ISSN:2790-1688 Volume-15-(2025)

A Review of Population Structure Dynamics and Risk Warning Based on Time Series Models

Yicheng Li

Jiaxing Senior High School BCOS, Jiaxing 314000, China

Abstract. In the fields of population data analysis and sociological research, time series models, particularly ARIMA and VAR models, are widely applied and highly regarded for their stability and interpretability. This paper systematically reviews the application of ARIMA models in modeling and forecasting key demographic indicators, such as total population size and aging rates. This model excels at capturing temporal patterns and trends within univariate sequences, making it a core tool for population forecasting. Furthermore, this paper employs VAR models to explore the multifaceted societal impacts of population aging. VAR models excel at analyzing the interactive dynamics among multiple variables such as aging, economic growth, and social welfare expenditures. By integrating the risk assessment capabilities of VAR models, this study aims to summarize the challenges posed by demographic shifts and the practical applications of ARIMA and VAR models in population data, providing foundational theoretical support for evidence-based policy design.

Keywords: ARIMA; VAR; Population; Aging; Risk Alert.

1. Introduction

Systematic attention and research on population data are not new, but their strategic importance has been elevated to unprecedented levels in recent years. Countries around the world have long maintained population statistics, such as China's household registration system and church records in Europe. However, modern-style population censuses and systematic data collection based on scientific methods began in the late 18th to 19th centuries, exemplified by the 1790 census in the United States and the 1801 census in Britain [19,20,21]. Scientists explore population characteristics by studying the interactions between individuals within a population and the relationships and effects of the entire population on the environment [18]. The field of science that collects and analyzes these data is known as population ecology, or demography, which describes the patterns of change in these characteristics over time through mathematical modeling [23]. Demography encompasses all the statistical factors that influence population growth and decline, such as population size, density, age structure, fertility (birth rate), mortality (death rate), and sex ratio [22].

However, since the mid-to-late 20th century, particularly in the 1970s and 1980s, global demographics have undergone significant shifts: persistently declining fertility rates coupled with markedly extended life expectancy have collectively driven a fundamental transformation in population structure: aging [24]. The impact of demographic shifts is no longer a gradual sociological phenomenon, but directly affects the economy, social security, the labor market, public assets, and even the nation's overall competitiveness [25,26]. Therefore, shifting from static descriptions to dynamic predictions and from post-event analysis to preemptive warnings has become an urgent requirement for population research and management decision-making. This provides a powerful practical impetus for the application of time series models in this field.

To address these demands, time series analysis methods from econometrics and statistics have been introduced into the field of population studies, with ARIMA models and VAR models serving as two core and representative tools. In 1970, statisticians George Box and Gwilym Jenkins introduced the theoretical framework and modeling process for the ARIMA (Autoregressive Integrated Moving Average) model in their work Time Series Analysis: Forecasting and Control. The core idea of this method is to make the sequence stationary through differencing and capture its dynamic patterns using a combination of AR and MA models [27]. Many demographic data points, such as birth rates, death rates, and total population figures, are inherently time series data. The

ARIMA model excels at handling such univariate time series and making short-term, precise forecasts. Its flexible structure relies solely on the historical information of the variable itself for prediction, proving highly effective when data is limited or causal relationships between variables are unclear [14]. This model is widely used to forecast key indicators, including total population, birth rates, death rates, and aging rates [2]. The VAR model was proposed and advocated by econometrician Christopher Sims in his paper "Macroeconomics and Reality," utilizing Vector Autoregression models. The VAR model jointly models all variables using their lagged values, capturing the dynamic interactive relationships within a multivariate system [28]. ARIMA models cannot explain the complex relationships between demographic variables and other socioeconomic variables, whereas VAR models are suitable for analyzing the interdependent mechanisms and dynamic pathways within such multivariate systems [29]. Through impulse response and variance decomposition, the intensity and duration of an impact from one variable on another can be quantified. This is crucial for understanding risk transmission mechanisms, conducting multi-scenario simulations, and evaluating policy effects, making it a core analytical tool for risk early warning systems [30].

