MJRET

OPEN ACCESS

Multidisciplinary Journal of Research in Engineering and Technology

www.mjret.in

eISSN:2348-6953

SOFTWARE SYSTEM FOR POPULATION PREDICTION

Harshit Handa, Sahil Sharma, Shivam Prajapati, Prof. Swetha S

Department of Information Science and Engineering
R V College of Engineering
Bengaluru, India

Abstract: *Human population growth rate is an important parameter for real world planning. It is dependent on various demographic factors such as population infant mortality rate, life expectancy at birth and total fertility rate. Several livability factors such as the number of education institutes and health facilities also play an important role in shaping the population. To predict the population of Singapore these factors are taken into account and best performing algorithms based on previous works that are linear regression, K Nearest Neighbour (KNN), random forest classifier and ensemble learning techniques are used to train models. These models can then be used through an interface where users can enter either (i) the year whose population is to be predicted or (ii) the values of the factors taken. In the first case, the values of the factors are predicted for that year and then the population based on those predicted factors, whereas in the second case the population is predicted based on the input factor values directly. The interface presents the information in graphical form for better understanding.*

Keywords*: Population, Demographic factors, Livability factors, Multiple Linear Regression, Random Forest, K – Nearest Neighbour Regression, XGBoost, Bagging, Boosting, Stacking*.

1. INTRODUCTION

Population is the collection of members of a species in a particular geographical area. Demographic information can be quantified using population statistics, which can also be used to evaluate the connections between ecosystems, human health, and infrastructure. Numerous creatures' distribution, development, and migration can all be described in terms of population. When referring to people, urbanisation, immigration, and population demographics are frequently brought up in conversation about population. For each component, multiple evolution assumptions are used to create various scenarios.

[1] Population prediction is crucial because it aids people in making future decisions, such as government researchers. People may use the outcome of population prediction in a variety of ways, including.

1. Calculate the demand for fundamental human needs such as food, water, power, and transportation.
2. Create plans for building projects like homes, roads, etc.
3. Calculate the labour force sizes in different locations.
4. Calculate the potential consumption in different areas.

[2] For long-term thinking, especially in terms of collective development, the predictions provide a solid foundation. Total predictions are estimates produced for the entire nation. However, they are referred to as regional or sectoral predictions when they are created for an area, state, province, district, or ethnic group. Comparatively speaking to regional estimates, total projections are simpler. This is due to the difficulty of obtaining reliable historical data for a region on internal migration, birth and death rates, and other demographics. Population predictions are based on a number of assumptions about migration, birth rate, and death rate. The projected population is high if it is anticipated that the birth rate is high, the death rate is low, the immigration rate is high, and the emigration rate is low. This estimate applies to less developed nations that are now going through a demographic shift.

It is characterised as a medium projection of population if it is believed that there would be a medium increase in immigration and emigration rates, as well as a medium increase in birth and death rates. Due to the effectiveness of family planning and health services, this prediction indicates a moderate increase in the growth rate of the population. Such projections are helpful in nations that are rapidly developing. It is a difficult challenge to predict the trends in the human population. The population of each nation is subject to a number of uncertainties. In the past, countries' demographics have been predicted using statistical methods. However, these methods are not particularly effective in forecasting a chaotic system like the population. First, the techniques make various assumptions that are occasionally thought to be unfounded. Second, statistical population forecasting techniques are unable to handle the inherent unpredictability. Additionally, it can be challenging to predict population increase due to certain occurrences that can drastically change a location's demographic profile in a short amount of time. Such as migration, birth rate, and death rate. A population can change quickly as a result of migration, particularly when it is brought on by life-altering events like war. In this instance, the demographic makeup of one area is altered as the population of another area grows. Using historical data and making certain assumptions, some mathematical models, such as the geometrical and arithmetical rise models, may predict the population of either new

or existing cities over the course of a decade. Although they are universally acknowledged, their applicability is constrained. It is essential to improve these models or switch to a more efficient model in order to obtain a more accurate result at the lowest possible cost.

