Forecasting Labor-Force Participation Rates by Ethnicity

Karl Hetzke and Madalyn Beban

***Introduction***

For our final project, we’ve chosen to find one model to forecast two related macro-economic series: labor force participation rate (LPR) in the United States among the white population and among the black population. These series are published online by the St. Louis Federal Reserve,[[1]](#footnote-1) contain monthly estimates of LPR from January 1972 to March 2021, and are not seasonally adjusted. Our interest in these series is due to the ongoing debate around racial inequality in the United States, and by finding an optimal forecast for these two series we hope to contribute something to that conversation. The main purpose of this paper, however, is in presenting forecasting techniques and we offer no conclusions about the future of racial inequality beyond a forecast of labor force participation rate.

The **Plots and Patterns** section begins with a presentation of graphs, correlograms, and diagnostic tests that examine historic series from January 1972 through March 2021. Visually, plots show strong seasonality whose magnitude changes over time, as well as polynomial trends that influence long-term behavior. The former lead the team to model with monthly dummy variables and with seasonal AR(12) terms. The latter led the team to model with deterministic trends and with first-differences. Augmented Dickey-Fuller tests indicate the two series likely contain unit roots in levels, but not in first differences; this information led the team to compare models of the series in levels with models of the series in first-differences.

We move on to present the results of three models evaluated on the two series of interest. For each model, we split the series into a training series and testing series using the first 80% of observations as training data and the remainder as testing data. For each model, we estimate model parameters with training data, then use those parameters to produce a forecast as long as our testing data set and calculate the root mean squared error (RMSE) of our forecast. In tables for each model, we present a full summary of regression results and fit-statistics, the value of RMSE for forecast series, a plot of true values, fitted values and forecast values, and a correlogram of the residuals from the training data.

**Model 1** is an OLS regression on trend (t1), trend-squared (t2), and monthly dummy variables. We interpreted these models' residual correlograms to indicate that both series contain an AR(1) process after control for trend, and that monthly dummy variables do not sufficiently capture all seasonal autocorrelation. This led us to Model 2.

**Model 2** is a Seasonal AR(1) model with quadratic trend and a single 12-period lag. This sufficiently removes most auto-correlation in residuals, but leaves some seasonal patterns uncaptured. On black LPR, forecast performance is weak as the deterministic trend appears to follow the series poorly. Relatively poor performance of these models for fitting the stochastic trend of black LPR lead us to Model 3.

**Model 3** evaluates the series in first-differences using Seasonal-ARIMA(1,1,1)(1,0,1)12. Rather than differencing the series and then estimating ARMA processes, we used the I(1) term to treat the series as ordered of integration I(1).

We conclude that Model 3 is the most appropriate model for forecasting these two series, despite having the worst performance by RMSE for Black\_LPR. For these series, it is inappropriate to treat trend purely deterministically, and values forecast when treating the series in first-differences as I(1) processes have much stronger forecasting potential.

***Plots and Patterns***

After visual inspection of the two series in levels (see Figure 1), three things are made very clear:

1. the series are not covariance stationary, and likely have a unit root
2. there is strong seasonality in both series, and that seasonality changes over time
3. any long-run trends are complicated, we can try to approximate them deterministically or treat them as random-walk processes

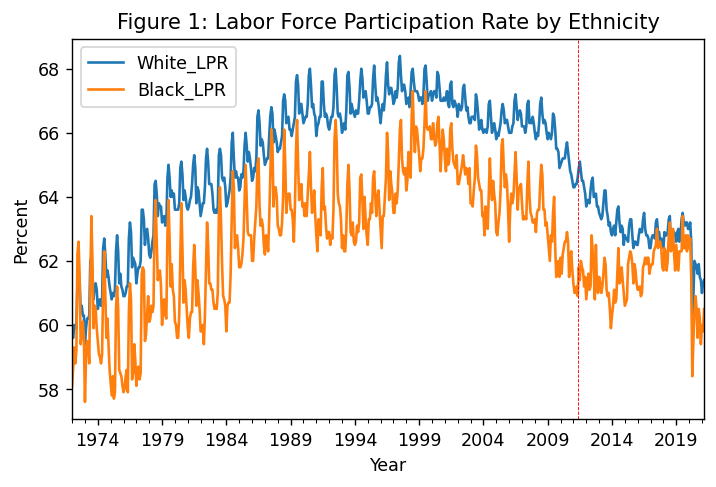
**On the presence of a unit root** the two correlograms of the two series (Figures 2 and 3) are typical of series that contain a unit root with ACFs that damp very slowly. From augmented Dickey Fuller (ADF) tests on both series including a constant and trend term (see Table 1) we fail to reject the null hypothesis of a unit root, so we conclude these two series are not covariance stationary. After taking first differences of both series and conducting another ADF test, we reject the null hypothesis of a unit root for both series; this along with a visual inspection of two series in first differences (see Figure 4) leads us to conclude the series are made stationary in first differences and therefore have an order of integration I(1).

