Forecasting Unemployment Trends: A Comparative Time Series Analysis of Colombia and the United States

Aye Nyein Thu, Mazhar Bhuyan, Yuqi Yang, Jisup Kwak

2025-04-24

Contents

0.1 1. Introduction, Motivation, Relevance, and Objectives

Unemployment remains one of the most visible and impactful indicators of economic well-being, directly shaping the lives of citizens and the priorities of governments. This report provides a comparative analysis of unemployment rates in the United States and Colombia, offering insight into how economic shocks and labor market structures influence short-term employment outcomes.

The United States, with its deep capital markets and relatively flexible labor regulations, often exhibits smooth unemployment cycles responsive to monetary policy. In contrast, Colombia faces more structural unemployment challenges, including higher informality and vulnerability to global commodity price swings.

From a development policy perspective, analyzing and forecasting these trends can inform interventions in education, training, and employment services. This analysis leverages publicly available ILO datasets and advanced time series models to forecast future labor market trajectories and highlight structural contrasts between the two economies.

The main objectives of the report are:

- To visualize and describe the historical unemployment trends in the US and Colombia
- To detect and treat anomalies such as missing values and outliers
- To compare the forecasting performance of multiple time series models
- To produce robust 12-month forecasts using the best-performing model(s)

0.2 2. Data Description and Preprocessing

The dataset used in this study is sourced from the International Labour Organization (ILO) and contains monthly unemployment data for the United States and Colombia, segmented by sex and age group. To conduct a comparative and consistent analysis, we have constructed a unified time series using the age-group data by summing unemployment counts and calculating a weighted average percentage.

Since percentages are normalized and directly comparable across countries and time, the main variable used for forecasting is the **total unemployment percentage (Total.Per)**. All data were checked for missing values and structural inconsistencies before modeling.

Table 1: Summary Statistics: United States and Colombia

	United States					Colombia				
Variable	US_Mean	US_SD	US_Min	US_Max	US_N	Col_Mean	Col_SD	Col_Min	Col_Max	Col_N
Age15to24.Per	12.008681	3.488402	5.7	26.9	288	21.563768	3.441774	15.1	34.4	276
Age25above.Per	4.761458	1.774522	2.6	12.8	288	8.460145	2.002818	5.3	17.9	276
Female.Per	5.561458	1.876113	2.9	15.7	288	13.959420	2.916427	9.1	25.5	276
Male.Per	5.949306	2.148008	3.2	13.3	288	8.736594	1.947194	5.3	17.9	276
Total.Per	5.765972	1.982789	3.1	14.4	288	10.914130	2.277091	7.1	20.1	276
Age15to24.Thou	2584.030208	739.483180	1232.8	4869.8	288	852.173188	144.136962	551.9	1265.4	276
Age25above.Thou	6368.664931	2360.277964	3711.0	17686.9	288	1494.322101	399.159524	962.7	3267.4	276
Female.Thou	4034.548611	1344.285566	2243.2	11494.3	288	1253.461594	237.993991	864.1	2193.1	276
Male.Thou	4918.150000	1734.999926	2842.0	11009.8	288	1093.032971	235.619647	698.0	2223.8	276
Total.Thou	8952.700347	3016.384542	5146.1	22504.1	288	2346.493116	461.038716	1641.0	4373.4	276