2. LITERATURE SURVEY

With the aid of various time series forecasting machine learning algorithms, including Linear regression, Support Vector Regression, Multilayer Perceptron, and Decision Tree Classifier, the paper [3] examined the growth of the Indian population using official population data. The I instances in datasets that were subjected to an analysis were those whose inclusion or exclusion had no impact on the effectiveness of the procedures. In order to predict the rise of the Indian population, linear regression performed better than the other classifier. Areas without historical data are not addressed by existing methodologies, and the absence of the feature is a problem that is frequently encountered. A study in paper [4] examines 17 machine learning methods, including base learners and ensemble learners, in estimating the nation's population growth rate. Among all the other strategies, random forest did the best. Currently, projections of the population over multiple regions can be made, mostly using the Interregional Cohort-Component model. The strategy to anticipate multi-regional population growth using machine learning is essentially what is proposed in [5]. They created a machine learning technique that, by efficiently utilizing the benefit of the XGBoost algorithm, predicts and analyses the population growth of Taiwan's major cities.

In a deep learning model, the research [6] suggested leveraging satellite photos to produce high-resolution population estimates. Using 1-year composite Landsat images, they specifically trained convolutional neural networks for population prediction in the USA at a 0.01° 0.01° resolution grid. Worldwide, it is noted that a large number of people choose not to participate in their nation's census. In order to reliably anticipate the population density of a region, the authors of study [7] provided two Convolutional Neural Network (CNN) architectures that effectively and efficiently incorporate satellite image inputs from various sources. They made use of population labels from the 2011 SECC census and satellite pictures of rural Indian villages.

The authors of research [8] compared the global estimates of the sizes of urban areas for the years 2000 and 2010 using high-resolution pictures of nighttime illumination. They used recently-proposed approaches to address issues with the nighttime luminosity data that are currently accessible, such as blurring, pixel stability over time, and a decreased ability to compare night light images between satellites and across time. Paper [9] reviews a systematic methodology for DMSP/OLS (nighttime light based) urban extent mapping, focusing on four

aspects: brightness saturation, blooming impact, intercalibration of time series, and correction of their temporal pattern. It has long been common practice to estimate population distribution using nighttime light (NTL). Its use in predicting population density has been constrained by the overglow effect of the images brought on by reflection of light from nearby locations and the varied population distribution patterns between urban and rural areas. In order to lessen the overglow effect and represent urban and rural population densities separately, a strategy was suggested in paper [10].

3. METHODOLOGY

Data Gathering and Processing

Reference number [14] is used to download population data of Singapore of the past several years. The data collected from multiple sources is then integrated and subsequently cleaned by removing irrelevant information. Data imputation is done by filling the missing values in a column by the mean of all the values in that column and by using the interpolation of data frame technique with each column.

Model Building and Training

After the data cleaning and data imputing process we divide our dataset into two parts. 80% of the dataset is used to train our ML model using various algorithms which include mainly linear regression, K nearest neighbour, Random forest and various ensemble learning methods like Boosting, Stacking and Bagging and the remaining 20% of the dataset is used for testing the created model. The input to our model are the liveability defining attributes such as number of schools, colleges, hospitals, entertainment zones and many more, demographic features such as mortality rate, birth rate etc and the year for which the prediction is required. The output is the predicted population for that year.

Software System

In this we are having two methods for the prediction purpose, In method 1 of prediction, the input to our ML model are various liveability defining facilities. On the basis of these inputs, the ML model will run the regression and predict the approximate value of population. On the other hand in the 2nd method, only the year for which the population is to be predicted is entered. Using regression, the values of individual parameters will be predicted for the given year first and then these predicted values are given as an input to the method 1 of prediction. The system also asks the user to choose the algorithm to use for the prediction purpose. Hence the ML model corresponding to the selected algorithm is used for estimating. The user can also compare the estimations of the algorithm with the real values in order to learn about the accuracy of the selected algorithm.

The comparison of algorithms' performance, trends of population, liveability attributes and demographic features over the years are displayed graphically using various graphs and visual elements. The terminal is used to provide the input to the system and the output is also displayed on the terminal in a well-defined and clean format.

4. RESULTS

The multiple algorithms used for the prediction produced different accuracies when tested upon the available dataset. The basic approaches used are multiple linear regression that attained an accuracy of 79.8%, K nearest neighbour with an accuracy of 90.49% and Random forest regression with an accuracy of 99.75% when tested upon a subset of the dataset.

