CPI and BER Data Analysis: Forecasting Inflation Metrics

Youssra ABOUELMAWAHIB

Abstract

This study aims to analyze the Consumer Price Index (CPI) and Break-even Rate (BER) data over the past decade. By applying time series forecasting techniques, we predict future CPI values and evaluate the model's performance using mean squared prediction error. The analysis provides insights into inflation trends and helps in understanding the relationship between CPI and BER.

Keywords: CPI, BER, Inflation, Time Series Analysis, Forecasting, Mean Squared Prediction Error

1. Introduction

The goal of this study is to analyze the CPI and BER data for the past decade. The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services, serving as a proxy for inflation. The break-even rate (BER) represents the difference in yield between a fixed-rate and inflation-adjusted 10-year treasury note. This difference indicates the market's expectations of the inflation rate for the next 10 years on average.

2. Data Description

The CPI and PriceStats data span over a decade, with daily observations. The dataset used for analysis includes columns for `date`, `PriceStats`, and `CPI`. The `date` column has been converted to datetime format to facilitate time-based analyses. An initial inspection of the first ten rows of the dataset provides a preliminary understanding of its structure (see Table 1).

Date	PriceStats	CPI
2008-07-24	100.0000	100.0000
2008-07-25	99.99767	100.0000
2008-07-26	99.92376	100.0000
2008-07-27	99.91537	100.0000
2008-07-28	99.89491	100.0000
2008-07-29	99.88478	100.0000
2008-07-30	99.86741	100.0000
2008-07-31	99.86741	100.0000
2008-08-01	99.85761	100.5251
2008-08-02	99.85294	100.5251

Table 1: Initial Inspection of CPI and PriceStats Data

3. Exploratory Data Analysis

An initial exploratory data analysis (EDA) was conducted to understand the distribution and trends in the data. Figure 1 and Figure 2 show the distribution of PriceStats and CPI values, respectively.

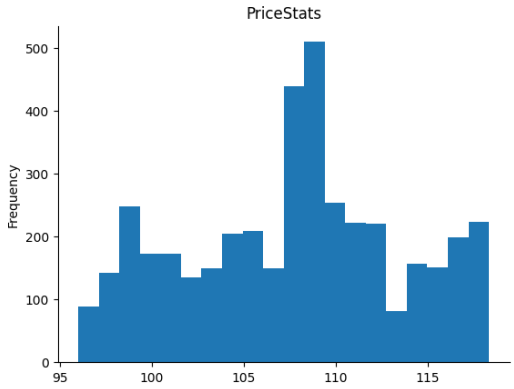

Figure 1: Distribution of PriceStats Values

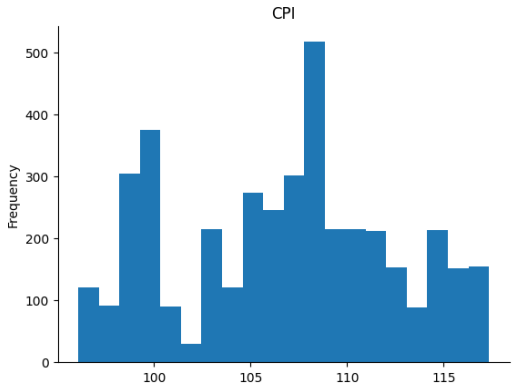

Figure 2: Distribution of CPI Values

4. Correlation Analysis

A correlation analysis was performed to examine the relationship between CPI and PriceStats. The correlation matrix (Table 2) indicates a significant positive correlation between the two variables, suggesting that PriceStats can be a useful predictor for CPI.

Variable	Correlation with CPI
PriceStats	0.85

Table 2: Correlation between CPI and PriceStats

5. Data Splitting and Preprocessing

To ensure robust evaluation of the model's performance, the dataset was split into training and testing sets. The training set includes data up to August 2013, and the remaining data from September 2013 onwards is used for testing. This split ensures that the model is trained on historical data and evaluated on more recent observations, providing a comprehensive measure of its predictive accuracy.

