Python-based Population Forecasting with Standard Deviation Analysis

Jiewen Zheng *

NingboHuaMao International School, Ningbo, China

* Corresponding Author Email: Jacob.zheng@nbhis.com

Abstract. Changes in the world's population are a constant problem, such as the aging crescent, declining fertility and so on. Having an accurate forecast can chart a better and more useful plan for future development. The accuracy of the python prediction is obtained by computing the standard deviation of the python prediction and the raw data. The code runs within an extremely basic set of read, judge, and compute rules. This data set has been obtained from more than 10, 000 data sets of populations between 1950 and 2021 in different age conditions and countries. Each individual age segment gets a different shift during this period, especially for segments with large historical events, so it will be a challenge to see if Python can or cannot produce accurate predictions. In the experiments, python was used to predict the future population and the standard deviation was obtained by comparing the predicted results with the source data.

Keywords: Standard deviation, Population, Accuracy.

1. Introduction

The world population is constantly a central issue for the governments of different countries. The ageing of the population, the fall in fertility, is now giving the present population equilibrium a huge jolt. The UN's prediction of 10.9 billion by 2100 is based, at least in part, on "the unprecedented ageing of the world's population", as well as "rapid population growth driven by elevated fertility" in some countries and regions. The world should expect to see considerably more grey hairs by 2050: by then, the number of people worldwide aged 65 or over is expected to be more than double the number of children under five, and about the same as the number of under-age children. Through reasonable population forecasting, we can understand the population size and age structure of a certain period in the future, which helps the authorities to understand the future labor resources of the country, do a good job in labor distribution and balance, allocate social resources with maximum efficiency, and formulate more reasonable social welfare, culture, education and health, urban development and construction planning. A virtuous circle for society as a whole. Python has a wealth of methods for making predictions, and by using data that exists in reality, we can understand the standard deviation of the predicted data and the true figure. In summary, the population is not well predicted by search and using Python and finding out its accuracy can be used as a reference to make better plans for the future.

2. Related work

'A modular neural network-based population prediction strategy for evolutionary dynamic multi-objective optimization' written by Li Sanyi, Yang Shengxiang, Wang Yanfeng, Yue Weichao, Qiao Junfei shows a novel population prediction algorithm based on modular neural network (PA-MNN) for handling dynamic multi-objective optimization (DMOP) [1].

'Detection and prediction of land use changes and population dynamics in the Gorganrud River basin, Iran' by Diba Ghonchepour, Amir Sadoddin, Abdolrassoul Salmanmahiny, Abdolreza Bahremand, Anthony Jakeman, Barry Croke. This paper illustrates that analysis of changes in land use and land cover is a fundamental tool for assessing the impact of human activities on the environment..The aim of this paper is to detect and predict possible land use change in the Gorgan

Rud River basin in Iran, and to estimate the role of past and future population growth as drivers of land use change and degradation [2].

Written by Kilburn KH, Thornton JC and Hanscom B. In this paper, researchers argue that brain function testing is conceptually similar to lung function testing..The authors used these tests to evaluate 293 adults from three unexposed groups in different regions of the United States..Subjects were contacted randomly from the voter registration rolls and compensated for their time..The tests included balance, reaction time, strength, hearing, visual performance and cognitive recall, perceptual motor and memory functions..The regression equations simulate the performance of each test and the effect of demographic factors..The resulting prediction equations will help researchers conduct quantitative tests on chemically exposed individuals and others with brain damage [3].

Written by B Justin, Echouffo-Tcheugui, J Stephen, Greene, Lampros, Papadimitriou, Faiez, Zannad, W Clyde and Yancy..The main thrust of this paper is to effectively screen high-risk patients and implement appropriate, cost-effective preventive interventions, otherwise the prevalence of heart failure is expected to rise significantly. A systematic review of the predictive properties of heart failure risk prediction models published up to August 2014 was performed using the MEDLINE and EMBASE databases. Eligible studies report the development, validation, or impact assessment of the model. Two researchers conducted an independent review to extract data on study design and characteristics, risk predictors, model differentiation, calibration and reclassification capabilities, and validation and impact analysis, and used these to reach the goals [4].

