Advanced Programming 2025

Swiss Population Forecasting : Baseline vs Linear vs AR Model

Final Project Report

Ali Yassir Dhina
AliYassir.Dhina@unil.ch
Student ID: 20861886

January 7, 2026

Abstract

Population forecasting is a central task in economics and public policy, as it directly affects planning in areas such as infrastructure, healthcare, housing, and pensions. This project studies the problem of forecasting the total resident population of Switzerland using historical demographic data. The objective is not only to produce forecasts, but to compare the performance of increasingly sophisticated models against a simple baseline.

Using yearly population data from the Swiss Federal Statistical Office covering the period 1860–2024, three approaches are evaluated. First, a naive baseline model assumes a constant population growth rate estimated from recent historical data. Second, a linear regression model predicts next-year population using current population levels, growth rates, and lagged values. Third, an autoregressive (AR) model forecasts population based solely on past population values.

Models are evaluated using a strict time-based train–test split and root mean squared error (RMSE) as the performance metric. Results show that both statistical and machine learning-based models substantially outperform the constant-growth baseline. Among the tested approaches, linear regression achieves the lowest forecasting error, while the AR model provides a competitive but slightly less accurate alternative. The findings highlight the importance of formal model comparison even in data-limited time-series settings.

Keywords: population forecasting, times series, regression, autoregressive models, model comparison, Switzerland

Contents

1 Introduction

Population dynamics play a crucial role in economic analysis and public policy design. Accurate population forecasts are required for long-term planning in areas such as education, labor markets, healthcare systems, and pension schemes. In countries with reliable historical data, population forecasting also provides an opportunity to evaluate and compare different statistical and machine learning methods in a controlled setting.

Switzerland represents an interesting case study for population forecasting. The country has a long and well-documented demographic history, stable institutions, and a population size that evolves gradually over time. At the same time, demographic trends are influenced by migration, fertility, and mortality patterns, making long-term forecasting inherently uncertain.

Many population projections rely on simple assumptions, such as constant growth rates or trend extrapolation. While these approaches are easy to implement, they often fail to capture changes in demographic dynamics. More sophisticated statistical and machine learning models may improve forecasting accuracy, but their added complexity must be justified through formal evaluation.

The objective of this project is to compare the performance of a simple baseline population growth model with more advanced forecasting approaches. Specifically, three models are evaluated: a constant-growth baseline, a linear regression model with lagged features, and an auto-regressive (AR) model. All models are trained and tested using a consistent time-based evaluation framework. The main contribution of the project is a transparent and reproducible comparison of these methods, highlighting the gains and limitations of machine learning approaches in a data-constrained time-series environment.

The remainder of this report is organized as follows. Section 2 reviews related work on population forecasting and time-series modeling. Section 3 describes the data, modeling approaches, and implementation details. Section 5 presents the empirical results and visualizations. Section 6 discusses the findings and limitations, and Section 7 concludes with directions for future research.

2 Literature Review / Related Work

Traditional approaches often rely on simple extrapolation techniques, such as constant-growth models, which assume a stable long-run growth rate estimated from historical data.

More flexible time-series methods, including autoregressive (AR) models, allow population dynamics to depend explicitly on past realizations, capturing persistence and inertia in demographic trends. Regression-based forecasting models further extend this framework by incorporating additional explanatory variables or transformations of historical population levels, such as growth rates or lagged values.

Recent applied work emphasizes model comparison rather than reliance on a single forecasting approach, particularly when data availability is limited. In such settings, transparent and interpretable models are often preferred over complex machine learning algorithms, whose performance gains may be constrained by small sample sizes.

This project follows this comparative approach by evaluating a baseline constant-growth model alongside linear regression and autoregressive specifications using Swiss population data.

3 Methodology

3.1 Data Description

The analysis relies on publicly available population statistics provided by the Swiss Federal Statistical Office (FSO). The dataset contains annual observations of the total permanent resident population in Switzerland, covering the period from 1860 to 2024. The long historical span makes the dataset suitable for time-series analysis, while also highlighting challenges related to structural changes over time.

For this project, the focus is placed on the aggregate national population level rather than disaggregated demographic components such as age, gender, or migration flows. This choice reflects the goal of evaluating forecasting methods under limited data availability, a common situation in applied settings.