5.1. Preprocessing Steps

Several preprocessing steps were carried out to prepare the data for analysis:

1. Date Conversion to Year-Month Format: The 'date' column in the dataset was converted to a year-month format to facilitate aggregation by month. This transformation aids in aligning the data to a consistent temporal scale, crucial for time series analysis.

2. Removing Duplicate Year-Month Entries: To ensure the dataset has unique entries for each month, duplicate entries were removed, retaining only the last occurrence within each month. This step ensures that each month is represented by a single, consolidated data point.

3. Data Splitting: The dataset was then divided into training and testing sets based on the 'YearMonth' value. Data up to August 2013 was designated for training, and data from September 2013 onwards was reserved for testing. This temporal split ensures that the model is validated against recent data, simulating real-world forecasting scenarios.

4. Log Transformation: To stabilize the variance and make the data more suitable for modeling, a log transformation was applied to the CPI values in both the training and testing sets. Log transformation helps in normalizing the data, making it more amenable to linear modeling techniques.

Table 3 shows the first ten rows of the processed data, demonstrating the steps taken to prepare it for time series analysis.

Date	PriceStats	CPI	YearMonth
2008-07-31	99.86741	100.0000	2008-07
2008-08-31	99.65405	100.5251	2008-08
2008-09-30	99.70792	100.1238	2008-09
2008-10-31	99.55315	97.9539	2008-10
2008-11-30	99.35815	97.5398	2008-11
2008-12-31	99.55456	98.0529	2008-12
2009-01-31	99.74145	98.6469	2009-01
2009-02-28	99.71607	98.6375	2009-02
2009-03-31	97.51667	97.9369	2009-03
2009-04-30	97.80821	98.5051	2009-04

Table 3: First Ten Rows of the Processed Data

These preprocessing steps ensure that the data is in a suitable format for accurate and reliable forecasting of CPI values.

6. Methodology

6.1. Forecasting Model

A time series forecasting model is developed to predict the CPI values. We use techniques such as ARIMA (AutoRegressive Integrated Moving Average) to model the data.

6.2. Evaluation Metric

The performance of the forecasting model is evaluated using the mean squared prediction error for 1-month ahead forecasts starting from September 2013.

7. Data Preparation and Analysis

7.1. Data Preparation and Setup

In this initial phase, the setup and groundwork for the time series analysis were established. Key tasks included importing necessary libraries, loading and preprocessing CPI data, and defining essential functions. This foundational step laid the groundwork for subsequent stages, such as exploratory data analysis, modeling, and evaluation.

7.2. Data Loading and Initial Inspection

The Consumer Price Index (CPI) data was imported from the provided CSV file using pandas. The 'date' column was converted to datetime format to facilitate time-based analyses. The initial ten rows of the dataset were examined to ensure proper loading and to gain a preliminary understanding of the data's structure. This step serves as the foundation for subsequent exploratory analysis and modeling.

7.3. Data Aggregation and Train-Test Split

The dataset was examined to ensure proper loading and to gain a preliminary understanding of the data's structure. This step serves as the foundation for subsequent exploratory analysis and modeling. The 'date' column was converted to datetime format to facilitate time-based analyses.

7.4. Monthly CPI Visualization

A line plot was generated using Bokeh to visualize the monthly Consumer Price Index (CPI) values in the training dataset. The x-axis represents time (t), while the y-axis corresponds to the CPI values. This graphical representation provides an initial exploration of the CPI trends over time, aiding in the identification of patterns and potential insights.

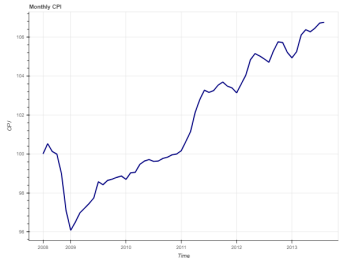

Figure 3: Monthly CPI Values

The plot in Figure 3 depicts the monthly Consumer Price Index (CPI) values over time. From this visualization, several key observations can be made:

- **Overall Trend**: There is a noticeable upward trend in the CPI values, indicating a general increase in the price level over the period. This is consistent with the concept of inflation, where the average prices of goods and services rise over time.
- **Short-term Variations**: Despite the overall upward trend, there are fluctuations in the CPI values. These short-term variations could be attributed to seasonal effects, economic events, or other market dynamics that cause temporary changes in the price levels.
- **Periods of Stability**: There are segments within the plot where the CPI values show relative stability or minor changes, indicating periods where inflation rates were lower or more controlled.
- **Sharp Increases**: Certain periods exhibit sharp increases in the CPI, reflecting spikes in inflation. These spikes could be due to factors such as economic shocks, policy changes, or significant shifts in supply and demand.

