. Using the Adam optimizer to adjust and optimize the model, the learning efficiency is set to 0.01 to approach the lowest point of the loss function. When the training frequency is set to 5000 times, the model is optimal.(Figure 3.)

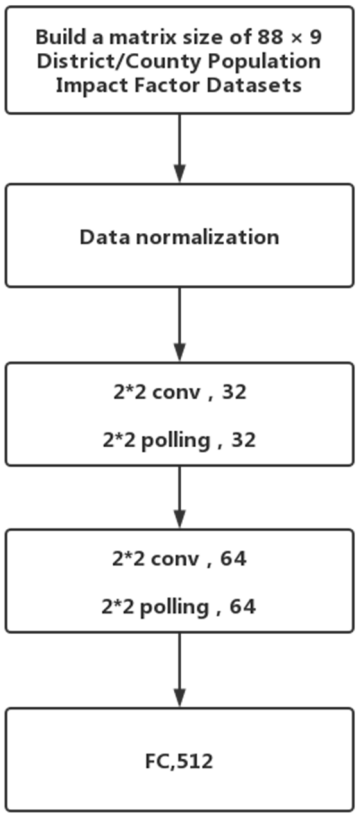

Figure 3. Convolutional neural network structure diagram

After extensive computer operation, the accuracy of this model is 95%. Then apply the trained model to grid population simulation, input the grid values of each distribution influence factor, and output the population at the corresponding grid positions. Thus achieving population spatialization.

4. Conclusion

The commonly used accuracy verification method for convolutional neural network models (CNN) is to evaluate the fit of statistical models by comparing the predicted values of 30% of the predicted data with the initial values. Unlike commonly used precision analysis, the precision analysis method of population spatialization mainly includes summarizing the population density of the subdivision grid onto the corresponding minimum administrative unit, and comparing the population on the minimum administrative unit generated by the experimental summary. This article substitutes grid feature data into a trained convolutional neural network (CNN) model to obtain the population density of each grid. Then, the population data of each grid unit is summarized in the corresponding townships, and compared with the population of the seventh population census of Guizhou Province in 2020 to output the fitting degree. Most of the predicted data values in this experiment are distributed near the trend line, with a goodness of fit of R^2=0.852, indicating that the prediction accuracy of this experiment is relatively high. The Convolutional Neural Network (CNN) model can accurately reflect the characteristics of population distribution in urban and rural areas, and better reflect the distribution patterns of urban and rural populations. Using census, survey, satellite, and mobile data to generate a consistent grid with high accuracy can distinguish the shape of rivers, with more obvious graininess, higher fragmentation[7]., more prominent textures, and more detailed information, which is more in line with the actual population distribution in complex surface environments. The simulation results of convolutional neural network models (CNN) have high accuracy and rich details.

References

[1] A Methodology for Deep Learning and its Application to Convolutional Neural Networks for Sentiment Classification", 2016, by J. Devlin, M.W. Chang, K. Lee and K.T. Wu..

[2] Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", 2015, by G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger.

[3] Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", 2015, by S. Ioffe and C. Szegedy.

[4] ImageNet Large Scale Visual Recognition Challenge", 2015, by V. Voulodimos, N. Doulamis, N. Doulamis and A. Antonopoulos.

[5] Object Detection with Countless Objects", 2014, by P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun.

[6] Li, Y., & Chen, Y. (2021). Random Forest: A Literature Review. Journal of Big Data Science and Engineering, 6(3), 67-79.

[7] Wang, Y., He, Q., & Wu, J. (2019). Predicting the Spatial Distribution of Population Using Remote Sensing Images and Machine Learning Algorithms: A Case Study in China. Applied Sciences, 9(17), 3867.