Written by Li Dong, Yu Yanyan, Wang Bo. This paper focuses on how to use multi-objective lioness optimization calculations and system dynamics models to predict urban populations. In the case of Xi'a, the paper simulates three scenarios consisting of five policy factors: fertility, employment, science and technology, healthcare and education. Their impact on the future population is derived, population size is projected for 2019-2050 and it is concluded that the employment policy and the fertility policy are the two most effective policies to promote population growth [5].

Written by Zhang Haiming and Xi Xiaoli. In this paper, we focus on predicting the trend of Guangzhou's preschool population. In this paper, the seventh census data of Guangzhou City and statistical yearbook data over the past years are used and put into the international population software PADIS-INT to predict the change trend of the preschool population aged 0-6 years old in Guangzhou from 2021 to 2035 under nine schemes with different net migration levels and fertility levels. The corresponding schemes and suggestions for this case are also given in the following paper [6].

Written by Pang Mengyin, WANG Haining, Wan Tongming, Ma Miao. This paper focuses on how predictive models can be used to predict the number of confirmed cases. Based on the prediction results of the Logistic model and the long short-term memory deep learning network model, it selects the cumulative number of confirmed cases over a certain period of time to train the linear combination parameters and obtain the final combined prediction model. Finally, the predictive performance of the proposed model is compared with that of Logistic, LSTM and SEIR models using RMSE and other predictive performance evaluation metrics. The results calculated by MAPE outperform other models and can provide technical support for subsequent epidemic prediction and prevention and control [7].

Written by Li Yanru, Li Meng and Liu Shuang. In this paper, the development trend of Shanghai's birth population is predicted and its influencing factors are mainly written. The GM (1,1) gray prediction model is used to forecast the population. With the help of the results, it is analyzed that the number of newborns in Shanghai will drop off the cliff in the next ten years and there will be an obvious phenomenon of aging of fewer children [8].

Written by Ma Xuesong. In this paper, we mainly use the Leslie model and the factor regression model to predict the future population of China. In the process, the grey GM (1, 1) model is also used to predict the sex ratio of birth population. The Pearson correlation coefficient, the gray correlation and the XGBoost machine learning algorithm determine the main influence factors, which are then modeled using regression analysis. Finally, it is fed into the trained BP neural network model, which then outputs the predicted values for the Chinese male and urban populations [9].

3. Methodology

Standard deviation, in mathematical terms, is the arithmetic square root of the arithmetic mean of the squared difference from the mean, expressed as σ. Standard deviation, also known as standard deviation or experimental standard deviation, is most commonly used in probability statistics as a measure of the extent of a statistical distribution. Standard deviation is the arithmetic square root of the variance. Standard deviation reflects the dispersion of a data set. Two sets of data with the same mean may not have the same standard deviation. In the formula below x_i means to each of the values of the data, $\bar{x}$ means the mean of x_i, and n means the number of data points.

Standard deviation is in form of:

$$Standard\ deviation = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad (1)$$

NumPy is the basic package for scientific computing using Python. It is a powerful n-dimensional array object. There are complex (broadcast) functions. Tools for integrating C/C ++ and Fortran code. It is useful for linear algebra; Fourier transform and random number functions.

Pandas is Python's core data analysis support library, providing fast, flexible, and unambiguous data structures designed for simple and intuitive handling of relational and labeled data. Pandas' main data structures are Series and DataFrame, which are sufficient to handle most typical use cases in finance, statistics, social sciences, engineering, and more.

Matplotlib is a Python drawing library that makes it easy for users to graph data and provides a variety of output formats. It can be used to draw a variety of static, dynamic, interactive charts. It is a very powerful Python drawing tool, and we can use this tool to present a lot of data more intuitively through the form of charts.