Prior to modeling, the data were cleaned and transformed into a consistent time-series format. Population levels were converted to numerical values, sorted chronologically, and checked for missing observations. No external covariates were added in order to keep the comparison focused on time-series dynamics.

3.2 Baseline Model: Constant Growth

As a reference point, a constant growth model is used as a naive baseline. The model assumes that population grows at a fixed annual rate equal to the historical average growth observed over a recent period.

Formally, let P_t denote the population in year t. The baseline forecast is defined as:

$$P_{t+1} = P_t \times (1 + g)$$

where g is the average annual growth rate estimated from historical data between 1980 and the start of the test period. This baseline provides a simple benchmark against which more flexible models can be evaluated.

3.3 Linear Regression Model

The second approach uses a linear regression model to predict next-year population levels. Unlike the baseline, this model incorporates multiple explanatory variables derived from historical data.

The features used include:

- Current population level
- Annual population growth rate
- Lagged population values (one-year and two-year lags)

The target variable is the population level in year $t+1$. This formulation allows the model to capture linear relationships between recent trends and future population changes. Although simple, linear regression serves as a useful bridge between naive extrapolation and more structured time-series models.

Mathematical Formulation. Let y_t denote the population at year t. The target variable is the next-year population y_{t+1}. The linear regression model takes the form:

$$y_{t+1} = \beta_0 + \beta_1 \, y_t + \beta_2 \, g_t + \beta_3 \, y_{t-1} + \beta_4 \, y_{t-2} + \varepsilon_{t+1},$$

where:

- y_t is the current population level,
- $g_t = \frac{y_t - y_{t-1}}{y_{t-1}}$ is the annual growth rate,
- y_{t-1} and y_{t-2} are lagged population values,
- ε_{t+1} is an error term.

The model parameters $(\beta_0, \ldots, \beta_4)$ are estimated using Ordinary Least Squares (OLS).

3.4 Autoregressive (AR) Model

The third model is an autoregressive (AR) model, which predicts future population values based solely on past observations. Specifically, an AR(2) specification is used, meaning that the population in year $t+1$ is modeled as a linear function of population levels in years t and $t-1$.

This approach explicitly exploits temporal dependence in the data and is widely used in economic and demographic forecasting. The AR model can be viewed as a restricted version of the linear regression model that focuses exclusively on lagged values, thereby offering a clearer interpretation of time-series dynamics.

Mathematical Formulation. The autoregressive model of order 2 (AR(2)) assumes that the future population depends only on its two most recent values:

$$y_{t+1} = \alpha_0 + \alpha_1\, y_t + \alpha_2\, y_{t-1} + u_{t+1},$$

where:

- $(\alpha_0, \alpha_1, \alpha_2)$ are model coefficients,
- u_{t+1} is a white noise error term with zero mean.

This specification can be viewed as a restricted case of the linear regression model in which only lagged values are included as predictors.

3.5 Evaluation Strategy

Model performance is evaluated using a time-based train–test split to respect the chronological structure of the data. All observations prior to the year 2000 are used for training, while observations from 2000 onward form the test set.

Forecast accuracy is measured using the Root Mean Squared Error (RMSE), defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

RMSE penalizes large prediction errors and provides an intuitive measure in the same units as the population variable. By comparing RMSE values across models, it is possible to assess whether increased model complexity leads to improved out-of-sample performance.

4 Implementation and Reproducibility

All models are implemented in Python using standard scientific computing libraries, including `pandas`, `numpy`, `scikit-learn`, and `matplotlib`, `pyaxis`, `openpyxl`. The project is organized into modular source files for data loading, feature engineering, model estimation, and evaluation. All core functionality is organized within the `src/` directory. A single entry point (`main.py`) orchestrates the full pipeline, from data loading to model comparison and visualization. Running `python main.py` reproduces all tables and figures reported in this study, ensuring full reproducibility.

```
train = df[df["year"] < test_start_year]
test  = df[df["year"] >= test_start_year]
```

Listing 1: Time-based train/test split used in all models

5 Results

This section presents the empirical results obtained from the three forecasting approaches : the baseline constant-growth model, the linear regression model, and the autoregressive model. The comparison focuses on out-of-sample predictive performance using a common test period.

5.1 Experimental Setup

All models are trained on historical population data up to 1999 and evaluated on observations from 2000 onward. This time-based split ensures that forecasts are generated strictly using past information, avoiding any look-ahead bias.