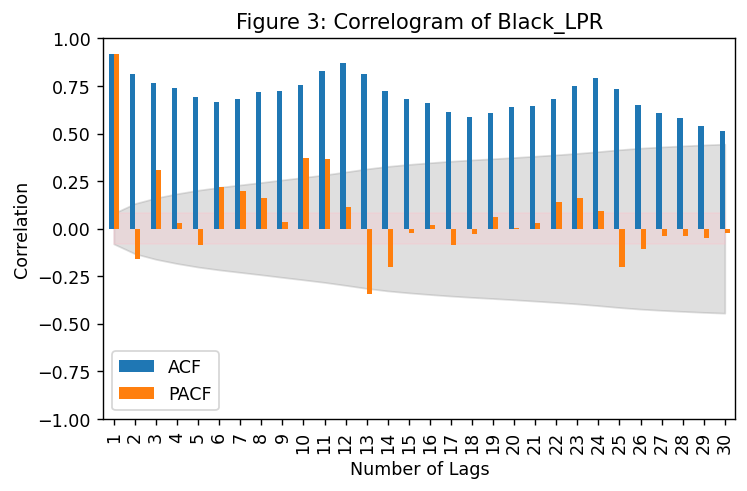
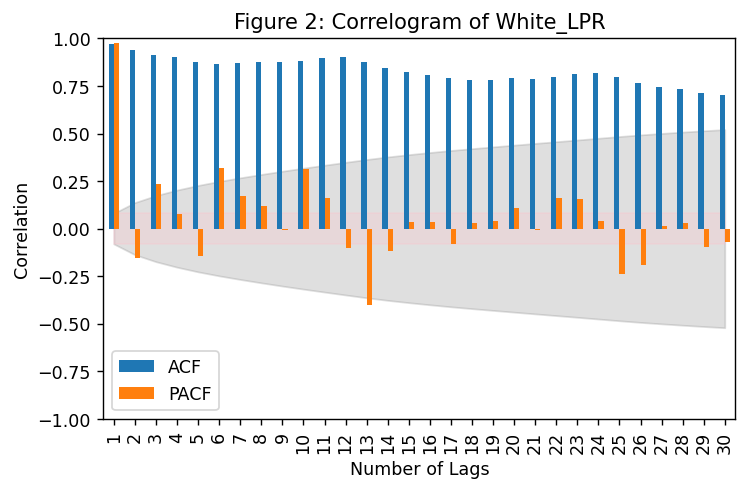
**Regarding seasonality**, the correlograms also show evidence of seasonality with increases ACF at 12 and 24 lags, and increases in PACF at 13 and 25 lags. From Figure 1 we can see that the magnitude of seasonal fluctuation is not constant and appears to be decreasing over time. Figures 1-3 also reveal that the magnitude of seasonal fluctuation is greater in the black population than white population, which may be an indicator that a larger share of the black population engages in work with seasonal variability.

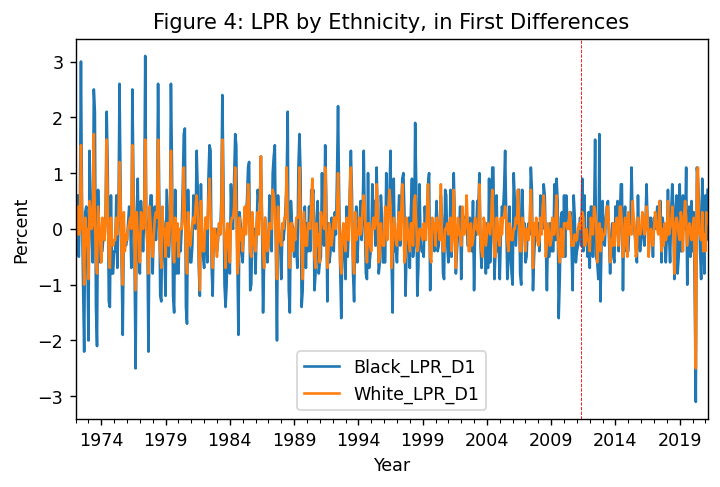
The diminishing magnitude of seasonal fluctuation we observe in Figure 1 suggests that we cannot adequately model seasonality in this series with just monthly dummy variables, but must either interact those seasonal dummy variables with a trend variable, or use a seasonal ARMA method where we include a 12 or 24 period lag-term in our model.

**Regarding trend,** it is likely this trend is best understood as a random-walk process with drift, but if we were forecasting with a relatively short time-horizon it might make sense to treat each series deterministically with a quadratic trend. The vertical red line in Figure 1 represents where our training series ends and testing series begins. The series follows a persistent downward trend in the last months of the training series, but this trend reverses shortly after the start of the training set until 2020 where both series display a sharp drop, obviously related to the COVID-19 pandemic and associated economic shutdowns. Any deterministic trend model is unlikely to forecast the late upswing beginning around 2014, and no model based purely on these series could predict the drop due to COVID-19. In first-differences, trend is not a consideration we’ll need to make.

|  |  |  |
| --- | --- | --- |
| **Table 1: ADF Test Results** | | |
| **Series** | **Test-Stat** | **PValue** |
| White\_LPR | -0.51 | 0.98 |
| Black\_LPR | -0.82 | 0.96 |
| White\_LPR\_D1 | -4.10 | 0.01 |
| Black\_LPR\_D1 | -7.25 | 0.00 |