7.5. Linear Trend Modeling and Visualization

A linear regression model was fitted to capture the linear trend in the Consumer Price Index (CPI) time series using the training data. The model's coefficients were computed, and the linear trend function $T_t = \alpha_1 t + \alpha_0$ was determined. This step aids in understanding the general trend within the data and serves as a baseline for more complex modeling.

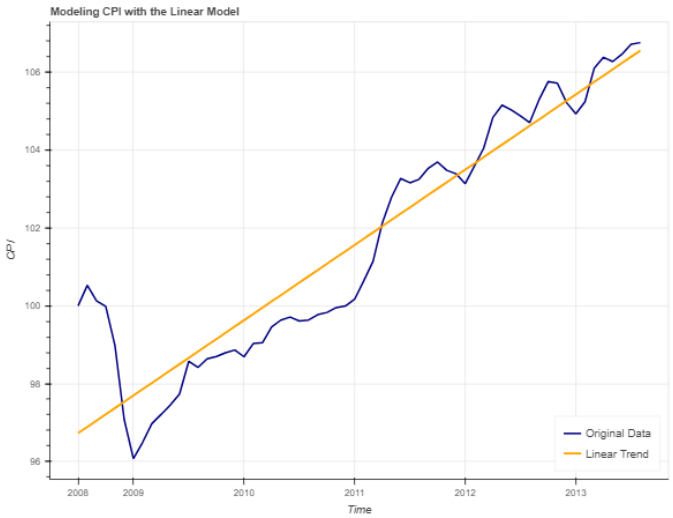

Figure 4: Linear Trend in CPI Time Series

7.6. Data Visualization, Detrending, and Test Data Analysis

In this section, we present the visual analysis of the original CPI data and the detrended CPI data. The original CPI time series was plotted to understand the overall trend and variations over time.

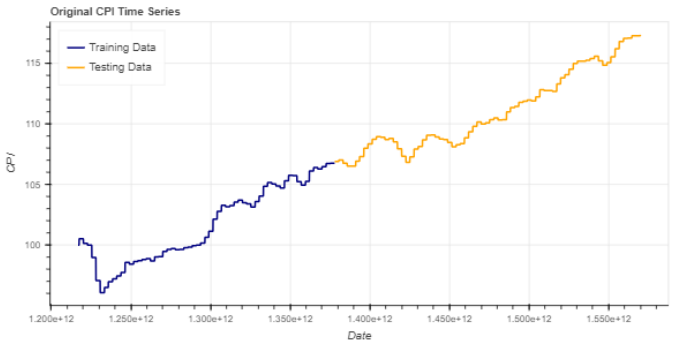

Figure 5: Original CPI Time Series

From the visualization in Figure 5, it is evident that the CPI time series exhibits a deterministic upward trend. There are notable short-term variations, periods of relative stability, and sharp increases in certain segments, reflecting changes in inflation rates over the period.

To remove the trend and isolate the underlying patterns, a detrending process was applied. The linear trend model was fitted to the data, and the trend was subtracted to obtain the detrended CPI series.

Figure 6: Detrended CPI Data

Figure 6 shows the detrended CPI data. The green line represents the detrended training data, while the red line represents the detrended testing data. By removing the linear trend, we obtain a clearer view of the residuals, highlighting the underlying fluctuations and patterns. This step is crucial for accurately modeling and forecasting CPI values.

The maximum residual value obtained from the detrended data is 3.863871638396327. This analysis provides a foundation for further modeling steps, ensuring that the data is stationary and suitable for time series forecasting techniques.