Model performance is assessed using the Root Mean Squared Error (RMSE), which measures the average magnitude of prediction errors in population units. Lower RMSE values indicate better predictive accuracy.

5.2 Model Comparison

Table 1 summarizes the performance of the three models. The baseline model serves as a naive benchmark, while the linear regression and autoregressive models introduce increasing levels of structure and flexibility.

Table 1: Model comparison based on RMSE (test period)

Model	**Train RMSE**	**Test RMSE**
Baseline constant growth	–	329,711
Linear regression	12,835	21,846
AR(2) model	28,528	39,102

The baseline constant-growth model performs poorly in terms of predictive accuracy, exhibiting a test RMSE that is an order of magnitude larger than those of the other models. This indicates that assuming a fixed historical growth rate is insufficient to capture recent population dynamics.

The linear regression model achieves the lowest test RMSE among the three approaches. By incorporating current population levels, recent growth rates, and lagged values, the model is able to adapt more effectively to changes in population trends.

The autoregressive AR(2) model improves substantially over the baseline but underperforms relative to the linear regression. While the AR model captures temporal dependence, it appears less flexible than the regression model that combines multiple explanatory features.

5.3 Baseline Constant-Growth Forecast

Figure 1 displays the historical Swiss population series together with the baseline forecast obtained by extrapolating a constant average growth rate estimated from data since 1980.

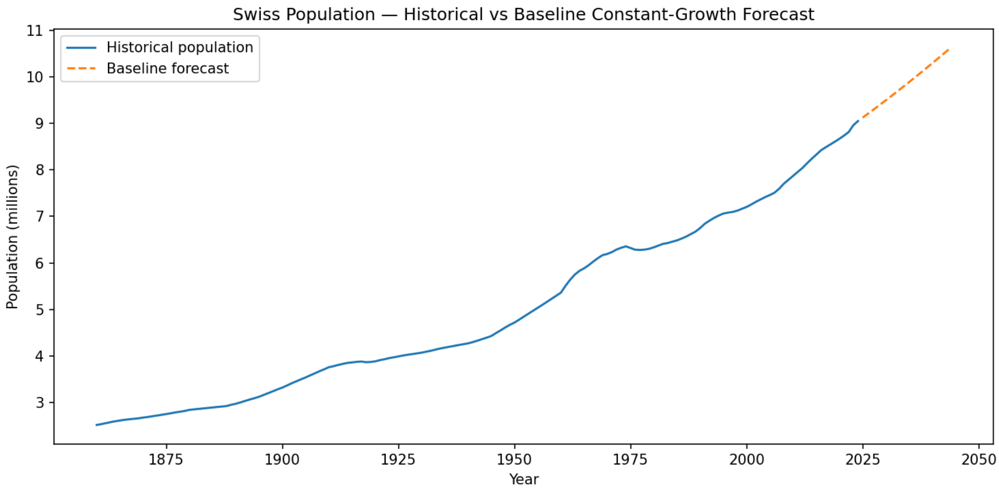

Figure 1: Historical population and baseline constant-growth forecast

The baseline model produces a smooth and deterministic trajectory that diverges progressively from observed population levels in the test period. This behavior highlights the limitations of assuming a fixed growth rate in the presence of demographic slowdowns and structural changes. As reflected in the large test RMSE reported earlier, this model fails to adapt to recent population dynamics.

5.4 Linear Regression Results

Figure 2 compares the actual population values with predictions from the linear regression model over the test period.