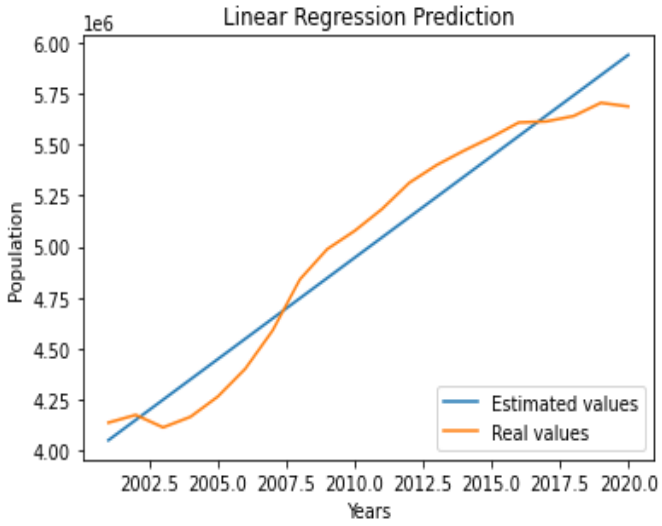

Fig.1: Linear regression prediction vs real values

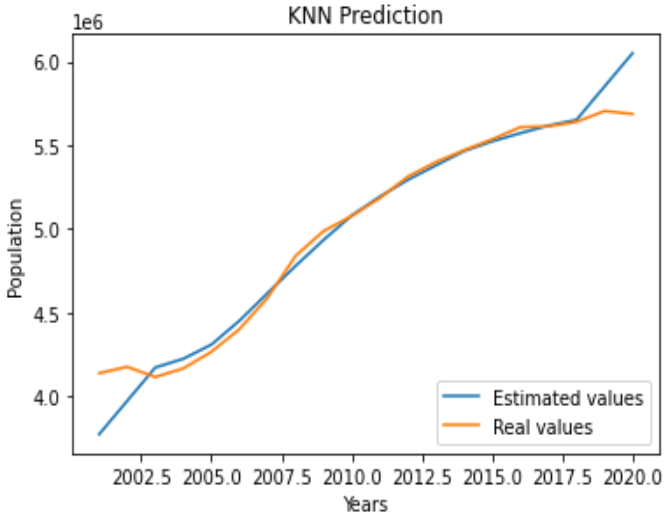

Fig.2: KNN prediction vs real values

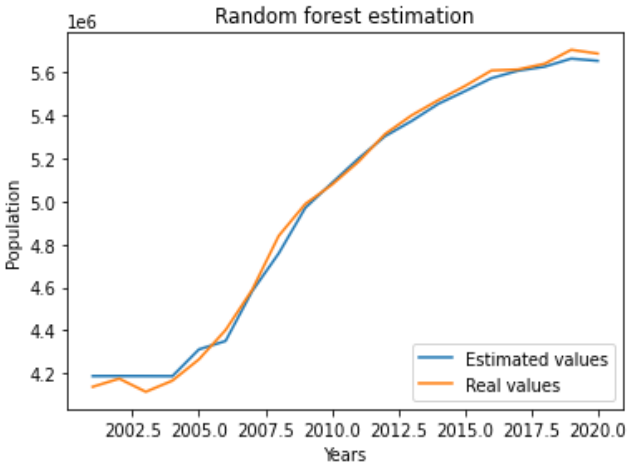

Fig.3: Random forest prediction vs real values

Along with the basic techniques, to improve the performance of the system some ensemble learning techniques were used like XGBoost that attained a peak accuracy of 88.63%, Stacking with an accuracy of 99.49%, Bagging with an accuracy of 98.98% and Boosting with an accuracy of 99.39%. A comparative study of accuracies is shown in table 1. Population vs Years graphs were plotted for all the algorithms used. Figure 1 shows such a graph with a predicted population using multiple linear regression curve and an actual population curve. Figure 2 shows this comparison with algorithm used as KNN, Figure 3 shows this comparison with algorithm used as random forest, Figure 4 shows this comparison with algorithm used as stacking, Figure 5 shows this comparison with algorithm used as bagging, Figure 6 shows this comparison with algorithm used as boosting and Figure 7 shows this comparison with algorithm used as XGBoost. Figure 8 shows the comparison between the estimations of all the algorithms for the test data. A comparison of the population prediction for future years from 2020 to 2050 is shown for all the algorithms in Figure 9,