The profound shifts in global population structures and the systemic risks they entail demand that researchers examine demographic issues from dynamic and predictive perspectives. Although numerous studies have applied ARIMA or VAR models to population domains separately, there remains a lack of systematic review, comparison, and outlook regarding the application of these two model types in simulating demographic dynamics and conducting risk early warning. This review aims to fill this gap by synthesizing existing research, evaluating the applicability and limitations of different models, and guiding future research directions. It provides a theoretical reference for constructing a scientifically sound and effective population risk early warning system.

This study reviews the modeling and forecasting applications of ARIMA models in time series data such as total population, aging rates, fertility rates, and migration rates. It also integrates the extension of VAR models in population risk assessment, including concepts like "risk thresholds for accelerated aging," "risk probabilities of negative population growth," and "social carrying capacity risks associated with large-scale population movements." This paper provides a detailed account of the functions and modeling processes of ARIMA and VAR models. It reviews a substantial body of prior literature and conducts a comparative analysis of existing research findings. Finally, it discusses the practical application of ARIMA and VAR models in population-related data and the importance of tailoring data forecasting approaches to specific contexts, offering guiding recommendations for future research.

The second section of this paper briefly introduces the application of two time series models in demography: the ARIMA and VAR models. In the third section, previous research findings on both the ARIMA and VAR models are reviewed, followed by a concise discussion and analysis.

2. Relevant Model for Population Data Analysis

2.1 ARIMA (Autoregressive Integrated Moving Average model)

The ARIMA model is one of the most classic models in time series forecasting. It is an extension of the ARMA model and is specifically designed to handle non-stationary time series [13,14]. The ARIMA model represents the linear relationships within a time series. If the data contains complex nonlinear patterns, the ARIMA model may not capture them effectively. The ARIMA model makes predictions based on its own historical data and historical errors, so it typically performs well in short to medium-term predictions [13]. However, for long-term predictions, prediction errors accumulate over time, leading to a rapid increase in prediction uncertainty and a decline in accuracy. A core requirement of the ARIMA model is that, after integration, the sequence must become a stationary sequence, which is a prerequisite for the model's effectiveness.

The ARIMA model consists of three components: The Autoregressive (AR) model, the Moving Average (MA) model, and Integrated (I) method for data processing. Below is a detailed introduction to the three models:

AR (Autoregressive model)

AR models use the lagged effects of lagged parameters from past population observations to predict current or future population observations. This captures the "inertia" or "memory" effect, which means that past values influence present values. The model represents the current value as a linear combination of the values from the past "p" moments, which can be expressed by the following formula:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + c + \varepsilon_t \quad (1)$$

X_t represents a population observation at time t, $\varphi_1, \ldots, \varphi_p$ represent the first p moments of the lag parameter, $\varphi_1 X_{t-1}, \ldots, \varphi_p X_{t-p}$ represent the lagged effect of the first p observations on the current predicted value, c represents a constant term, and ε_t represents the error term at moment t (random fluctuations).

MA（Moving Average model）

The MA model uses the values of past prediction errors to predict current or future observations of the population. This captures the impact of historical unpredictability (e.g., sudden accident) on current or future values. The model expresses the value at the current or future moment as a linear combination of random errors (white noise) from the past "q" moments, which can be expressed by the following formula:

$$X_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \mu \quad (2)$$

$\theta_t, \ldots, \theta_{t-q}$ represents the lag parameter, $\varepsilon_t, \ldots, \varepsilon_{t-q}$ represents the prediction error at the first q moments, μ represents the average of previous population observations, $\theta_1 X_{t-1}, \ldots, \theta_p X_{t-q}$ represents the lag effect.

ARMA

Combining the AR model and the MA model results in the ARMA model, which is expressed by the following formula:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \theta_{t-1}\varepsilon_{t-1} + \theta_{t-2}\varepsilon_{t-2} + \cdots + \theta_{t-p}\varepsilon_{t-q} + c + \varepsilon_t + \mu \quad (3)$$

Integrated

This is the key to the ability of ARIMA models to handle non-stationary time series. By applying differencing operations (calculating the differences between adjacent observations), a non-stationary series with trends or seasonality can be transformed into a stationary series, which can then be fitted using AR and MA. The Integrated calculation is as follows:

First-order integrated:

$$Y_t = X_t - X_{t-1} \quad (4)$$

$$Y_{t-1} = X_{t-1} - X_{t-2} \quad (5)$$

Second-order integrated:

$$Z_t = Y_t - Y_{t-1} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \quad (6)$$

Y_t is the first-order integrated of $X_t - X_{t-1}$, Y_{t-1} is the first-order integrated of $X_{t-1} - X_{t-2}$, Z_t is the second-order integrated of $Y_t - Y_{t-1}$.