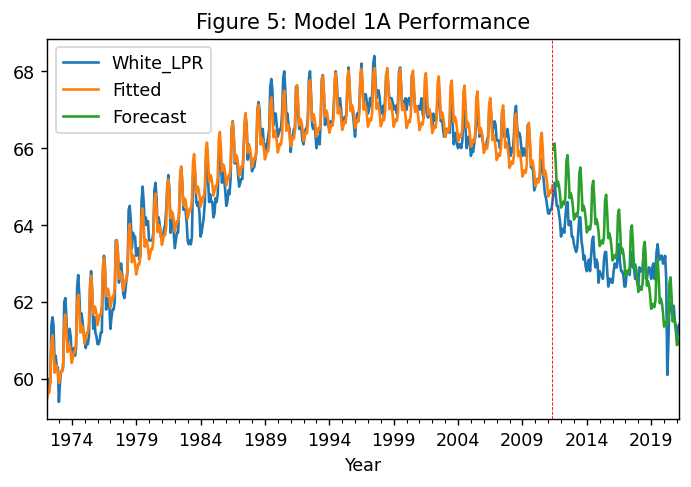


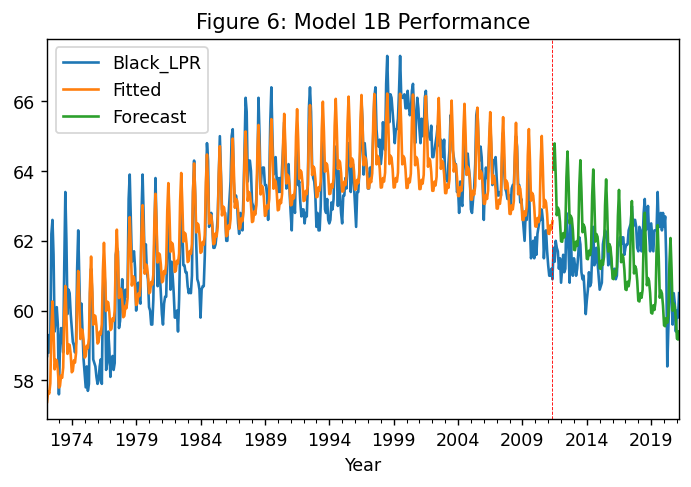


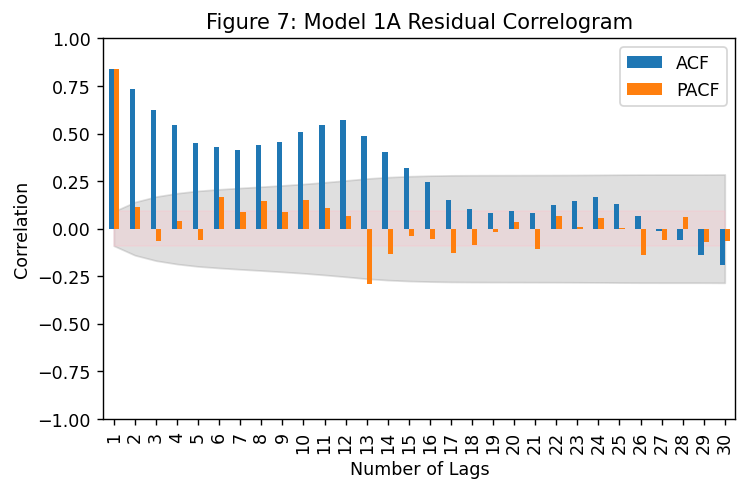
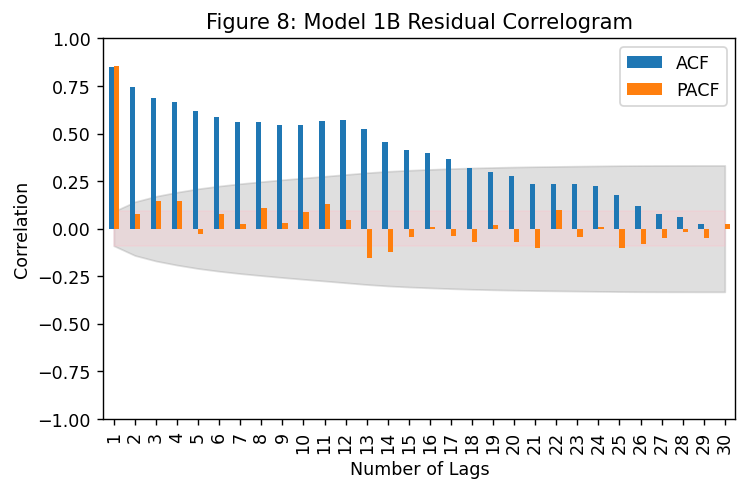
***Model 1: Quadratic Trend with Monthly Dummies***

For our first model, we regressed each series on a trend variable, trend-squared, and monthly dummy variables. A table of RMSE values, plots of forecast performance and correlograms of residuals are presented below, while full tables of estimation results for Models 1A (White\_LPR) and 1B (Black\_LPR) are presented in the appendix.

|  |  |  |
| --- | --- | --- |
| **Table 2: RMSE for Models 1A-1B** | | |
| **Model** | **Series** | **RMSE** |
| 1A | White\_LPR | 0.86 |
| 1B | Black\_LPR | 1.34 |





Model 1 performs surprisingly well given that we know before estimating the series does not actually follow a perfect quadratic trend with constant seasonal fluctuation. All independent variables are statistically significant, though this is unsurprising given the series' visible trend and seasonality. The relatively strong forecast performance is perhaps driven by coincidence: the COVID crisis likely impacted LPRs for both groups, sustaining a downward trend in rates beyond what may have occurred naturally. Note the rise in rates at the end of the testing window.