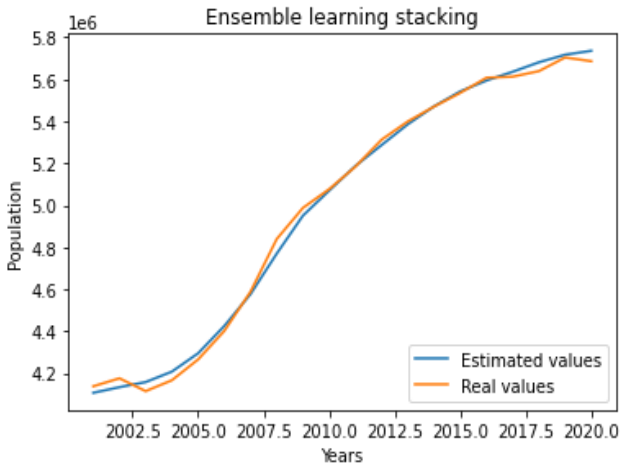

Fig.4: Stacking prediction vs real values

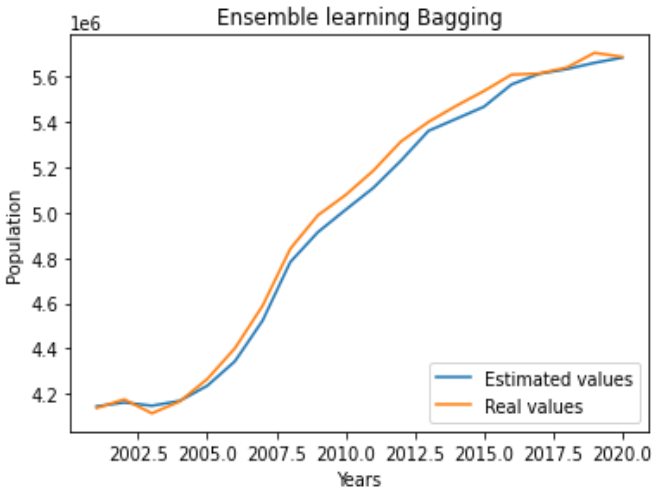

Fig.5: Bagging prediction vs real values

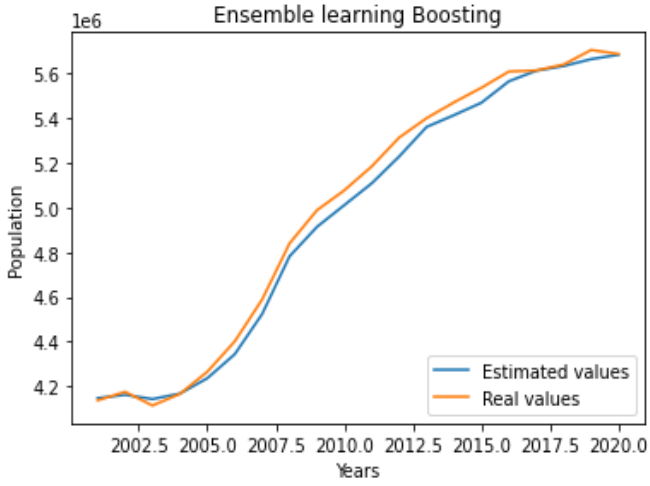

Fig.6: Boosting prediction vs real values

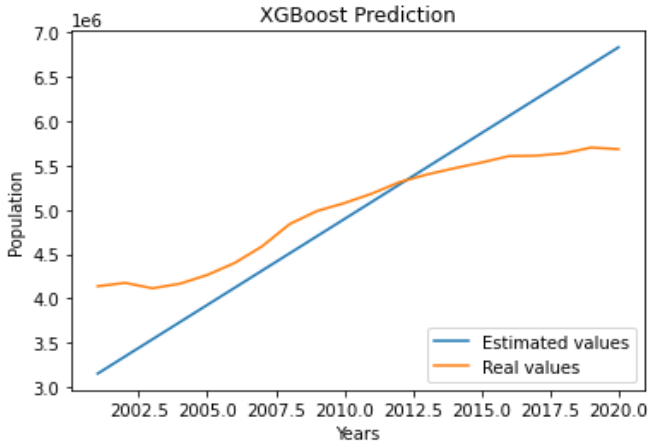

Fig.7: XGBoost prediction vs real values

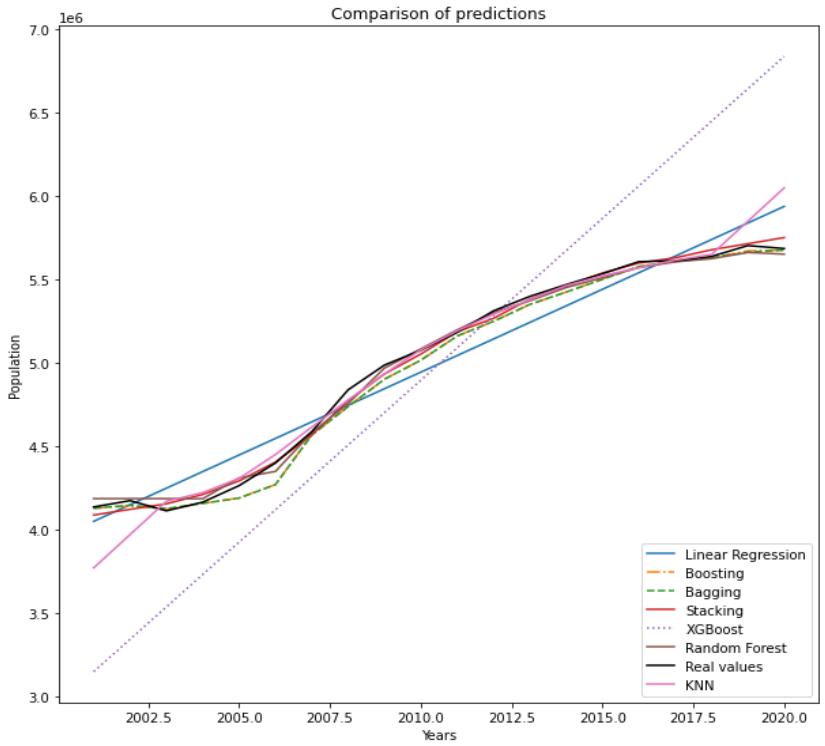

Fig.8: All algorithms' prediction vs real values

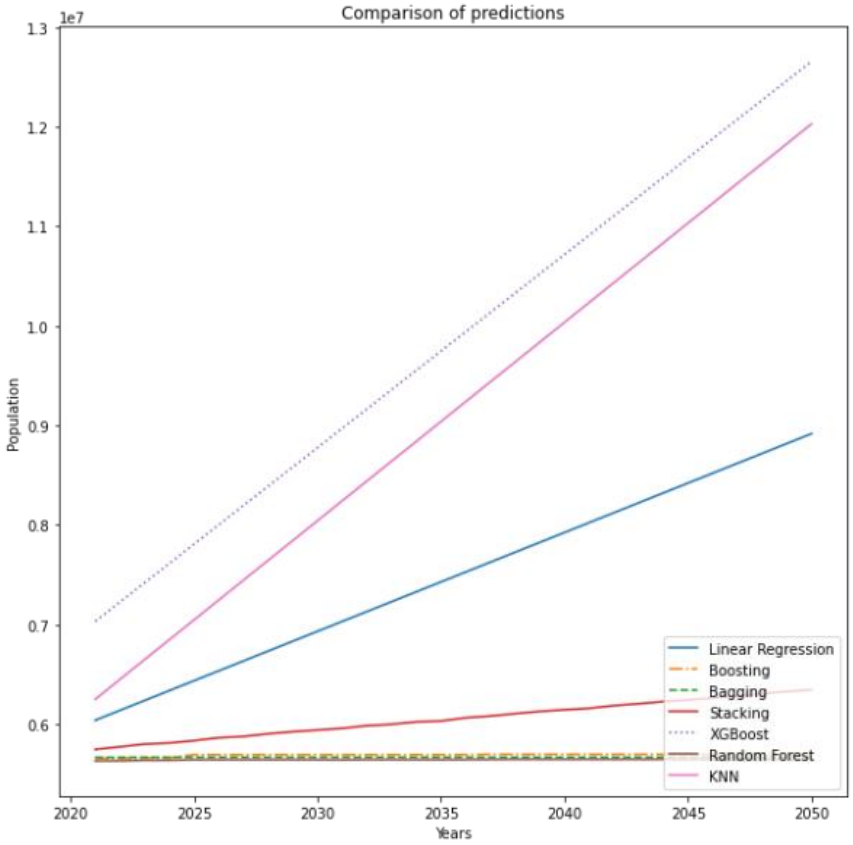

Fig.9: Comparison of predictions by all the algorithms for future years

Table 1 *Comparison of accuracies of various algorithms*

Algorithm	**Accuracy**
Multiple Linear Regression	79.8%
K Nearest Neighbour	90.49%
Random Forest	99.75%
XGBoost	88.63%
Stacking	99.49%
Bagging	99.98%
Boosting	99.39%

Out of all the algorithms used, the random forest regression produced the best results on the testing data. All the ensemble learning algorithms obtained excellent accuracies, each one of them having the accuracy above 99%. As it is evident from the figures, the difference in the real and the predicted values is very less in the case of all ensemble learning techniques. Figure 8 represents a comparison of the predictions of all the algorithms compared with the real values.