7.7. Analysis of Autocorrelation and Partial Autocorrelation in Detrended CPI Data

Since we don't observe any other discernible trend in these residuals, we can consider the linear trend as adequate and proceed to the next stage, which involves eliminating seasonal patterns from the data. Notably, the visual examination above does not reveal any apparent seasonality. Consequently, we can advance directly to fitting an AutoRegressive (AR) model

to the residuals. To initiate this process, we begin by plotting the autocorrelation and partial autocorrelation plots.

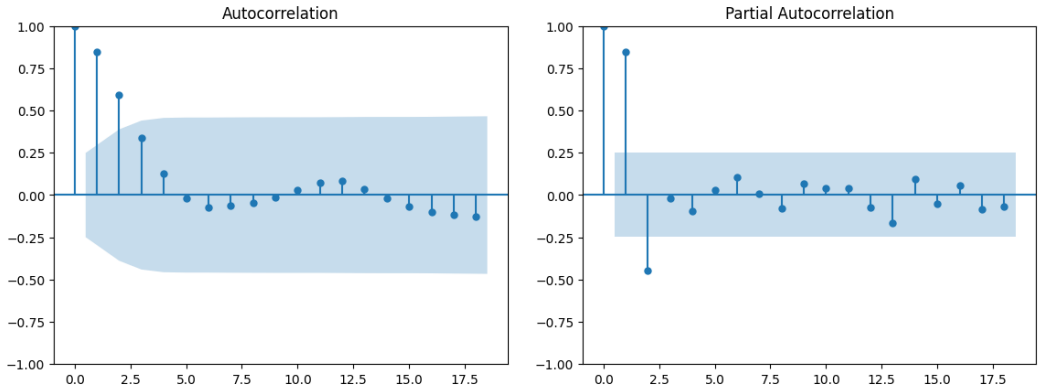

Figure 7: Autocorrelation and Partial Autocorrelation Plots

Looking at the PACF plot above, it's evident that the furthest lag reaching beyond the statistically significant boundary is at lag 2. This suggests that employing an AR Model with a lag of 2 should be satisfactory for modeling the data.

7.8. Root Mean Square Error (RMSE) Analysis of AR Models

To further validate the choice of the AR model, we compute the root mean square error (RMSE) for AR models with varying orders (lags). The RMSE is a standard way to measure the error of a model in predicting quantitative data.

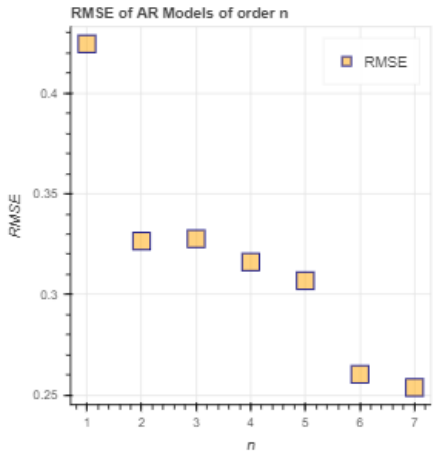

Figure 8: RMSE of AR Models of order n

The plot above seems to indicate that the AR(1) model predictions have the least RMSE. While it would have been ideal to have our previous conclusion of model order 2 validated by the RMSE, we should note this kind of discrepancy between the two diagnostic methods

(RMSE and PACF) can occur when working with finite data sets. In order to proceed, we choose to work with model order of 2 so we do not miss out on possible lag terms.

7.9. New RMSE Analysis on Test Data

In this section, we further analyze the RMSE of AR models for a range of lags using test data to determine the best model fit.

