العلوم الاقتصادية
العدد ثمانية وسبعون/ المجلد عشرون/ أيلول 2025

Some types of linear and nonlinear models used to forecast population numbers

HAYDER RAAID TALIB
University of Sumer / College of Administration & Economics
Noor Abdul-Kareem Fayadh
College of Administration and Economic, Sumer University, Iraq
Sarah Sabah akram
College of Administration and Economic, Wasit University, Iraq

Abstract :

Population forecasting is a fundamental pillar of development planning and public policy formulation, as strategic decisions are based on accurate estimates of future population size and its age and geographic distribution. Statistical models used in this field vary from linear models, which assume a direct proportional relationship between time or influencing factors and population size, to nonlinear models, which allow for the representation of more complex dynamics involving variable growth rates or environmental and economic constraints.

This study reviews the theoretical foundations and mathematical equations of both linear and nonlinear models, including the model, the multiscale linear regression-iterative multiscale model, the dynamic linear model (DLM), the spatio-temporal linear model, the exponential growth model, and the logistic model (nonparametric convolutional models). It also discusses the application conditions, parameter estimation methods, and the limitations of each model in light of the statistical characteristics of population data. This study concluded that the selection of the optimal model depends on the nature of the available data, the time horizon for forecasts ion, and the level of spatial analysis required.

Keywords: : population forecasting, linear models, nonlinear models, exponential growth model, dynamic linear model.

Introduction

Population forecasting is a specialized branch of population statistics and quantitative demography. It aims to estimate future population size and distribution by age, sex, and geographic location based on historical data and future assumptions about fertility, mortality, and migration rates. Population forecasting is considered one of the fundamental tools in economic and social planning. There are many population forecasting methods, including the models used in this article, which are divided into:

Linear models: It is assumed that the change in population is linearly proportional to time or to a set of explanatory variables. These models depend on simplifying assumptions that make the rate of change constant or close to constant, such as a multi-scale model, and sometimes multi-dimensional linear formulas, such as the Leslie Matrix, which represents age changes using matrix multiplication operations.

Nonlinear models: These models deal with population growth at a variable rate that increases or decreases over time due to dynamic factors or environmental and economic constraints. They include the exponential model, the logistic model, and more complex models such as the Gompertz Richards model. These models are able to represent population saturation and the slowdown in growth rates when approaching the carrying capacity.

The development and use of these models is not limited to the academic aspect but extends to public policy and resource management, making population forecasting a vital tool for sustainable development and strategic decision-making.

Linear models: There are several linear models used to forecasts population numbers, including:

Multiscale linear regression- iterative multiscale: The multiscale regression model is a statistical model that aims to integrate the information received, meaning to integrate interpretive signals at different levels of space/scale (such as: the local neighborhood level, the regional level, the national level) into a single linear formula, allowing each scale to explain part of the variance in the dependent variable (population or growth rate) from different levels of measurement in a single forecasts ive equation for the purpose of improving the accuracy of population estimation. This model is based on the assumption that the change in population size does not depend on a single factor, but is affected by several factors that may be measured at different levels of geographical or temporal analysis. For example, the levels of measurement may include the following [2]:

Local level: Data from neighborhoods or municipalities, such as the number of schools, health centers, or service provision.

Regional level: Data from provinces or states such as unemployment rates, average income, or regional infrastructure.

National level: Data at the country level such as economic growth rate or population and migration policies.

This model expresses the population or its logarithm in area i and time t where we consider $(y_{i,t})$ to be the target variable for area i at time t (e.g., the logarithm of the population or the growth rate). We assume M scales and each scale has a vector of variables $(X_{i,t}^{(m)})$ and coefficients $(\beta^{(m)})$ where the model has the following form:

$$y_{i,t} = \sum_{m=1}^{M} X_{i,t}^{(m)} \beta^{(m)} + \varepsilon_{(i,t)} \qquad \varepsilon_{(i,t)} \sim N(0, \sigma^2)$$

The most important assumptions of this model are that the coefficients $(\beta^{(m)})$ are linear, the errors $(\varepsilon_{(i,t)})$ are independent with a certain structure, and the variables at different scales may be multilinear so the model needs regularization procedures or an iterative algorithm to disentangle the effects [6].