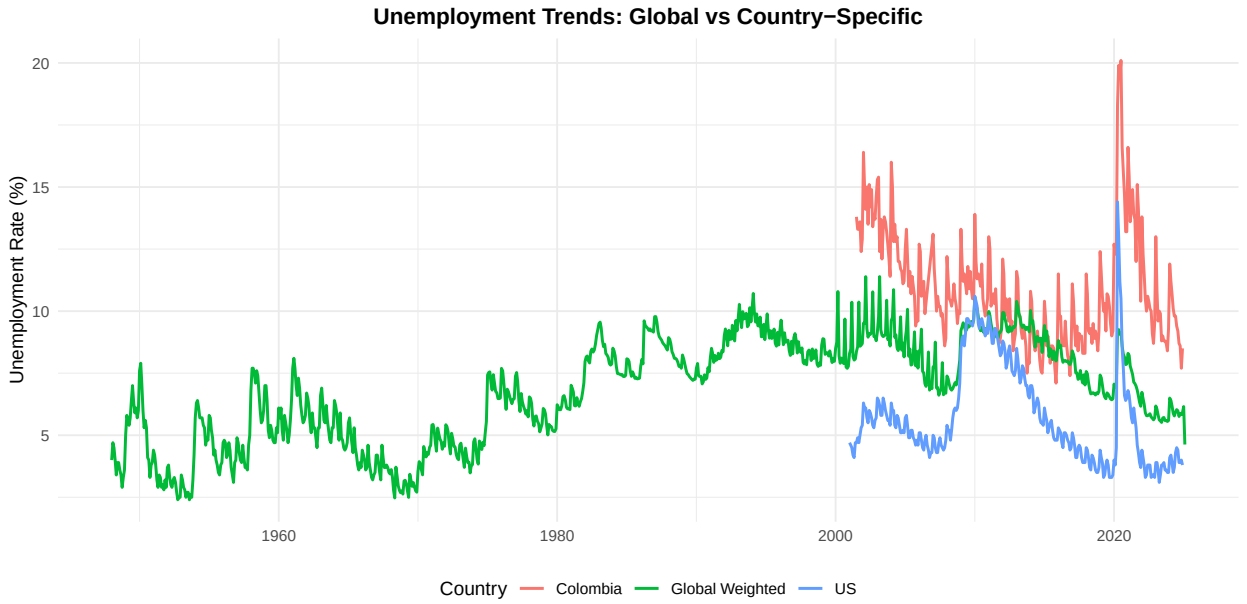


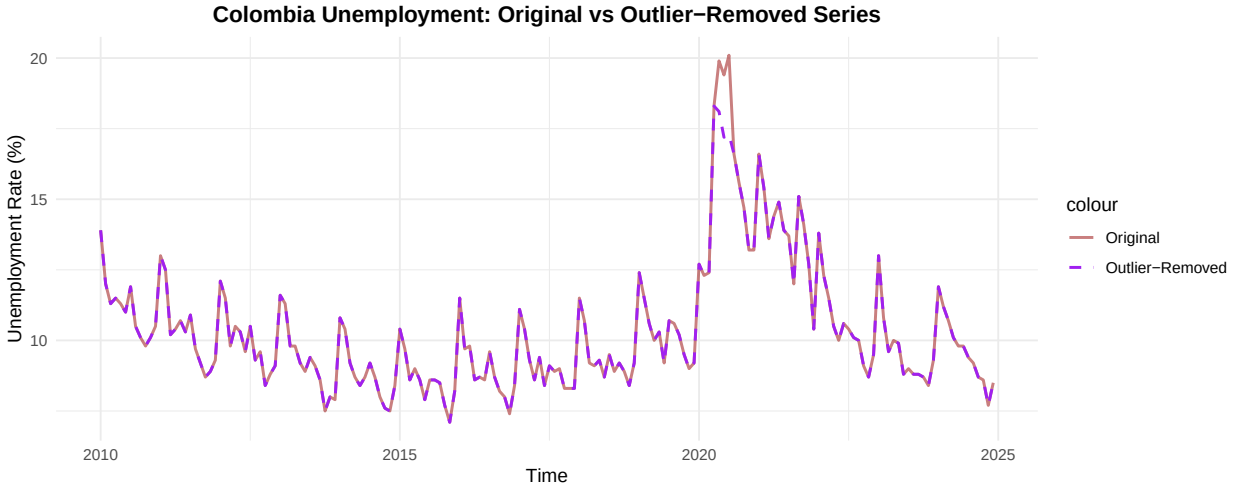
Figure 1: Figure: Global Weighted Average vs US and Colombia Unemployment Trends

Unemployment rates across the globe have seen significant fluctuations over the decades, often reflecting the impact of global economic crises, policy shifts, and technological change. Historically, an unemployment rate of around 4% is often regarded as full employment, meaning most individuals willing and able to work can find jobs. The global plot shows that many economies experienced sharp spikes during major recessions (e.g., 2008) and more recently during the COVID-19 pandemic in 2020.

However, what's striking is the rapid recovery of unemployment rates post-COVID in many regions. As reflected in the global and country-level panels, the US shows a pronounced spike in 2020 followed by a fast recovery, thanks to aggressive fiscal and monetary responses. Colombia, while also experiencing a peak, exhibits greater volatility and slower normalization, likely due to structural vulnerabilities such as labor informality and limited social insurance coverage.

These plots collectively set the stage for understanding the differences in labor market resilience between a high-income country and a developing one — an essential motivation for this forecasting exercise.