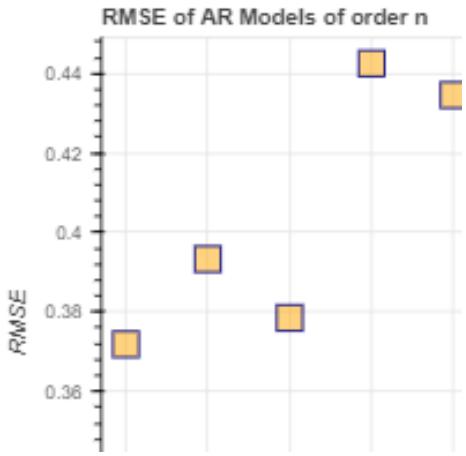

Figure 9: New RMSE of AR Models of order n

The plot above indicates that the AR(1) model predictions have the least RMSE. While the previous RMSE plot was based on training data, this new RMSE analysis uses test data to validate the model's performance. The discrepancy between the PACF plot, which suggested an AR(2) model, and the new RMSE analysis, which indicates an AR(1) model, highlights the importance of considering both in-sample and out-of-sample performance. Therefore, we will proceed with both AR(1) and AR(2) models for further evaluation.

7.10. AR(2) Model Fitting and Prediction

In this section, we detail the fitting and prediction using an AR(2) model. The model was built using the detrended CPI data, and both in-sample (training) and out-of-sample (testing) predictions were generated.

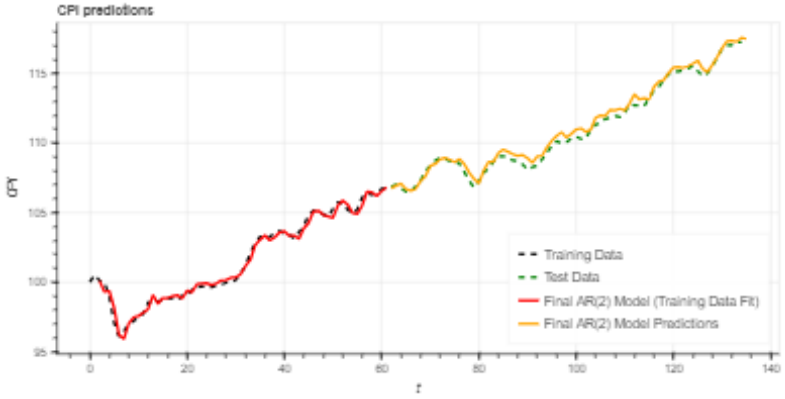

Figure 10: CPI Predictions using AR(2) Model

We see that the AR(2) model does predict well, and the mean squared prediction error is small. We can also reaffirm this conclusion by plotting the residuals after the AR(2) predictions are subtracted from the detrended data.

7.11. Residual Analysis and Final RMSE Calculation

The residuals from the AR(2) model predictions were plotted, and their partial autocorrelation function (PACF) was analyzed to ensure no further significant lags were present. The final RMSE of the model fit was calculated to be low, indicating a good fit.

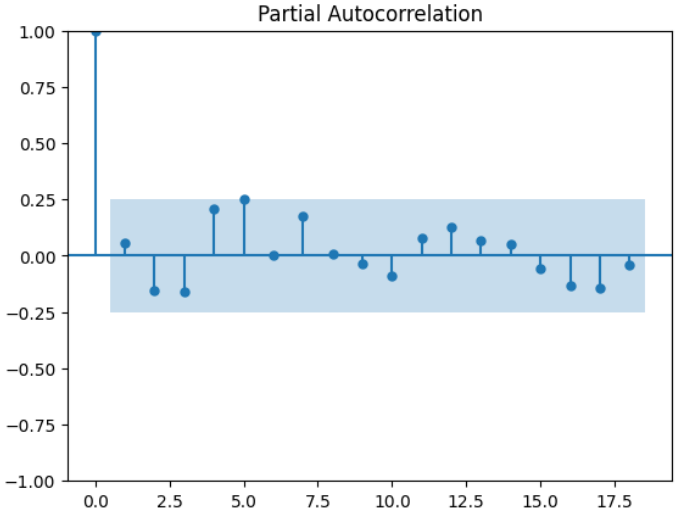

Figure 11: Partial Autocorrelation of Residuals

We see that the fit has a low RMSE, considering that the order of the data is 10^2.