5. CONCLUSION & FUTURE SCOPE

By analyzing the performances of various machine learning algorithms, it was observed that all the algorithms performed well with similar results while different methods of ensemble learning like boosting, stacking and bagging achieved the best results. The random forest regression resulted with the best accuracy and its graph was the most similar to the actual graph obtained by plotting the dataset points. All the tried algorithms gave different graphical trends of population for the future years. With time as the actual population values are made public, it can be known which algorithm estimated best. For now a rough estimate of the range of population for a particular year can be made using the algorithm giving maximum and the one giving minimum population. Many of the machine learning forecasts have performed well, which is encouraging. However, further research is necessary before they may legitimately be applied in reality. The majority of the time, machine learning techniques provide predictions that are opaque to end users. It will need more effort to create and test machine learning

techniques that have been shown to be effective across numerous small area demographic datasets.

REFERENCES

[1] *Population projection [Online] available at: https://www.sociologydiscussion.com/demography/population-projections/population-projections-meaning-types-and-importance/*

[2] *Importance of population projection [Online] available at: : https://www.projectrhea.org/rhea/index.php/Importance_of_population_projection*

[3] *Brintha Rajakumari S, Padmanabhan P, Christy S, Nandhini M, "Prediction Of Population Growth Using Machine Learning Techniques", European Journal of Molecular & Clinical Medicine, ISSN 2515-8260 Volume 7, Issue 5, 2020*

[4] *Mohammad Mahmood Otoom, "Comparing the Performance of 17 Machine Learning Models in Predicting Human Population Growth of Countries", IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.1, January 2021*

[5] *Chian-Yue Wang and Shin-Jye Lee, "Regional Population Forecast and Analysis Based on Machine Learning Strategy", https://doi.org/10.3390/e23060656, 24 May 2021*

[6] *Caleb Robinson, Fred Hohman, Bistra Dilkina, "A Deep Learning Approach for Population Estimation from Satellite Imagery", arXiv:1708.09086v1 [cs.AI] 30 Aug 2017*

[7] *Wenjie Hu, Jay Harshadbhai Patel, Zoe-Alanah Robert, Paul Novosad, Samuel Asher, Zhongyi Tang, Marshall Burke, David Lobell, Stefano Ermon, "Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery", arXiv:1905.02196v1 [cs.CV] 4 May 2019*

[8] *Rafael Ch, Diego A. Martin, Juan F. Vargas, "Measuring the Size and Growth of Cities Using Nighttime Light", CAF Development Bank of Latin America, Working paper No. 2018/14, September 2018*

[9] *Xuecao Li and Yuyu Zhou, "Urban mapping using DMSP/OLS stable night-time light: a review", International Journal of Remote Sensing, 2017 http://dx.doi.org/10.1080/01431161.2016.1274451*

[10] *Weichao Sun, Nan Wang, Yi Cen, "Estimating Population Density Using DMSP-OLS Night-Time Imagery and Land Cover Data", IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, Vol. 10, No. 6, June 2017, DOI: 10.1109/JSTARS.2017.2703878*

[11] *Xi Chen, "Nighttime Lights and Population Migration: Revisiting Classic Demographic Perspectives with an Analysis of Recent European Data", Remote Sens. 2020, 12, 169; doi:10.3390/rs12010169*

[12] *Paul C. Sutton, Christopher D. Elvidge, "Relationships between Nighttime Imagery and Population Density for Hong Kong", Proceedings of the Asia-Pacific Advanced Network 2011 v. 31, p. 79-90., http://dx.doi.org/10.7125/APAN.31.9, ISSN 2227-3026*

[13] *Bin Li, Tianfei Wang, Liping Jia, "Application of Improved Logarithm Logistic Models in Population Prediction", 2012 Eighth International Conference on Computational Intelligence and Security, DOI 10.1109/CIS.2012.30*

[14] *(2022) Census dataset of Singapore [Online] available at: https://tablebuilder.singstat.gov.sg/*