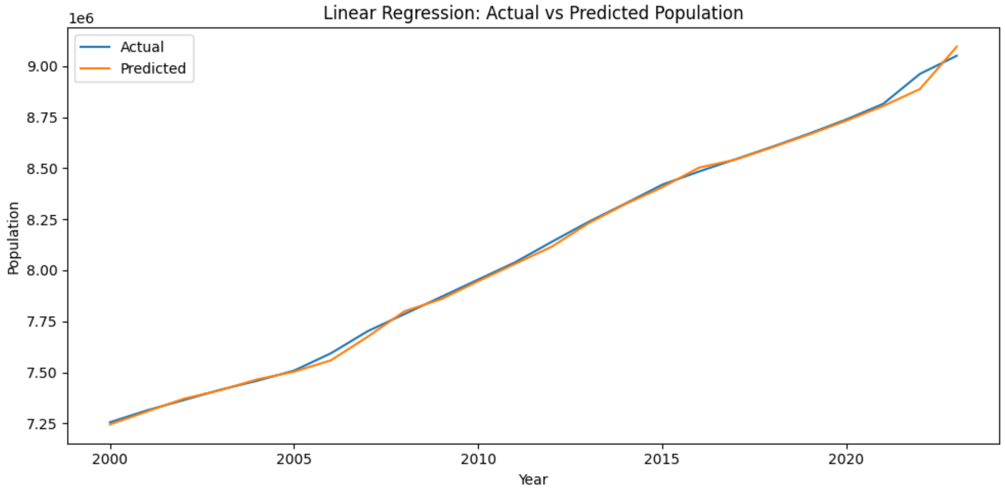

Figure 2: Linear regression: actual vs predicted population (test period)

The linear regression model closely tracks the observed population series, capturing both the level and short-term fluctuations. This improved performance is consistent with the substantially lower test RMSE reported in Table 1. By combining contemporaneous population levels, recent growth rates, and lagged values, the model benefits from a richer information set than purely autoregressive approaches.

Figure 3 presents the residuals over the test period.

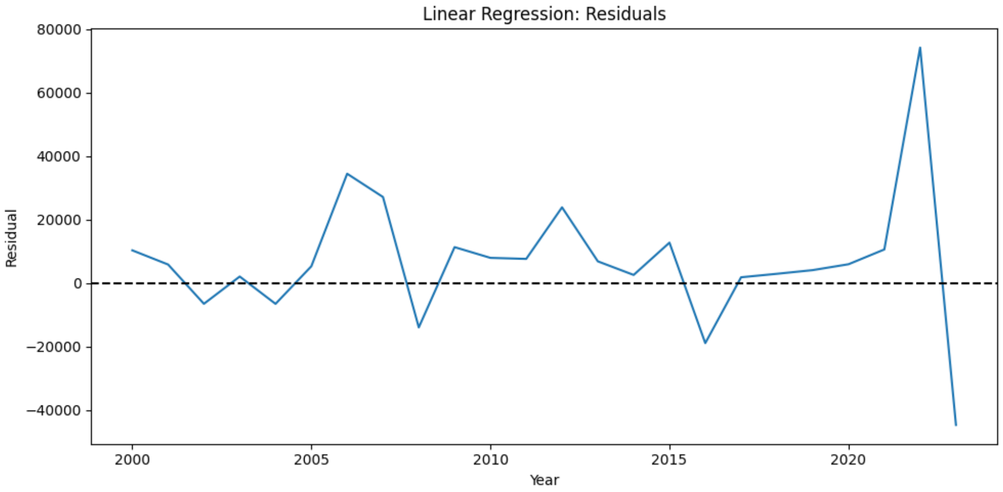

Figure 3: Linear regression residuals over the test period

Residuals are centered around zero with no obvious systematic pattern, suggesting that the model captures most of the predictable structure in the data.

5.5 Autoregressive Model Results

Figure 4 shows the test-period performance of the AR(2) model, which predicts population levels using only lagged values.

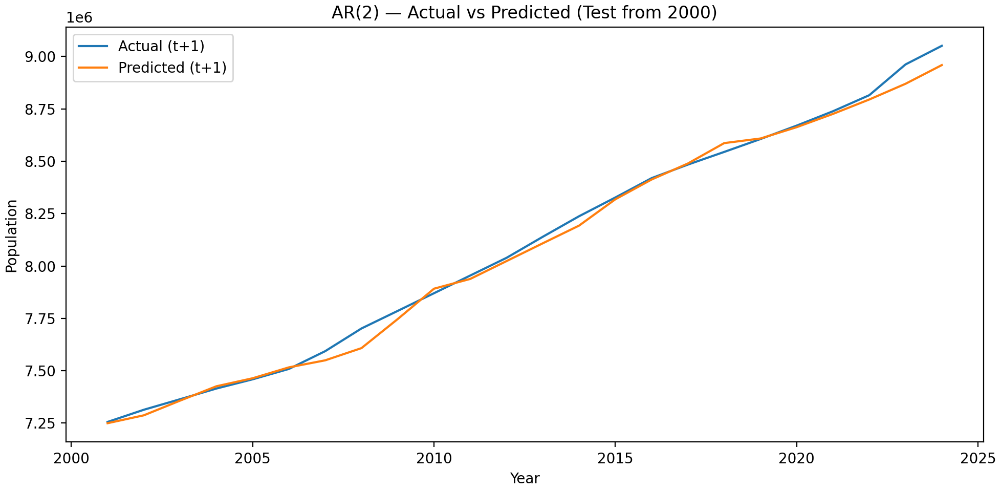

Figure 4: AR(2) model: actual vs predicted population (test period)

While the AR(2) model captures the overall upward trend, its predictions lag behind turning points and exhibit larger deviations from observed values compared to the linear regression model.