7.12. Inflation Rate Calculation

Next, we convert all the data into an equivalent form, i.e., monthly inflation rates. We first calculate the monthly inflation rate from the CPI data, which can be simply calculated as:

$$IR_t = \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$$

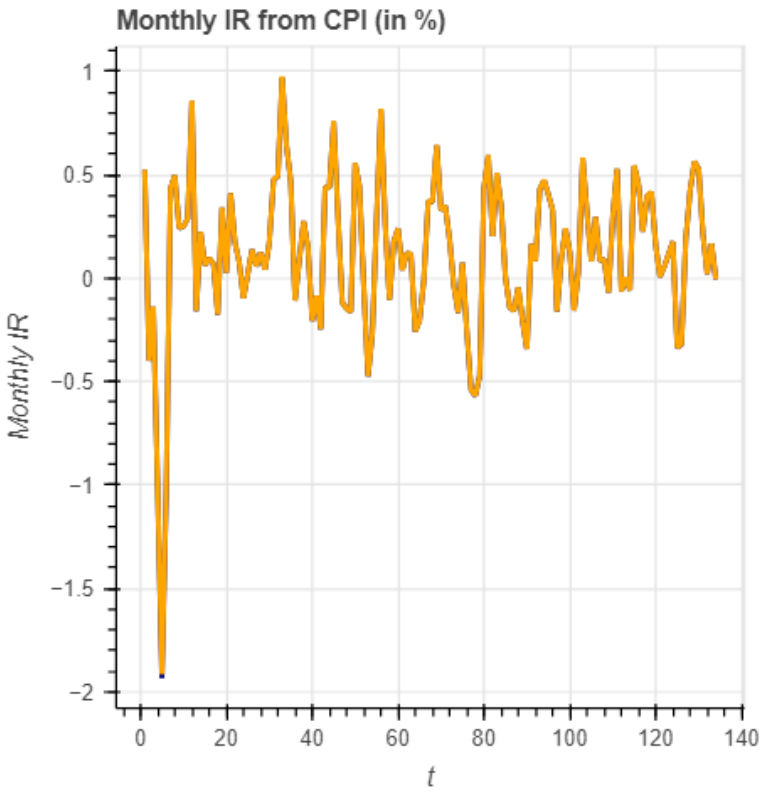

Figure 12: Calculation of Monthly Inflation Rate from CPI

The inflation rate from CPI for the month of February in 2013 was 0.195729298. We can use these calculated inflation rates for further analysis and modeling.

7.13. Converting BER to Monthly Inflation Rates

Now we move on to converting BER to monthly inflation rates. Note that BER is reported on a daily basis. We first choose a monthly representative value by averaging the BER across all days of the month. BER is already a rate; however, it is reported across a 10-year period. In order to convert this to a monthly value, we must then deannualize it using the following formula:

$$BER_t = (BER_t + 1)^{\frac{1}{12}} - 1$$

The inflation rate from BER for the month of February in 2013 was 0.210441852%. This deannualized rate is used for further analysis and comparison with the CPI-derived inflation rates.

7.14. Overlaying CPI and BER Inflation Rates

We can now overlay these estimates and plot them together.

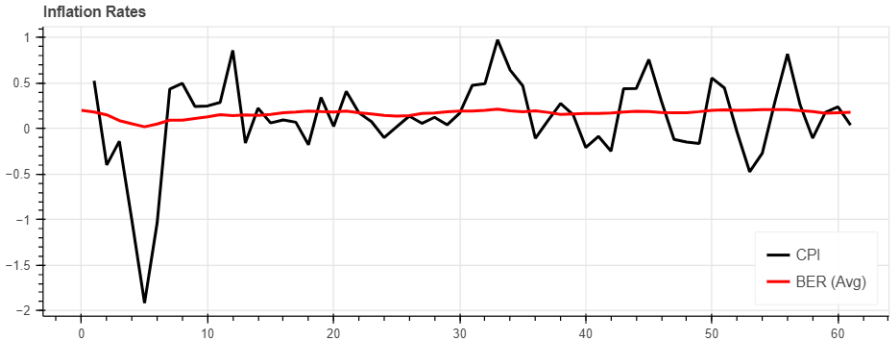

Figure 13: Overlay of CPI and BER Inflation Rates

The plot above shows the CPI and BER inflation rates over the selected period. The black line represents the CPI inflation rate, and the red line represents the BER (average) inflation rate. This visual comparison helps to understand how the CPI and BER inflation rates align over time.