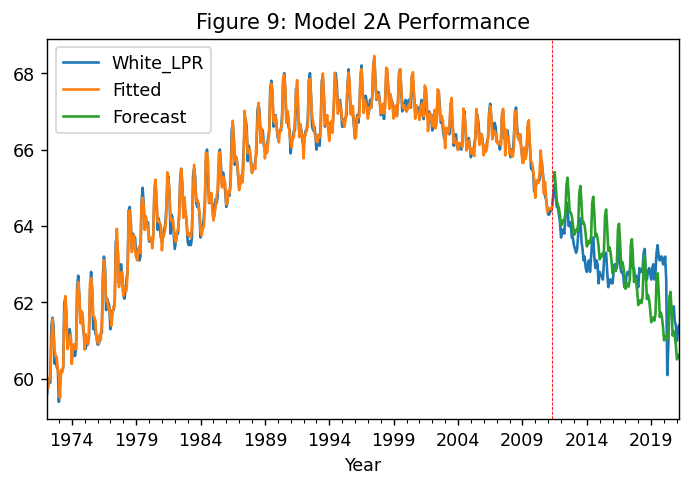
We can observe from residual correlograms of both series that there are seasonal patterns yet unmodeled. This may be caused by time-varying seasonality that constant dummy variables do not capture. The monthly dummy variables did remove much seasonal autocorrelation and from residual correlograms we see more clearly that these two series appear to follow an AR(1) process. For model 2, we attempt to capture this AR(1) process and capture more seasonality by estimating a seasonal AR(1) model with quadratic trend.

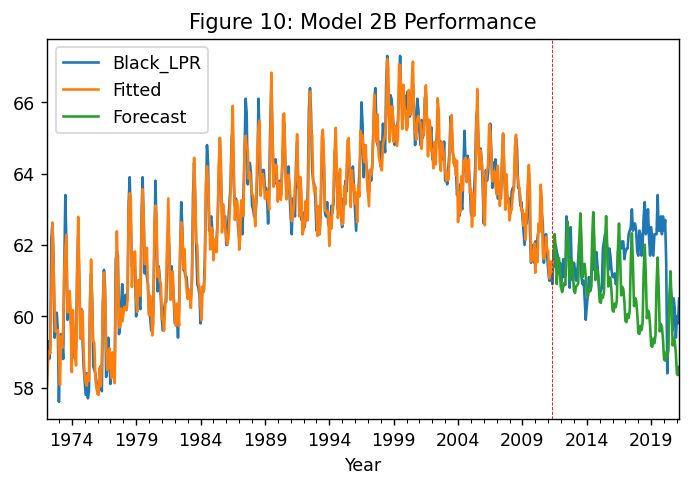
It is interesting to note how much better this simple model performs for White\_LPR than Black\_LPR. This appears to be driven by a higher degree of trend-volatility in Black\_LPR which may suggest a higher level of job insecurity in the black population compared to the white population.

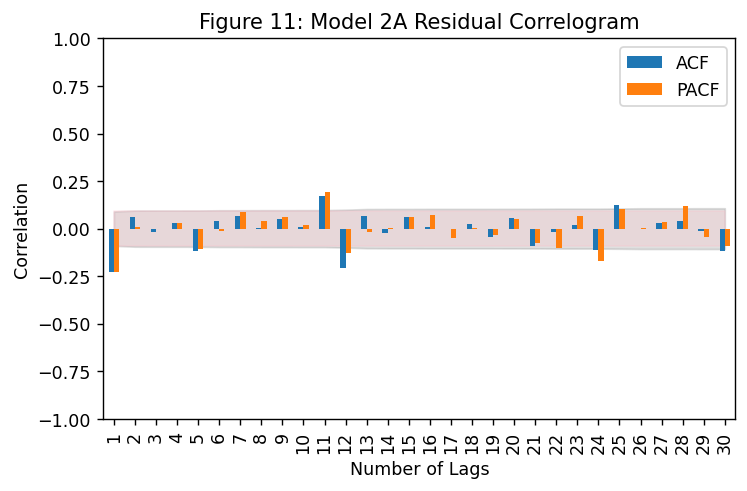
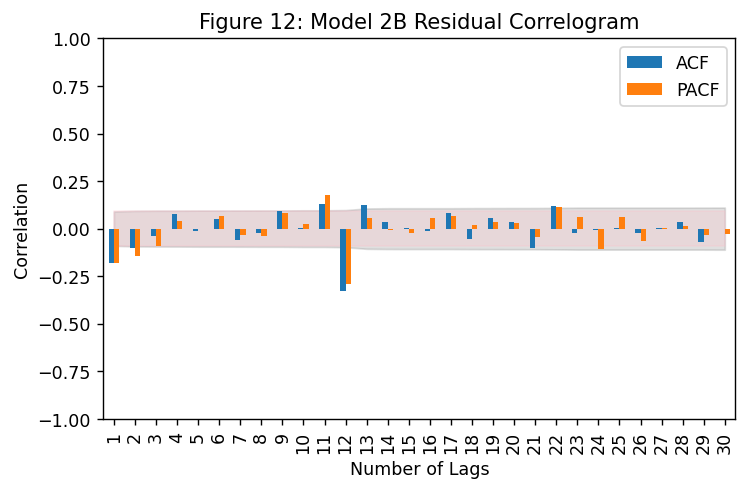
***Model 2: Seasonal ARIMA(1,0,0)(1,0,0)12 with Quadratic Trend and Monthly Dummies***