The forecasting mechanism depends on estimating the coefficients $(\hat{\beta}^{(m)})$ using one of the estimation methods (ordinary least squares, backfitting) and then entering the future or expected values of the explanatory variables at all scales in the previous equation. This results in the expected estimate of the population size or the expected value of the explanatory variables at all scales in the previous equation. This results in the expected estimate of the population size or its growth rate in the future periods [2]. In the case that the logarithmic transformation is used during the modeling, the expected values are returned to the actual population scale using the exponential transformation, taking into account the bias correction if necessary.

Dynamic Linear Model- DLM: This dynamic time-linear model is one of the powerful statistical models used to analyze time series and forecast future values, especially in cases where the target variable is not directly observed but is monitored through data containing noise or measurement errors. In the context of forecasts ing measurement numbers, this model assumes that the actual population number in any time period is a latent state that develops over time, while the population data or statistics represent observations of this state.

This model is based on two interconnected levels: the state level, which describes the true, unobserved values of the population, and the observation level, which links these true values to the recorded data (State equation that describes the evolution of the real population from one period to another), where they are expressed through the following formulas [5]

$$x_{t+1} = F_t x_t + U_t$$

Where:

x_{t+1} is the state vector at time t, which may contain other elements such as the general population trend m_t and the rate of change b_t.

F_t: time transition matrix that specifies how the state transitions between periods.

U_t: Random error vector.

Observation equation that describes the relationship between the condition and the recorded values:

$$y_t = H_t x_t + U_t$$

x_t: The value of the observed population in time period t, the value we actually obtain from the census or population statistics and may contain measurement error.

H_t: Link matrix (a matrix that specifies how to convert the state elements into the value that can be observed if the state x_t consists of more than one component such as the population level m_t and the slope b_t).

U_t: Random error vector.

Population numbers in this model are forecasted using a Kalman filter. The process starts by estimation the current state $(\hat{x}_t)$ and its covariance matrix (P_t) based on all previous observations. Then, the equation of state is used to forecasts the future state after one-time period:

$$\hat{x}_t + \frac{1}{t} = F_t \hat{x}_t$$

If we want to forecast h future periods, we repeat the transition process.

$$\hat{x}_t + \frac{h}{t} = F_t \hat{x}_t$$

Then we use the observation equation to convert the expected state into the expected population:

$$\hat{y}_t + \frac{h}{t} = H_t \hat{x}_t + \frac{h}{t}$$

Where the uncertainty in the forecasts ion is calculated using the following formula:

$$\text{Var}(\hat{y}_t + \tfrac{h}{t}) = P_t H_t + \tfrac{h}{t} H_t^T + R_t$$

It allows confidence intervals to be set around the forecasts ed values and is an essential tool for estimating the accuracy of forecasting.

One of the most important advantages of this model is that it is dynamic, meaning that it updates estimates as soon as new observations become available, making forecasting more accurate over time, as it is flexible enough to represent long-term trends and seasonal or cyclical fluctuations in population numbers while maintaining the simplicity of the linear structure of the coefficients [5].

Spatio- Temporal Linear Mixed Model: It is a statistical framework that combines spatial and temporal structures in a single linear formula with the aim of representing and explaining changes in data across both time and space. In the context of population forecasting, this model assumes that population distribution is affected in each time period by several factors, including local factors specific to each region, in addition to spatial interactions that reflect the influence of neighboring regions on each other. The model is based on the assumption that previous population values in the same region (temporal dependence) and values in neighboring or nearby regions (spatial dependence) and additional explanatory variables may be fixed or change over time. The model can be expressed in the following mathematical formula [1]:

$$y_{i,t} = \alpha + p \sum_{j \epsilon N(i)} w \phi y_{i,t-1} + x_i , t^T \beta + \varepsilon_{(i,t)}$$

$y_{i,t}$: Population of the viewer at location i and time period t.

α: Model constant (intercept).

p: the spatial correlation coefficient measures the strength of the influence of values in neighboring areas.