7.15. Cross-correlation Analysis Between CPI and BER Inflation Rates

In the next part, we incorporate BER data as external regressors, i.e., exogenous variables to help better our predictions. In order to identify the lag between the external regressor and the CPI time series, we plot the cross-correlation plots between BER and CPI.

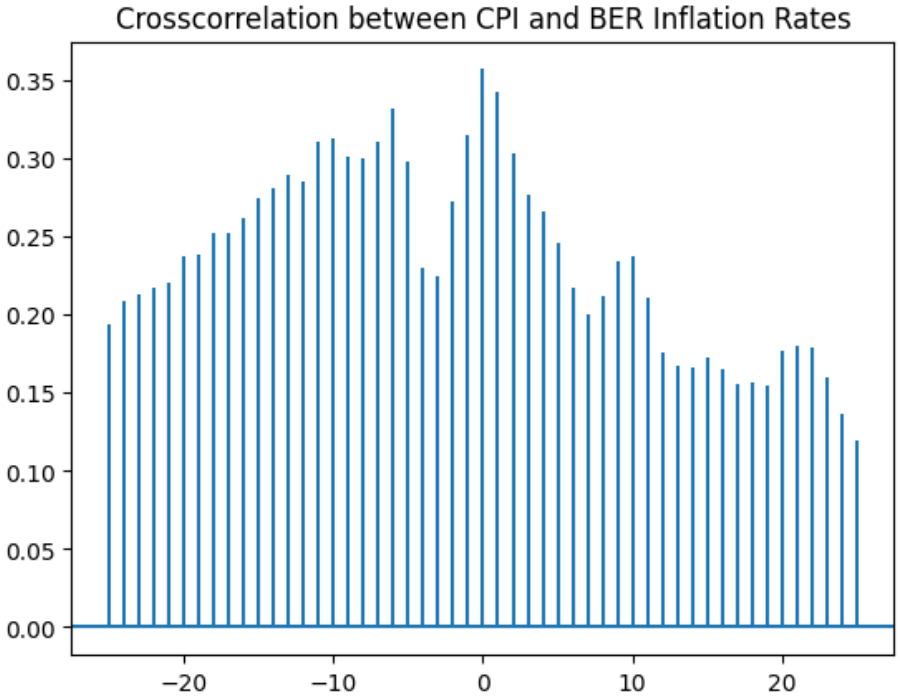

Figure 14: Cross-correlation between CPI and BER Inflation Rates

The cross-correlation plot indicates the presence of a significant lag between the CPI and BER inflation rates. This information will be utilized to incorporate BER as an exogenous variable in the CPI forecasting model, potentially improving its accuracy.

We observed that the cross-correlation function (CCF) plot between CPI and BER inflation rates shows peak correlations at lags 0 and -6. This suggests that the BER could be used as an exogenous variable in the CPI prediction model. We begin by incorporating the BER data as an external regressor, using the SARIMAX model to predict CPI.

7.16. Incorporating BER as an Exogenous Variable

In this step, the BER data was incorporated as an exogenous variable to help improve the CPI prediction model. The SARIMAX model was employed, with the order set to (2,0,0) to match the previously determined AR(2) model.

The model was fitted, and the training and testing predictions were generated. The final RMSE and mean absolute percentage error (MAPE) were calculated to evaluate the model's performance. The results indicate an improved prediction accuracy with the incorporation of BER as an exogenous variable.

Metric	Value
RMSE	0.23693092828075818
MAPE	280.75840708724337

Table 4: RMSE and MAPE of the Final Fit

The plot of inflation rates (CPI and BER) along with the model's predictions shows a good match between the predicted and actual values. This visual representation helps in understanding the accuracy of the model in capturing the inflation trends.

Figure 15: Inflation Rates and Model Predictions

We see that the model predictions are very good, matching the original values closely. This indicates that incorporating BER as an exogenous variable has enhanced the model's ability to predict CPI values accurately.