2.2 VAR (Vector Autoregression model)

The VAR model is an econometric model used to analyze multivariate time series systems. In the VAR model, all variables in the system are treated as endogenous variables, i.e., variables that influence each other. It does not preemptively distinguish between cause and effect but instead uses the data itself to reveal the dynamic interactive relationships between variables [16].

VAR models are suitable for multivariate time series data. Ideally, all variables should be stationary time series [17]. Since multiple parameters need to be estimated, a sufficiently long sample size is required to ensure the accuracy and stability of the estimates [16,17]. The data frequencies of all variables would be consistent.

The population system itself is a complex, multidimensional, and highly interconnected system, which aligns perfectly with the characteristics of the VAR model. Population variables exhibit strong bidirectional causal relationships with economic, social, and environmental variables. For example,

economic development, GDP growth, and increased income influence labor mobility, improved healthcare conditions reduce mortality rates, and rising opportunity costs affect fertility rates. Conversely, demographic structure, such as the proportion of the working-age population and the degree of aging, directly impacts economic growth potential and consumer markets. The VAR model can effectively capture this complex feedback mechanism, whereas the single-variable ARIMA model cannot.

The effects of demographic changes are often long-term and lagged. A peak in fertility rates takes approximately 20 years to translate into a peak in labor supply, which in turn affects the economy. The impulse response function of the VAR model can clearly illustrate the delay, duration, and eventual convergence of this effect. Population risks rarely occur in isolation. Aging simultaneously exerts pressure on pension systems, healthcare expenditures, and labor markets. The variance decomposition of the VAR model can help quantify the contribution of different population shocks to various risks, providing a basis for systemic risk assessment [15].

Suppose there are k time series variables, where t represents the time point. Then the $k \times 1$ column vector y_t can be expressed as follows:

$$y_t = (y_{1t}, y_{2t}, \dots, y_{kt})' \tag{7}$$

The VAR model also uses past observations to predict current or future values. Let m represent the observations from the past m moments. The VAR model can be expressed by the following formula:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_m y_{t-m} + u_t \tag{8}$$

y_t is a column vector of size $k \times 1$, A_m is a $k \times k$ matrix, usually determined using the least squares method (the number of columns in the matrix and the number of rows in the vector must be the same to multiply them). m represents the lag order and also represents the order of the model. u_t is a $k \times 1$ column vector, and is the error term or white noise (random error with zero mean and constant covariance).

3. Results and Discussion

In order to review the modeling and predictive applications of ARIMA models in time series data, such as total population and aging rates, it is necessary to summarize the findings of previous studies. As a country with a large population, China's demographic issues have always been a major factor influencing its economic and social development. Therefore, we will begin by examining population samples from some of China's provinces [2].

In order to model the time series of the total population of Guizhou Province, China, Tian Yingfu and Miao Boqi compared the relative error of the predicted sample range from 1952 to 2002 and the relative error of the predicted range (2002 and 2003), and ultimately selected ARIMA (2,1,3) as the optimal solution for modeling the total population of Guizhou Province [1]. In modeling the total population data of Zhejiang Province, the ARIMA (1,1,0) model was calculated to be the best fitting model, and the SBC and AIC principles were used to verify that this model was the most appropriate model [2]. In order to achieve more diverse results, in addition to sample data from two provinces in China, the population of Bandung, Indonesia, was also studied. Based on the results of AIC and SIC, the ARIMA (1,1,0) model was selected to predict population because its AIC and SIC values were lower than those of other ARIMA models [3]. Similarly, when predicting Pakistan's population, selecting the ARIMA (1,2,0) model is most reasonable [4]. In fact, selecting an appropriate model for time series prediction of population data is an art rather than a mathematical science [4]. In simulating real-world scenarios, the significance level must be relaxed to meet the assumptions [4]. In the population forecasting model for Madiun Regency, the most suitable model was calculated to be ARIMA (0,2,1) [5].