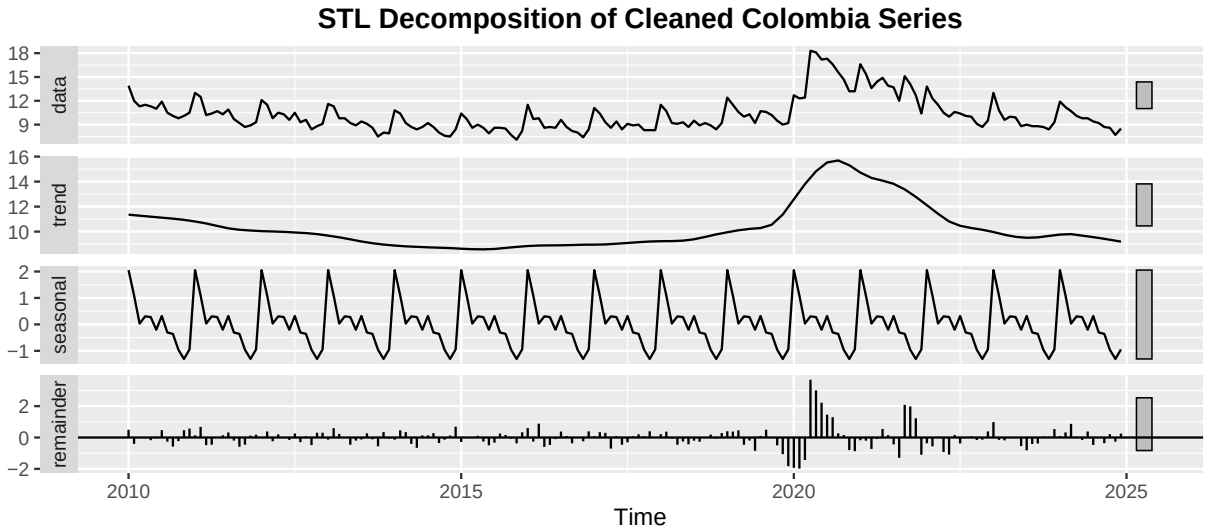
0.3 3. Outlier Detection and Pre-Forecast Diagnostics for Colombia



The original unemployment series for Colombia (2010–2024) contained several large, abrupt shifts not aligned with typical seasonal or trend patterns. These were likely due to structural shocks—such as policy interventions or labor market disruptions—that introduced high-frequency noise into the series.

Forecast models trained on this unprocessed data displayed:

- Poor residual diagnostics (e.g., autocorrelated errors, non-stationarity)
- Inflated prediction intervals due to volatility
- Difficulty capturing the seasonal signal amidst erratic fluctuations



Since the original unemployment series for Colombia contained several irregularities that could affect model performance. Rather than using an IQR-based approach, we employed the tsclean() function which simultaneously handles both outliers and missing values through a more robust procedure. This method:

- Identifies and replaces outliers using smoothing
- Interpolates missing values
- Preserves the overall trend and seasonality

- The cleaned series shows smoother transitions while maintaining the fundamental patterns observed in the original data. Thus, although both versions were tested, the **outlier-removed series was ultimately chosen** as the modeling base to ensure robust and interpretable forecasts.

0.4 4. Analyzing Decomposition and Stationarity

The decomposed unemployment series for both the US and Colombia (after outlier removal) revealed strong seasonal patterns and structural trends. For the US, the ADF test confirmed non-stationarity ($p = 0.5664$), while the Mann-Kendall and Seasonal Mann-Kendall tests indicated a significant downward trend and seasonal effects. Similarly, Colombia's series exhibited non-stationarity (ADF $p = 0.63$), strong seasonality (Kruskal-Wallis $p < 0.001$), but no significant seasonal trend (SMK $p = 0.136$). In both cases, one level of differencing was required to achieve stationarity, and deseasonalized, differenced series were used for robust model training and testing.

The training datasets for both the US and Colombia were constructed by transforming the monthly unemployment percentage into time series objects, excluding the most recent 12 months which were reserved for testing. For the US, data from 2001 to 2023 was used, while Colombia's training period covered 2010 to 2023 to ensure continuity after handling missing and outlier values. Deseasonalization and differencing by 1 lag were applied where necessary to stabilize the series and meet stationarity assumptions before model fitting.

0.5 5 Colombia's Unemployment Model Fitting:

0.5.1 Models 1–4: Baseline Approaches

Make sure to use plain ASCII backticks for chunk delimiters:

```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 514.29, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24

##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 30.927, df = 24, p-value = 0.1559
##
## Model df: 0.   Total lags used: 24

##
##  Ljung-Box test
##
## data:  Residuals from Simple exponential smoothing
## Q* = 30.803, df = 24, p-value = 0.1595
##
## Model df: 0.   Total lags used: 24
```

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0)(2,0,0)[12]
## Q* = 35.059, df = 22, p-value = 0.0382
##
## Model df: 2.   Total lags used: 24
```