$N(i)$: The set of regions adjacent to region i.

w_{ij}: The weight of the spatial relationship between region i and region j. These weights are determined based on the distance or contact boundaries.

Φ: The time dependence coefficient measures the influence of past population values in the same area.

x_i: Vector of explanatory variables in region i and period t (such as unemployment rates, income, or migration).

β: Regression coefficients that determine the effect of each explanatory variable.

$\varepsilon_{(i,t)}$: the random error component assumes a normal distribution with zero mean and constant variance.

The forecasting mechanism starts after estimating each of $(\hat{\alpha}\ ,\hat{\beta},\ \widehat{\Phi}\ ,\hat{P})$When these estimates are provided, future forecasts ions of the population in region i and time t+1 are calculated, where the spatial component is first calculated by adding the observed population numbers in the neighboring regions multiplied by the spatial relationship weights (w_{ij}) and then multiplying the result by the spatial correlation coefficient P. This reflects the statistical effect of the geographical neighborhood on the population in the target region. [1]

Then the temporal component is calculated by relying on the population value in the same area in the previous period multiplied by the temporal correlation coefficient Φ. This reflects the continuity of the temporal pattern of the population within the area. Then, the effect of the explanatory variables (migration rates, etc.) will be added by multiplying them by the regression coefficients β that were estimated from the previous data. Finally, all these components are combined into a single formula to obtain the expected value of the population:

$$\hat{y}_{i,t+1} = \hat{\alpha} + \hat{p}\sum j\epsilon N(i) w_{ij} y_{i,t} + \widehat{\Phi}\ \ yi,t + xi,t\ + 1^{T},\hat{\beta}$$

where $\hat{y}_{i,}t+1$ represents the expected population in area i and period t+1.

In this way, the model combines spatial and temporal effects in the forecasts ion process and at the same time provides estimates of confidence intervals, which helps in evaluating forecasts and determining the degree of statistical certainty associated with them [1].

Nonlinear models: There are several nonlinear models used to forecasts population numbers, including:

2-1 Exponential Growth Models: The exponential model is considered one of the basic nonlinear models in population growth analysis, as it is based on the assumption that the relative population growth rate is constant over time, meaning that the percentage of increase in the population in any period of time is directly proportional to the current population size. This assumption can be formulated using the differential equation[3]:

$$\frac{dP(t)}{dt} = r\ p(t)$$

Where p(t) represents the population as a function of time and r is the relative growth rate that determines the speed of change in population size. By solving this equation, we will obtain the closed form of the model:

$$p(t) = P_0\ e^{rt}$$

Where P_0 represents the population at the initial time t=0 and e is the natural base of the logarithms (≈ 2.71828)

The population forecasting mechanism is carried out after estimating the parameters P_0 and r using data for the population over time. When these estimates are available, future values can be calculated by entering the target

time into the equation. $\hat{p}(t+h) = P_0\, e^{r(t+h)}$ represents the statistical estimate of the population after h periods of time from the current time. The model assumes that the growth rate remains constant, which makes it suitable for short-term forecasts or cases where there are no clear constraints on growth [3].

2-2 Nonparametric Convolutional Models: This model is considered a development of the exponential model that takes into account the presence of a carrier capacity K, which represents the maximum number of people that the environment can support in the long term. The model imposes that the growth rate gradually decreases with the increase in population size, so that growth is rapid in the early stages and then gradually slows down when approaching K. This concept can be expressed by the differential equation[4]:

$$\frac{dP(t)}{dt} = r\, p(t)\left(1 - \frac{P(t)}{K}\right)$$

Where $P(t)$ is the population at time t, r is the maximum growth rate at the beginning of the period, and K is the carrying capacity. By solving the equation, we obtain the following equation [4]:

$$p(t) = \left(\frac{K}{1 + Ae^{-rt}}\right)$$

Where $A = \left(\frac{K-P_0}{P_0}\right)$ which determines the starting point and is the population at time t=0.