Having observed from residual correlograms in Model 1 lingering seasonal effects and an AR(1) process, for our second set of models, we estimated an AR(1) process with a seasonal AR(12) lag term and a constant, trend and trend-squared as exogenous variables to account for trend. RMSE of these new models and plots are presented below, full estimation output in the appendix.

|  |  |  |
| --- | --- | --- |
| **Table 3: RMSE for Models 1A-2B** | | |
| **Model** | **Series** | **RMSE** |
| 1A | White\_LPR | 0.86 |
| 1B | Black\_LPR | 1.34 |
| 2A | White\_LPR | 0.74 |
| 2B | Black\_LPR | 1.48 |





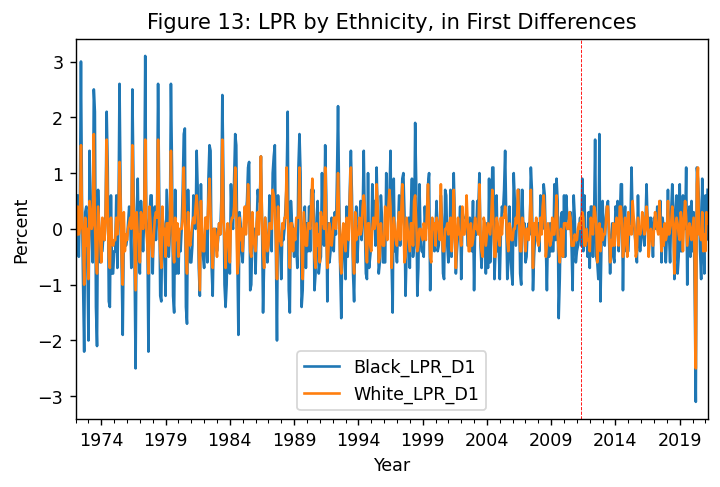
 

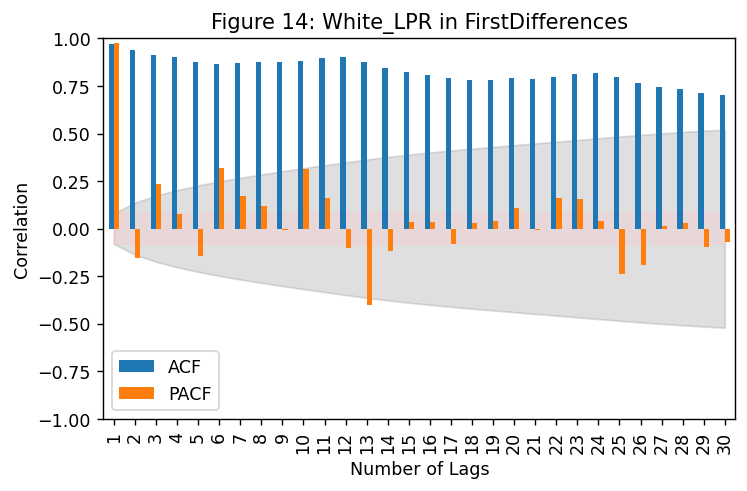
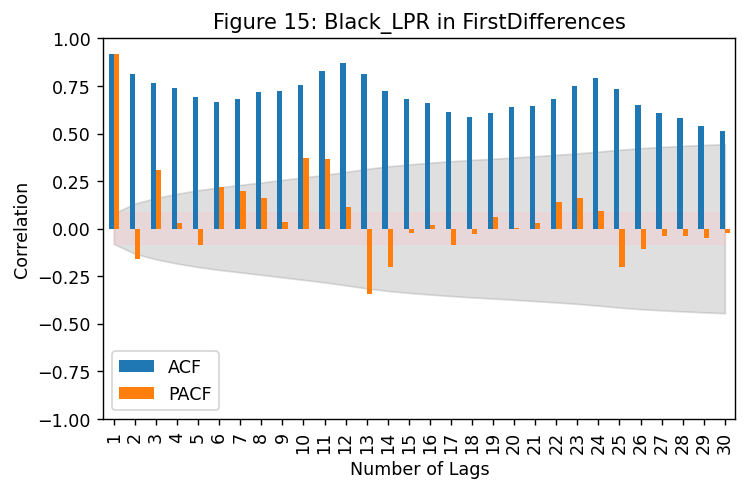
From Model 1, the forecast performance measured by RMSE improves for White\_LPR, but is worse for Black\_LPR. As with model 1, this appears to be due to trend-volatility where the deterministic trend works well enough for White\_LPR (given the down-turn due to COVID, otherwise this forecast would still perform poorly).