Residual diagnostics are shown in Figures 5 and 6.

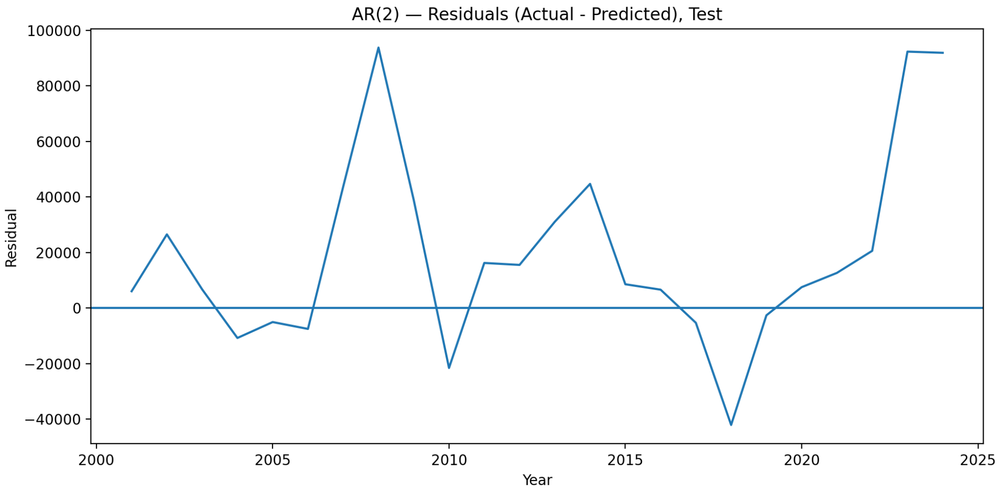

Figure 5: AR(2) residuals over the test period

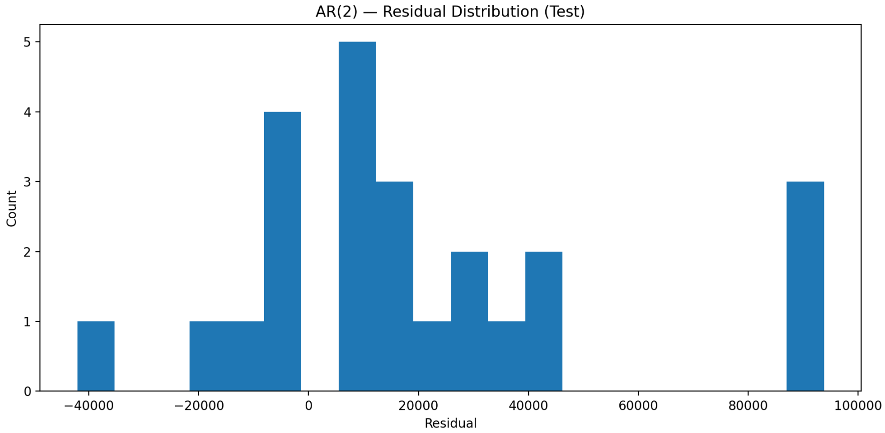

Figure 6: Distribution of AR(2) residuals (test period)

Residuals display higher variance and mild temporal structure, indicating that the autoregressive specification alone is insufficient to fully capture population dynamics.

6 Discussion

6.1 Baseline Model Limitations

The baseline constant-growth model performs poorly relative to more flexible approaches. By construction, it assumes that historical average growth rates persist unchanged into the future. While this assumption produces a smooth and interpretable forecast, it fails to account for demographic slowdowns, migration shocks, and structural changes observed in recent decades.

This limitation is reflected in the very large test RMSE, which is an order of magnitude higher than that of the data-driven models. The baseline forecast therefore serves primarily as a reference point rather than a competitive predictive model.