Seaborn is a graphical visual python package based on matplotlib. It provides a highly interactive interface that makes it easy for users to create a variety of attractive statistical charts. Seaborn is a more advanced API encapsulation based on matplotlib, which makes drawing easier.

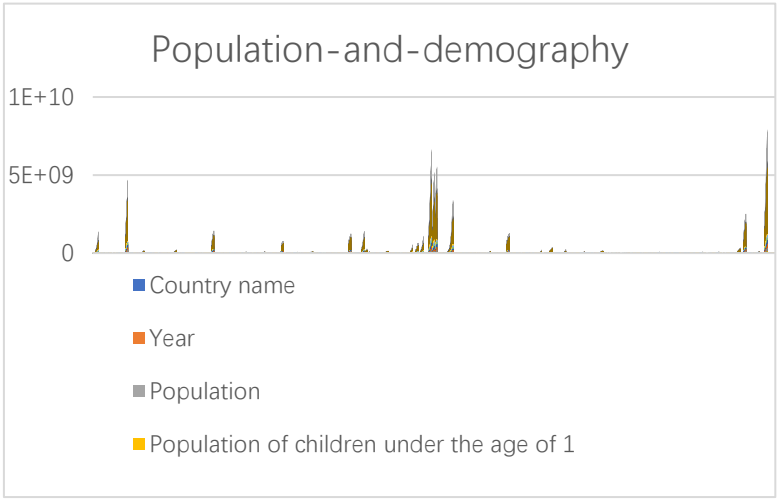


Figure 1. A figure from Age Structure - Our World in Data with data of population in different age trend and country.

Fig 1 is the data set that used in the code, the data have shown Country name, Year, Population, Population of children under the age of 1, Population of children under the age of 5, Population of children under the age of 15, Population under the age of 25, Population aged 15 to 64 years, Population older than 15 year, Population older than 18 years, Population at age 1, Population aged 1 to 4 year, 12 Population aged 5 to 9 years, Population aged 10 to 14 years and etc... Panda has since collected this data and analyzed it into:

The data after pandas disposal is the original data without any movement or changes, and after simply confirming that the data has been successfully imported, Python needs to read the data for further analysis. This project analyzes the data using information and description functions.

The functions described show the count, maximum and minimum of the values in the object, the mean of the values and the standard deviation of the observations. The info function prints a short summary of the DataFrame, showing information about the DataFrame, including the index data type dtype and column data type, the number of non-null values, and memory usage. These two data analyses can give python a hint of what the dataset is about in a clearer and deeper format. Then use a code to change the list that shows country name into orderly number, and create a categorical to show the original country name, like the country name in the data set had been replaced by number, and the code create a new data set to put the country name. The Lambda function is then used in the code. This code makes all countries or some huge cities into a list of numbers starting with 0 and ending with 253, and the result has become part of the original data replacing the ordered numbers.

Lambda: (A Lambda is an anonymous inline function that does not require a name (that is, an identifier) and consists of a single expression that is evaluated when called. It consists of the keyword Lambda, arguments, and function bodies.

The last part of the code was using data in the new graph that have just made, to calculate the standard deviation based on pythons' prediction and data already exist. For this calculation Linear regression had been used, this model fitted to minimize the sum of squares of residuals between the actual target values in the data set and the targets predicted by linear approximation. Next, the data organized in the past few steps need to be used.

Linear regression: (Statistical analysis, which uses regression analysis in mathematical statistics to determine quantitative relationships between two or more variables that are dependent on each other, is widely used. In regression analysis, only one independent variable and one dependent variable are included, and the relationship between them can be approximated by a straight line. This regression analysis is called unitary linear regression analysis. A regression analysis is called multilinear if it includes two or more independent variables and there is a linear relationship between the dependent and independent variables.)

The methodology section gives ideas and a brief introduction to how and what this project is about.