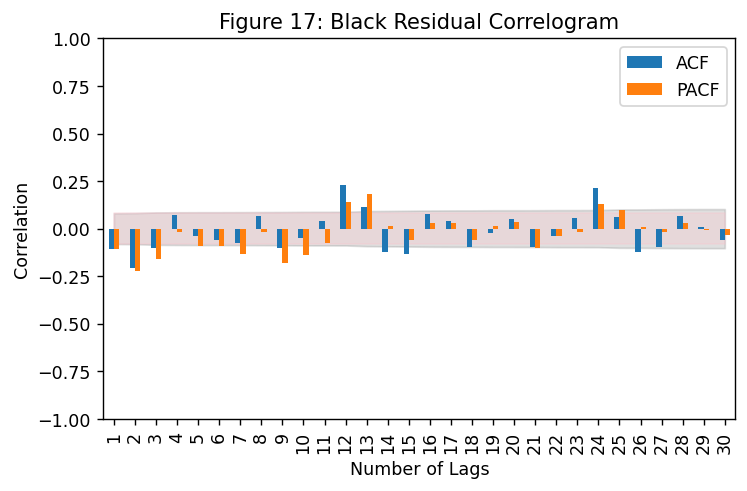
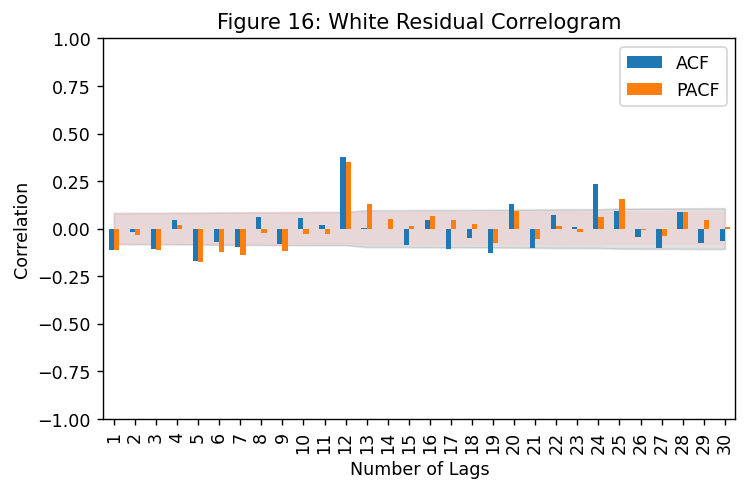
The residual correlograms suggest there aren't much more auto-regressive patterns to exploit other than fine-tuning seasonality, but the plot of our forecast performance clearly shows that much more of our forecast error is driven by our models missing the trend changes that occur around 2014. We are forecasting a consistent downward trend in both series, while in reality both series demonstrate an upward trend from late 2014 to the onset of the COVID-19 pandemic. Fine-tuning seasonality wont address this, and neither will fitting a better deterministic trend to our testing data, so we turn in model 3 to work with these series in first-differences.

***Model 3: Seasonal ARIMA(1,1,1)(1,0,1)12***

ADF tests in the Plots and Patterns section (Table 1) led us to conclude that first differencing removes the presence of a unit root, while visual inspection of the series in differences led us to conclude that first differencing makes both series trend-stationary (see Figure 13, below). Correlogram analysis of the series in first differences shows signs of strong seasonality (see figures 14 and 15), which must be adjusted before we can infer any kind of ARMA process; this led us to regress the series in first-differences on monthly dummy variables and evaluate correlograms of residuals of these regressions (see figures 16 and 17).