6.2 Linear Regression vs Autoregressive Models

Among the models considered, the linear regression approach achieves the best predictive performance on the test period. By combining contemporaneous population levels, recent growth rates, and lagged population values, the model leverages a richer information set than the purely autoregressive specification.

In contrast, the AR(2) model relies exclusively on past population levels. While it successfully captures the long-run upward trend, it responds more slowly to changes in growth dynamics. This lagging behavior is visible in both the actual-versus-predicted plots and the residual diagnostics.

6.3 Error Structure and Predictive Stability

Residual analysis provides additional insights into model behavior. The linear regression residuals are tightly centered around zero with no obvious time-dependent structure, suggesting that most predictable variation has been captured.

The AR(2) residuals, while still reasonably well-behaved, exhibit larger dispersion and occasional spikes. This indicates that purely autoregressive dynamics may be insufficient to fully explain short-term population fluctuations, especially in periods of accelerated growth or slowdown.

7 Conclusion and Future Work

7.1 Conclusion

Empirical results clearly demonstrate that the baseline constant-growth model performs poorly in a modern forecasting context. Its strong structural assumptions lead to large prediction errors and an inability to adapt to changing demographic trends. In contrast, the linear regression model achieves the best predictive performance, substantially reducing test RMSE by incorporating lagged population levels and recent growth dynamics.

The autoregressive AR(2) model performs better than the baseline but remains less accurate than the regression approach. While AR models capture long-run trends, they react slowly to turning points and structural changes. Overall, the results highlight the importance of combining demographic momentum with short-term growth information when forecasting population dynamics.

From a methodological perspective, this project demonstrates how classical time-series ideas and simple machine learning models can be rigorously compared using transparent evaluation metrics and out-of-sample testing.

7.2 Limitations

Several limitations must be acknowledged. First, the analysis relies on a relatively short annual time series, which constrains the complexity of models that can be reliably estimated. Second, only aggregate population data are used, excluding potentially relevant covariates such as migration flows, fertility rates, or economic conditions. Finally, the models are designed for short- to medium-term forecasting and should not be interpreted as long-run demographic projections.

7.3 Future Work

Future extensions of this project could proceed along several dimensions. Incorporating additional demographic variables, such as births, deaths, and net migration, would allow for richer multivariate models. More advanced time-series methods, including state-space models or Bayesian approaches, could improve uncertainty quantification. Finally, extending the framework to regional or age-specific population forecasts would provide more granular insights relevant for policy and planning.

References

[1] Anthropic. *Claude*. `https://www.anthropic.com/claude`. Large language model used for coding assistance and explanations. 2024.

[2] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. 3rd. OTexts, 2021. URL: `https://otexts.com/fpp3/`.

[3] Gareth James et al. *An Introduction to Statistical Learning*. 2nd. Springer, 2021.

[4] OpenAI. *ChatGPT*. `https://chat.openai.com`. Large language model used for debugging, code structuring, and writing assistance. 2024.

[5] Swiss Federal Statistical Office. *Permanent resident population by age, sex and year*. `https://www.bfs.admin.ch`. Accessed: November 2025. 2024.

A Additional Results

A.1 Residual Diagnostics

This appendix presents additional diagnostic plots for the linear regression and autoregressive models. These figures support the main results by illustrating the distribution and temporal structure of prediction errors.