4. Experiment

0.9999708017086113 is the result obtained by the code. In a report, He Chun, an associate professor at Guangdong University of Technology, wrote about using a Malthusian population model to predict Guangdong's future population and give a plan for population growth. She collected data on Guangzhou's total population, natural growth rate, birth rate, and death rate from 1975 to 2010. Also, the total number of men and women. She combined these with a Malthusian population model to predict the population between 2005 and 2014, and then used the known data from 2005 to 2010 to get the wrong numbers. The same steps are then performed again with the Logistic population growth model. The results of these two models are compared in the conclusion, and the 5-year forecast results with errors are favorable, but this forecast needs to be corrected according to the actual situation [7]. The prediction method she used was different from mine, because I made the prediction by Python program, while she made the prediction by mathematical formula. In contrast, I did not import, analyze and calculate additional data other than population data that could affect the population, so the resulting standard deviations are not minimal.

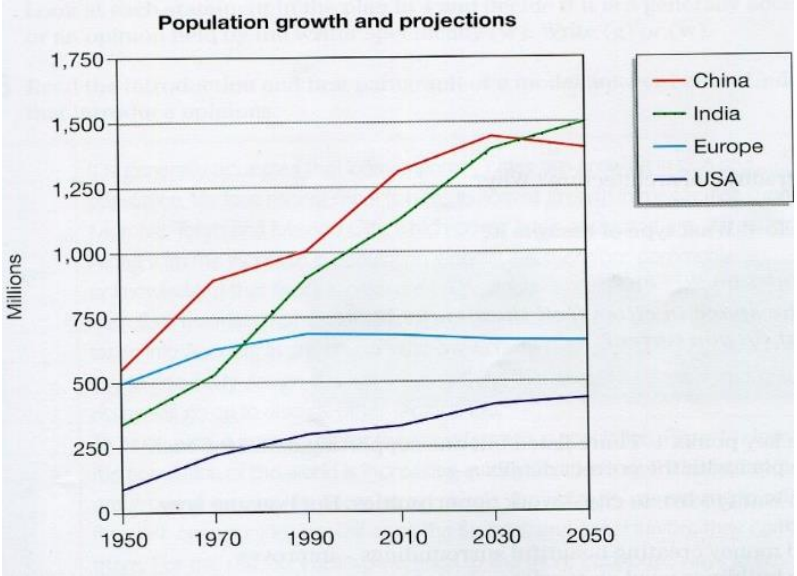


Figure 2. Picture from http://fastlearn.edu.vn/2019/11/25/line-chart-3-population-growth-and-projection/, about Population growth and projection.

Malthusian population model

(Malthus surveyed the demographic information of England for more than a hundred years.

Based on the assumption of a constant population growth rate, the well-known exponential population growth model is formulated: the population is x_0 in this year and x_k in k years with an annual growth rate r [10].)

$$x_k = x_0(1 + r)^k \tag{2}$$

Logistic population growth model

(The analysis shows that when r > 0 (r is the annual growth rate), the population will increase indefinitely, which is definitely impossible. When the population increases to a certain number, the rate of growth will decrease, because natural resources, the environment, food, medical and health conditions, and other factors act as a brake on population growth, and the brake becomes stronger as the population increases. Thus, logistic population growth models are built on top of the Malthusian model. [10].)

Figure 1 contains the data information imported into the program, including population data for different ages, ages and countries, up to 253 countries or cities, but it is still not possible to compute accurate population predictions with this 10, 000 + data information. First, the data is oversimplified, and although the amount of data is enormous, it is all about age, which is not sufficiently diverse. Demographic shifts are not a simple discipline that can be accurately predicted by age alone. Second, instead of using the most accurate population prediction module, the code writes a more efficient standard deviation calculation module that goes straight to the subject, thus missing many details such as data accuracy and persuasion. In summary, general data can be predicted directly using data, wanting to be able to use as reference data requires additional details, but the results also justify the Python prediction module.