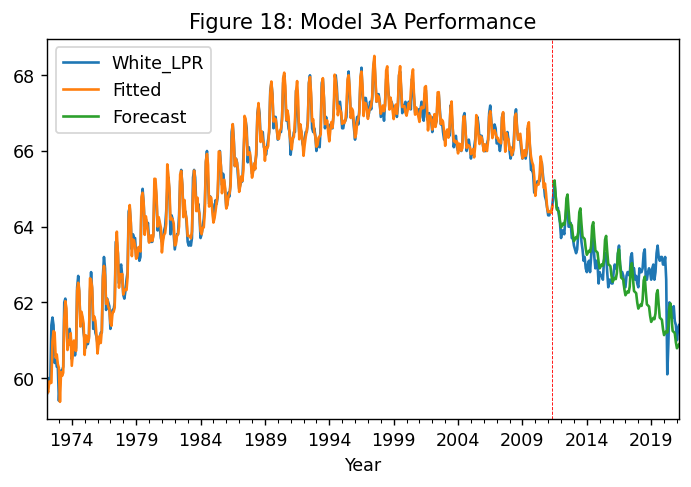


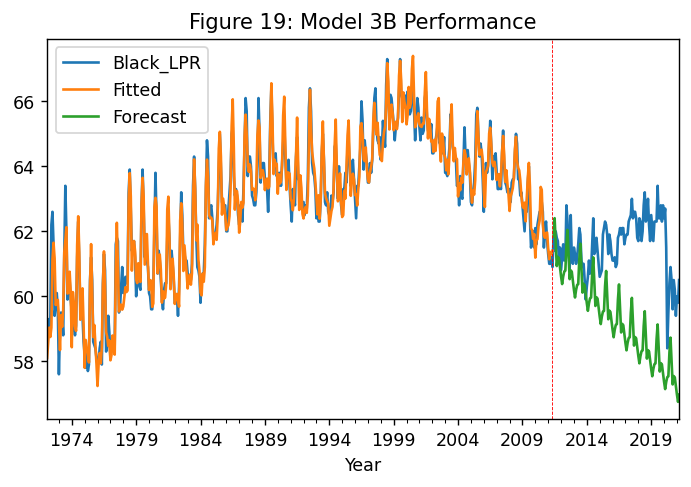
 

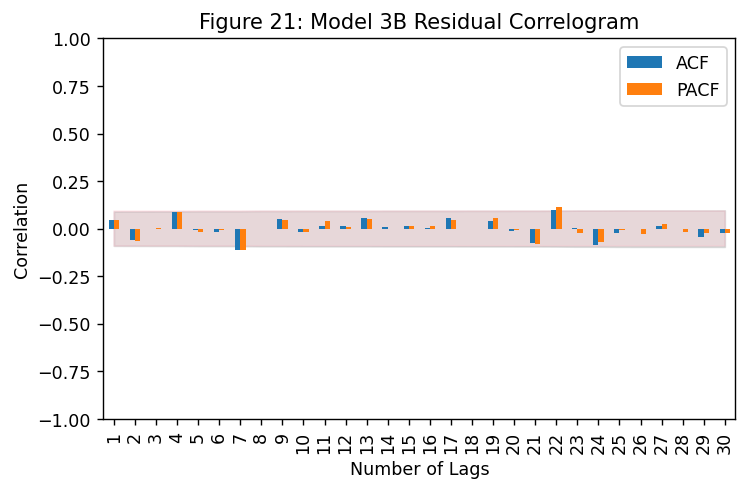
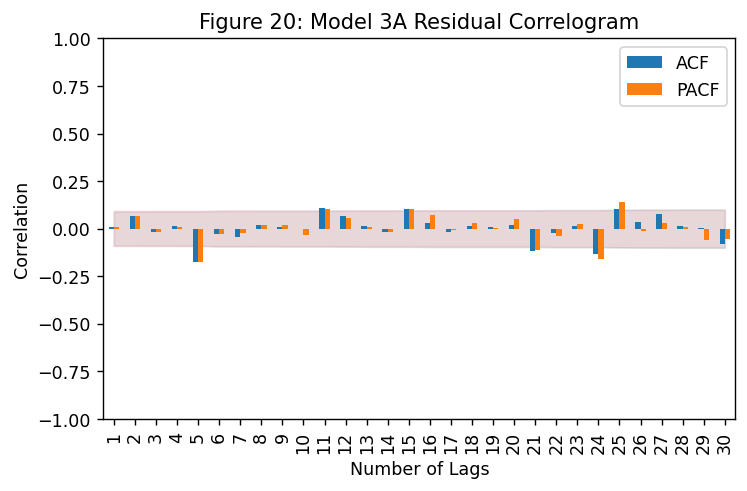


We see in the resultant correlograms significant spikes in ACF and PACF at 12 and 24 lags, indicating remaining seasonality; we also see smaller negative spikes in ACF/PACF that could be indicative of a low-order ARMA process. After evaluating a few low-order ARIMA processes with an I(1) term to difference the series, we chose ARIMA(1,1,1)(1,0,1)12 for our final model.

|  |  |  |
| --- | --- | --- |
| **Table 3: RMSE for Models 1A-3B** | | |
| **Model** | **Series** | **RMSE** |
| 1A | White\_LPR | 0.86 |
| 1B | Black\_LPR | 1.34 |
| 2A | White\_LPR | 0.74 |
| 2B | Black\_LPR | 1.48 |
| 3A | White\_LPR | 0.66 |
| 3B | Black\_LPR | 2.71 |







Correlograms of residuals (figures 20 and 21) suggest there are very few patterns left to exploit, but performance based on RMSE is much worse for Black\_LPR than in previous models. As with previous models 1B and 2B, the forecast error in Black\_LPR seems largely due to the true series over-performing the forecast-trend for most years in the testing data; when the model suggested an ongoing downward trend, the true series was trending upward. Forecast performance for White\_LPR is better than models 1A and 2A, though as with Black\_LPR the error seems largely derived from the series turning upward when the model predicts a continuing downward trend. For both models, were it not for the steep decline as a result of the COVID-19 pandemic, the forecasts would have diverged from the testing data by a larger margin.

***Conclusion***

For a model of LPR in either black or white populations, the strong-seasonality of the series makes a short-term forecast (within a calendar year) fairly straight forward and reliable enough; any seasonal ARMA model should perform reasonably well in these short-window forecasts. For longer term forecasts like the forecasts we tried to implement, the stochastic components of trend make this quite difficult. It would probably be more appropriate to attempt to forecast macroeconomic events that have a predictable impact on LPR than it is to forecast LPR using only auto-regressive techniques and fitting deterministic trends; however if limited to auto-regressive techniques we find our Model 3 to be most appropriate. Model 3 is the best performer for White\_LPR by RMSE, though it is the worst performer for Black\_LPR. The non-deterministic trend, however, cannot appropriately be modeled by assuming whatever trend fits at the end of a training set will continue indefinitely, and the I(1) process seems essential to include.

***Appendix***

*Model 1A: White\_LPR regressed on Trend, Trend-Squared, and Monthly Dummy Variables*

==============================================================================

Dep. Variable: White\_LPR R-squared: 0.972

Model: OLS Adj. R-squared: 0.972

Method: Least Squares F-statistic: 1247.