Through the population forecasting models developed for different regions, it can be observed that while there may be similarities in the models used across different regions or countries, the majority are distinct. Therefore, population forecasting models for different regions must be tailored to local

conditions. To identify the most suitable population forecasting model, it is first necessary to calculate the values of AR(p) and MA(q) to determine the approximate range of the ARIMA model. To select the most appropriate model, further calculations are required, such as the white noise test, and different computational criteria can be used, such as AIC or SIC. Finally, the model with the smallest calculated value is selected as the most suitable model.

Once the most suitable ARIMA model is obtained, predictions can be made. The population prediction data for each province and country mentioned above, predicted using the corresponding ARIMA models, is very close to the actual population data, further proving the accuracy of the ARIMA model.

The ARIMA model can not only predict the total population of a region but also predict the population infected with the novel coronavirus. Hernandez-Matamoros, A. et al. proposed an algorithm to calculate the optimal ARIMA parameters for each country with a low RMSE. The algorithm for calculating the optimal ARIMA parameters was tested using 10% of the original data. This method analyzes specific cases (countries) to create a general case (geographical region) [6]. In response to the above situation, VAR models can be used to integrate data from numerous countries and regions, as well as from different sources, to conduct risk assessments.

The primary function of the VAR model is to integrate the appropriate ARIMA model developed in the initial modeling phase into the extension of population risk assessment, such as the risk threshold for excessive aging rates, the probability of risk associated with negative population growth, and the social carrying capacity risk associated with large-scale population movements, etc..

In the study by Lopreite and Mauro, the focus was on enhancing understanding of the impact of population aging on the growth of healthcare spending in Italy[7]. They used a Minnesota-based B-VAR model to analyze the relationship between per capita healthcare expenditure, per capita GDP, aging index, and life expectancy, combining useful information to improve both short-term and long-term analysis [7]. In their paper, Houjian, L. et al. used a VAR model to study how China's aging population and renewable energy consumption affect agricultural green total factor productivity [11]. Population aging has only a short-term hindering effect on agricultural green total factor productivity. In the long run, an increase in the aging population will inevitably lead to an improvement in agricultural green total factor productivity [11]. The possible reason is that aging provides society with a large pool of skilled labor and deepens human capital reserves, and these accumulated conditions can enhance agricultural green total factor productivity [8,9,10]. As such, VAR models are frequently employed to forecast risks associated with population aging. Utilizing VAR models can facilitate the assessment of the relationship between longevity and national health conditions, evaluate the dampening effect of population aging on renewable energy consumption, and clarify the types of policies or research recommendations currently required [7,11].

In addition to showing the benefits of demographic changes for the future, VAR models can also be used to assess risks in certain areas. For example, using VAR models to assess socioeconomic risks such as labor shortages and wage reductions caused by the large decline in Romania's rural population due to political factors [12].

Based on the above articles, the VAR model is a powerful basic tool for population risk assessment. Using the system equations estimated by VAR, it is possible to artificially set an extreme but possible risk, and then simulate the impact of this risk on GDP growth rate, government pension expenditure, labor supply, and other aspects over the next 20 to 30 years through impulse response functions. At the same time, policy variables can be incorporated into the VAR system to simulate the corresponding effects and provide support for reasonable policies.

4. Conclusion

This paper briefly introduces the functions and modeling processes of the ARIMA and VAR models, reviews prior literature on their practical applications, and integrates both models for population forecasting and risk early warning. This review fills a gap in the literature by integrating

ARIMA and VAR models for population dynamics simulation and risk early warning, while also providing comparative analysis and future directions. It synthesizes existing research findings, evaluates the applicability and limitations of different models, and guides future research pathways. The study offers new insights for model extension and provides theoretical references for constructing scientifically sound and effective population risk early warning systems.

References

[1] Tian, Y., & Miao, B. (2006). ARIMA model of Guizhou Province's total population time series. Journal of Yunnan Minzu University: Natural Science Edition, 15(3), 4.

[2] Dai, J., & Chen, S. (2019). The application of ARIMA model in forecasting population data. Journal of Physics: Conference Series. IOP Publishing. 1324(1), 012100.