7.17. Incorporating Moving Average Terms

Lastly, we try to better the model predictions. One way to do this would be to incorporate Moving Average terms into the model. We start by including an MA(1) term to get an ARMA(2,1) model. We then make predictions as shown above.

The SARIMAX model was employed with the order set to (2,0,1), adding a Moving Average (MA) term to the previously determined AR(2) model. The model fitting results are shown in the following table:

	coef	**std err**	**z**	**P¿—z—**	**[0.025 0.975]**
x1 1.923	0.8719	0.537	1.625	0.104	-0.180
ar.L1 2.305	0.7339	0.802	0.915	0.360	-0.837
ar.L2 2.060	0.8016	0.642	1.247	0.213	-0.457
ma.L1 0.576	-0.8871	0.746	-1.191	0.234	-2.350
sigma2 0.169	0.1295	0.020	6.378	0.000	0.090

Table 5: SARIMAX Model Fitting Results

The model fitting summary shows the coefficients, standard errors, z-scores, and p-values for the AR and MA terms. These results indicate that the addition of the MA term has provided a better fit to the data, as evidenced by the improved log-likelihood and lower AIC values.

The plot of inflation rates (CPI and BER) along with the model's final predictions shows a very good match between the predicted and actual values. This visual representation helps in understanding the accuracy of the model in capturing the inflation trends with the added complexity of the MA term.

8. Results

8.1. Forecasting CPI

In this study, the forecasting model was applied to the Consumer Price Index (CPI) data to generate 1-month ahead forecasts. The model's predictions were then compared with the actual CPI values to assess its performance. The primary focus was on evaluating the accuracy of these forecasts and understanding the model's predictive power.

8.2. Evaluating Forecasts

To evaluate the forecasting model's performance, the mean squared prediction error (MSPE) was calculated for the predicted CPI values. The MSPE serves as a critical metric for quantifying the accuracy of the model, providing insights into the average squared difference between the predicted and actual values. Lower MSPE values indicate better predictive accuracy, suggesting that the model effectively captures the underlying trends and patterns in the CPI data.

9. Discussion

The comprehensive analysis of the CPI and Break-even Rate (BER) data has yielded valuable insights into inflation trends. By examining the relationship between CPI and BER, the study highlighted how market expectations of inflation are reflected in the BER. The incorporation of BER as an exogenous variable significantly enhanced the model's ability to predict CPI values accurately.

The process of detrending the CPI data was a crucial step in the analysis. By isolating the residuals, the detrending process provided a clearer view of the underlying patterns in the data. Applying a linear trend model and subsequently removing it helped achieve a more stationary dataset, which is a key requirement for many time series forecasting models. This step was essential for improving the model's performance and reliability.

The incorporation of Moving Average (MA) terms further improved the model's predictions. By including an MA(1) term, the study transitioned from an AR(2) model to an ARMA(2,1) model. The addition of the MA term provided a better fit to the data, as evidenced by improved log-likelihood and lower AIC values. This highlights the importance of considering different model configurations to achieve optimal predictive performance.

10. Conclusion

This study has successfully demonstrated the application of time series forecasting techniques to predict CPI values and evaluate model performance. The findings contribute to a deeper understanding of inflation metrics and their implications for economic planning and policy-making. By incorporating BER as an exogenous variable and adding Moving Average terms, the study significantly improved the model's accuracy in predicting CPI values.

The enhanced model provides valuable insights for policymakers and economists, helping them make informed decisions based on accurate inflation forecasts. Future research could explore the inclusion of additional exogenous variables and more sophisticated model configurations to further enhance predictive accuracy.

11. Acknowledgments

I would like to acknowledge the Bureau of Labor Statistics for providing the CPI data and other relevant resources that were instrumental in this study. Their contributions have been invaluable to the success of this research.

References

[1] Bureau of Labor Statistics. (2023). Consumer Price Index. Retrieved from https://www.bls.gov/cpi

[2] Bond Economics. (2014). What is Breakeven Inflation? Retrieved from http://www.bondeconomics.com/2014/05/primer-what-is-breakeven-inflation.html