The prediction mechanism is performed here after estimating the parameters (K, P_0, r) from the data using nonlinear regression methods. The equation is then used to calculate future values, giving the expression $\widehat{P}(t+h) = \frac{K}{1+Ae^{-rt}}$, which means estimating the population size after h time periods. This model is distinguished by its ability to represent the phenomenon of environmental saturation, making it the most suitable model for long-term forecasts.

Linear and nonlinear models can be compared through advantages, disadvantages, and Optimal areas of use:

The model	Advantages	Disadvantages	Optimal areas of use
Multiscale linear regression- iterative multiscale	Ability to interpret multilevel variance Integrates variables across different spatial and temporal dimensions Easily interpretable coefficients	Sensitive to Multicollinearity, requires advanced regulation.	Forecasting at local and regional levels when multi-source data is available
Dynamic Linear Model DLM	Forecasts are updated as new data arrives, representing seasonal trends.	Requires good time series data and is computationally complex.	Long-term population data with noise or missing data.
Spatio- Temporal Linear Mixed Model	Accurately combines spatial and temporal effects.	Difficulty in estimating many transactions Computational burden	Regional forecasting when there is interaction between adjacent areas
Exponential Growth Models	Simple and easy to estimate, suitable for short periods	Assumes a growth rate that ignores resource constraints.	Short-term germination or free-growing conditions
Nonparametric Convolutional Models	Reflects environmental saturation, estimates carrier capacity	Less accuracy in the short term requires good parameter data.	Long-term forecasts with economic/environmental constraints.

The most important conclusions reached in this article:

Conducting an actual and regular population census is the basis for obtaining accurate and reliable data.

It represents the basis for obtaining accurate and reliable data.

Allows monitoring of population changes over time.

Choosing the appropriate statistical model.

This is done by studying the behavior of historical population data and growth rates.

Linear models (such as linear regression) or non-linear models (such as exponential or logistic growth) can be used.

Comparison between linear and nonlinear models.

The predicted results of different models are compared using statistical comparison criteria.

One of the most important statistical criteria for comparing models is AIC to measure the goodness of fit, reduce information loss, and other criteria.

The accuracy of forecasts depends on the quality of the data:

Regular and comprehensive census reduces deviations in models.

Population growth does not always follow a single pattern.

It may be affected by economic, social, health factors, and internal and external migrations.

The necessity of choosing more than one model.

There is no ideal model for every case, but rather the most appropriate model varies according to the nature of society and the stage of development.

The importance of updating forecasts periodically.

Sudden changes such as wars, epidemics, or economic booms may alter population trends.

The role of external factors.

Natural disasters, pandemics (such as COVID-19), and wars can cause sudden and unexpected changes in population size.

The need to integrate more than one data source.

It is not enough to rely only on censuses, but rather it is possible to integrate vital records data (births, deaths), migration data, and field surveys.

spatial variation.

Population growth is not equal in all regions. Some cities may experience rapid growth, while other regions may experience stability or decline.

Recommendations:

Encouraging cooperation between disciplines (statistics, sociology, economics) to produce more comprehensive models that reflect the various dimensions affecting population change.
Focus on data quality by conducting a comprehensive and regular census supported by modern technologies such as geographic information systems and remote sensing.
Enhancing international cooperation and exchange of expertise in the field of population forecasting, particularly in addressing cross-border challenges such as mass displacement or climate change.
Choosing the appropriate prediction model that matches the nature of the data.

References

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers, 107(6), 1247-1265.

Pearl, R., & Reed, L. J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proceedings of the national academy of sciences, 6(6), 275-288.

Tsoularis, A., & Wallace, J. (2002). Analysis of logistic growth models. Mathematical biosciences, 179(1), 21-55.

West, M., & Harrison, J. (1997). Bayesian forecasting and dynamic models. New York, NY: Springer New York.

Xi, Y., Liu, Y., Li, T., Ding, J., Zhang, Y., Tarkoma, S., ... & Hui, P. (2023). A satellite imagery dataset for long-term sustainable development in united states cities. Scientific data, 10(1), 866.

United nations report , department of economic and social Affairs , population division , 2011

United nations Manual x , Department of International Economic and Social Affairs Population Studies , ST/ESA/SER.A/81 , 2000