[3] Setyawan, E. B., Novitasari, N., & Muttaqin, P. S. (2020). Multi-variable forecasting model using ARIMA (P, Q, N) method to project number of population in Bandung, Indonesia. IOP Conference Series: Materials Science and Engineering. IOP Publishing. 830(3), 032088.

[4] Zakria, M., & Muhammad, F. (2009). Forecasting the population of Pakistan using ARIMA models. Pakistan Journal of Agricultural Sciences, 46(3), 214-223.

[5] Farida, Y., Farmita, M., Ulinnuha, N., & Yuliati, D. (2022). Forecasting Population of Madiun Regency Using ARIMA Method. CAUCHY: Jurnal Matematika Murni dan Aplikasi, 7(3), 420-431.

[6] Hernandez-Matamoros, A., Fujita, H., Hayashi, T., & Perez-Meana, H. (2020). Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Applied soft computing, 96, 106610.

[7] Lopreite, M., & Mauro, M. (2017). The effects of population ageing on health care expenditure: a Bayesian VAR analysis using data from Italy. Health policy, 121(6), 663-674.

[8] Fougère, M., & Mérette, M. (1999). Population ageing and economic growth in seven OECD countries. Economic Modelling, 16(3), 411-427.

[9] Fougère, M., Harvey, S., Mercenier, J., & Mérette, M. (2009). Population ageing, time allocation and human capital: A general equilibrium analysis for Canada. Economic Modelling, 26(1), 30-39.

[10] Čiutienė, R., & Railaitė, R. (2015). A development of human capital in the context of an aging population. Procedia-Social and Behavioral Sciences, 213, 753-757.

[11] Li, H., Zhou, X., Tang, M., & Guo, L. (2022). Impact of population aging and renewable energy consumption on agricultural green total factor productivity in rural China: Evidence from panel var approach. Agriculture, 12(5), 715.

[12] Dumitru, E. A., Sterie, M. C., & Dragomir, N. (2023). Political events upon the Romania rural population using VAR model. Ciência Rural, 54(3), e20230066.

[13] Zhang, M. (2018). Time series: Autoregressive models ar, ma, arma, arima. University of Pittsburgh.

[14] Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. In Time series analysis and its applications: with R examples. Cham: Springer International Publishing. 75-163.

[15] Benninga, S., & Wiener, Z. (1998). Value-at-risk (VaR). matrix, 32, s33.

[16] Qin, D. (2011). Rise of VAR modelling approach. Journal of Economic Surveys, 25(1), 156-174.

[17] Benati, L., & Surico, P. (2009). VAR analysis and the Great Moderation. American Economic Review, 99(4), 1636-1652.

[18] Tarsi, K., & Tuff, T. (2012). Introduction to population. Nat Educ Knowl, 3(11), 3.

[19] Malthus, T. R., Bonar, J. (1965). First Essay on Population, 1798. US: A. M. Kelley, Bookseller.

[20] Wright, C. D. (1900). The history and growth of the United States census (Vol. 194). US: US Government Printing Office.

[21] Taylor, A. J. (1951). Taking of the Census, 1801-1951. British medical journal, 1(4709), 715.

[22] Dodge, Y. (Ed.). (2003). The Oxford dictionary of statistical terms. Oxford university press.

[23] Swedlund, A. C. (1978). Historical demography as population ecology. Annual Review of Anthropology, 7, 137-173.

[24] Rowland, D. T. (2009). Global population aging: History and prospects. In International handbook of population aging. Dordrecht: Springer Netherlands. 37-65.

[25] Bloom, D. E., Boersch-Supan, A., McGee, P., & Seike, A. (2011). Population aging: facts, challenges, and responses. Benefits and compensation International, 41(1), 22.

[26] Veras, R. (2009). Population aging today: demands, challenges and innovations. Revista de saúde pública, 43, 548-554.

[27] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. US: John Wiley & Sons.

[28] Sims, C. A. (1980). Macroeconomics and reality. Econometrica: journal of the Econometric Society, 1-48.

[29] Khan, M. S., & Khan, U. (2020). Comparison of forecasting performance with VAR vs. ARIMA models using economic variables of Bangladesh. Asian Journal of Probability and Statistics, 10(2), 33-47.

[30] Benninga, S., & Wiener, Z. (1998). Value-at-risk (VaR). matrix, 32, s33.