5. Conclusion

The program uses Python code to predict the future population and uses the results to compute with real data to obtain the standard deviation between the results and the actual data, which is an acceptable standard deviation and proves that Python can indeed predict approximations. From the results, the Python prediction can only be used as a reference and there is still a large deviation between it and the exact result, especially for data with a population in the hundreds of millions,

where the standard deviation of 0.9999708017086113 is not a small amount. However, the standard deviation of 0.9999708017086113 is not a small number if the death rate, the fertility rate and the fertility rate are adjusted before calculation. Other factors such as different data are also incorporated into the calculations and the results may become more accurate. The direct calculation of the standard deviation is not, after all, as accurate as the results given by the more refined population calculation model. In Fig. 9, the Python program already reminds the fact that the results may not be exact. Therefore, the results of directly using Python to compute the primary population and the projected population will not be available, but if more factors are added and certain types are calculated, the Python predictions may be able to reach the level of reference for future population trends.

In the past, the plotting tool Matplotlib has been used many times in attempted code writing. For example, the plot function used to compare the Python predictions with real data on the same line graph as in Fig. 2. However, the combination of Matplotlib and simulation results proved more difficult than expected, and was eventually forced to be abandoned under the pressure of time. However, this also allows the current theme to emerge, compared to reminding people of the reality of population problems such a huge topic, testing Python standard deviation is a lot more mundane, but only 20 lines of code to get such a calculation result makes people have to feel the progress of artificial intelligence and technology. In future code writing, the inaccuracies of the data analysis will be better adjusted, and more detailed data will be used to ensure the accuracy of the results. During the writing of the paper, additional knowledge related to the domain of expertise will be used to describe the results, and additional details will be given in the description of the results.

References

[1] Sanyi, Li. Elsevier B.V, A Modular Neural Network-Based Population Prediction Strategy for Evolutionary Dynamic Multi-Objective Optimization, https://www.sciencedirect.com/science/article/abs/pii/S221065022030482X?via%3Dihub. Accessed 25 Aug. 2023. (2019).

[2] KH, Kilburn. Population-Based Prediction Equations for Neurobehavioral Tests. Environmental Sciences Laboratory, pp. 257 – 263, University of Southern California School of Medicine, Los Angeles 90033, USA. (1998).

[3] Ghonchepour, Diba. Detection and Prediction of Land Use Changes and Population Dynamics in the Gorganrud River Basin, Iran, Wiley Online Library, https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.4662. Accessed 25. (2023) Aug.

[4] Justin, B. Population Risk Prediction Models for Incident Heart Failure: A Systematic Review. LIPPINCOTT WILLIAMS & WILKINS. (2015).

[5] Li Dong, Yu Yanyan, Wang Bo “Urban population prediction based on multi-objective lioness optimization algorithm and system dynamics model.” Journal | [J] Scientific reports. Volume 13, Issue 1. 2023. PP 11836 - 11836. (2023).

[6] Zhang Haiming, Xi Xiaoli, “Prediction of the change trend of preschool population in Guangzhou under the background of population agglomeration.” Journal of Shaanxi Preschool Teachers College. (2022).

[7] Li Yanru, Li Meng, Liu Shuang, “Prediction and Influencing factors of birth population development trend in Shanghai: Based on GM (1,1) grey prediction model.” Journal of Economic Research. (2022).

[8] Pang Mengyin, WANG Haining, Wan Tongming, Ma Miao, “Prediction of the number of confirmed cases based on the combined prediction model.” Journal of Computer technology and development. (2023).

[9] Ma Xuesong, “Research on Chinese population forecasting based on Leslie model and influencing factor screening.” Master electronic Journal Publishing House. (2023).

[10] Chun, He. The Application of Malthus Population Model in the Prediction of Guangzhou Population, China Academic Journal Electronic Publishing House. (2010).