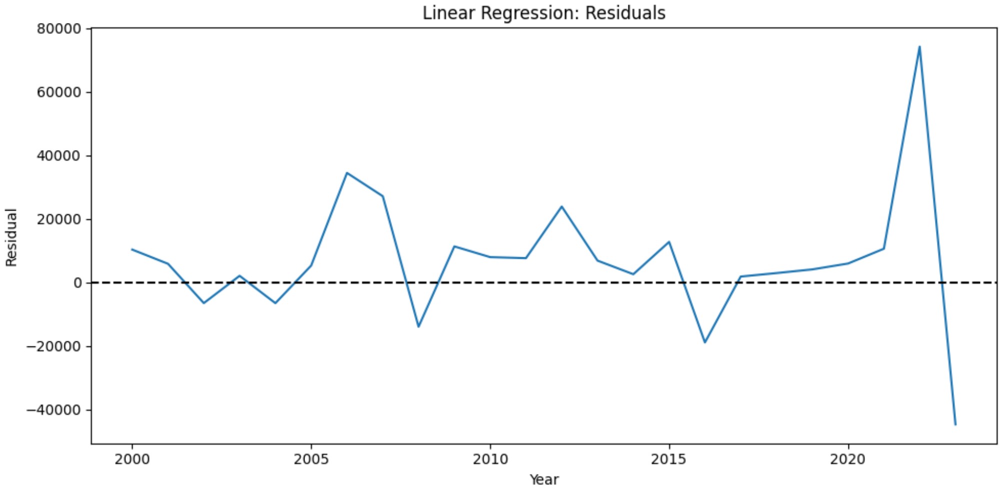

Figure 7: Linear regression residuals over the test period

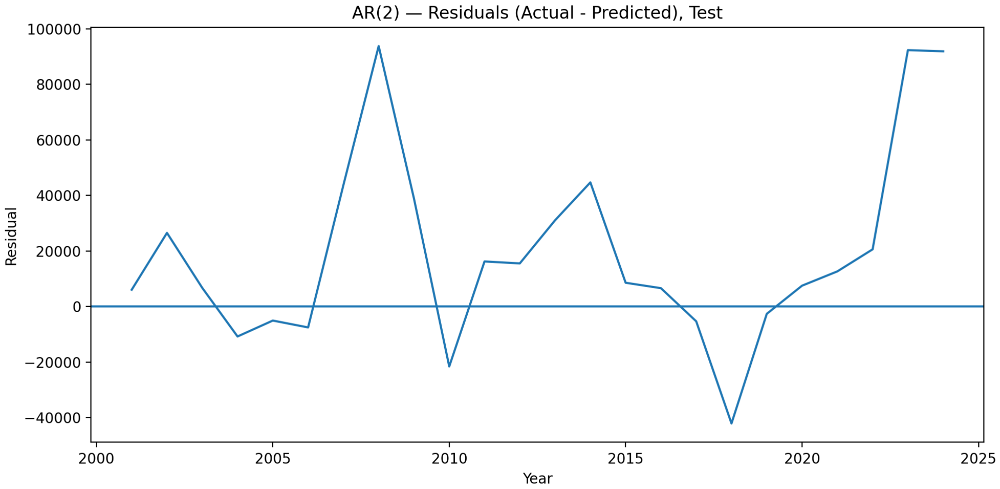

Figure 8: AR(2) residuals over the test period

A.2 Lag Sensitivity Analysis

Figure 9 illustrates the sensitivity of the AR model's test RMSE to the choice of lag order.

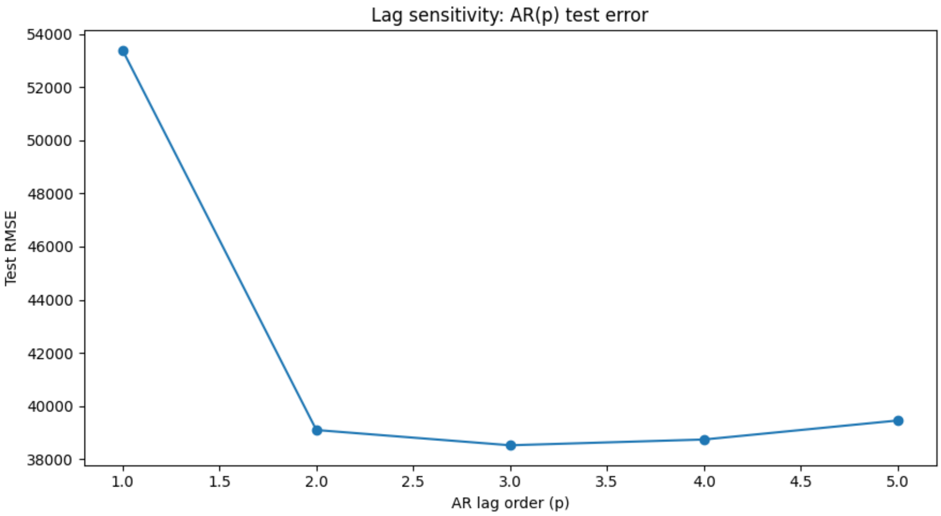

Figure 9: Test RMSE of AR(p) models for different lag orders

B Code Repository

GitHub Repository: `https://github.com/yassir-d/Population_growth/tree/main`