No. Observations: 473 Prob (F-statistic): 0.00

Df Residuals: 459 Log-Likelihood: -164.16

Df Model: 13 AIC: 356.3

Covariance Type: nonrobust BIC: 414.5

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

t1 0.0466 0.000 99.207 0.000 0.046 0.047

t2 -7.448e-05 9.59e-07 -77.692 0.000 -7.64e-05 -7.26e-05

month\_1 59.2951 0.071 835.820 0.000 59.156 59.435

month\_2 59.3887 0.071 836.585 0.000 59.249 59.528

month\_3 59.4998 0.071 837.607 0.000 59.360 59.639

month\_4 59.4562 0.071 836.459 0.000 59.316 59.596

month\_5 59.6677 0.071 838.909 0.000 59.528 59.807

month\_6 60.6308 0.072 843.336 0.000 60.490 60.772

month\_7 60.8141 0.072 845.331 0.000 60.673 60.955

month\_8 60.3642 0.072 838.538 0.000 60.223 60.506

month\_9 59.7477 0.072 829.450 0.000 59.606 59.889

month\_10 59.9032 0.072 831.094 0.000 59.762 60.045

month\_11 59.8229 0.072 829.475 0.000 59.681 59.965

month\_12 59.6146 0.072 826.094 0.000 59.473 59.756

==============================================================================

Omnibus: 16.871 Durbin-Watson: 0.322

Prob(Omnibus): 0.000 Jarque-Bera (JB): 10.855

Skew: -0.231 Prob(JB): 0.00439

Kurtosis: 2.420 Cond. No. 1.05e+06

==============================================================================

======================================================

RMSE of Forecast is 0.8603456267380386

======================================================

*Model 1B: Black\_LPR regressed on Trend, Trend-Squared, and Monthly Dummy Variables*

==============================================================================

Dep. Variable: Black\_LPR R-squared: 0.815

Model: OLS Adj. R-squared: 0.810

Method: Least Squares F-statistic: 155.7

No. Observations: 473 Prob (F-statistic): 6.64e-159

Df Residuals: 459 Log-Likelihood: -624.99

Df Model: 13 AIC: 1278.

Covariance Type: nonrobust BIC: 1336.

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

t1 0.0388 0.001 31.208 0.000 0.036 0.041

t2 -6.04e-05 2.54e-06 -23.782 0.000 -6.54e-05 -5.54e-05

month\_1 57.3001 0.188 304.878 0.000 56.931 57.669

month\_2 57.3073 0.188 304.714 0.000 56.938 57.677

month\_3 57.5721 0.188 305.923 0.000 57.202 57.942

month\_4 57.4695 0.188 305.184 0.000 57.099 57.840

month\_5 57.7245 0.188 306.346 0.000 57.354 58.095

month\_6 59.2232 0.190 310.939 0.000 58.849 59.597

month\_7 59.9922 0.191 314.770 0.000 59.618 60.367

month\_8 59.1152 0.191 309.969 0.000 58.740 59.490

month\_9 57.9716 0.191 303.781 0.000 57.597 58.347

month\_10 58.2102 0.191 304.842 0.000 57.835 58.585

month\_11 58.1541 0.191 304.363 0.000 57.779 58.530

month\_12 57.8391 0.191 302.534 0.000 57.463 58.215

==============================================================================

Omnibus: 6.535 Durbin-Watson: 0.287

Prob(Omnibus): 0.038 Jarque-Bera (JB): 6.534

Skew: 0.262 Prob(JB): 0.0381

Kurtosis: 2.760 Cond. No. 1.05e+06

==============================================================================

======================================================

RMSE of Forecast is 1.3443214679617723

======================================================

*Model 2A: Seasonal ARIMA(1,0,0)(1,0,0)12 with Quadratic Trend and Monthly Dummy Variables*

==========================================================================================

Dep. Variable: White\_LPR No. Observations: 473

Model: SARIMAX(1, 0, 0)x(1, 0, 0, 12) Log Likelihood 147.727

Sample: 01-01-1972 AIC -261.454

- 05-01-2011 BIC -190.750

Covariance Type: opg HQIC -233.645

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

month\_1 59.2951 0.383 154.787 0.000 58.544 60.046

month\_2 59.3887 0.375 158.245 0.000 58.653 60.124

month\_3 59.4998 0.381 156.334 0.000 58.754 60.246

month\_4 59.4562 0.384 154.931 0.000 58.704 60.208

month\_5 59.6677 0.387 154.282 0.000 58.910 60.426

month\_6 60.6308 0.376 161.170 0.000 59.894 61.368

month\_7 60.8141 0.384 158.377 0.000 60.061 61.567

month\_8 60.3642 0.386 156.281 0.000 59.607 61.121

month\_9 59.7477 0.380 157.103 0.000 59.002 60.493

month\_10 59.9032 0.385 155.527 0.000 59.148 60.658

month\_11 59.8229 0.381 157.118 0.000 59.077 60.569

month\_12 59.6146 0.384 155.134 0.000 58.861 60.368

t1 0.0466 0.004 13.220 0.000 0.040 0.053

t2 -7.552e-05 7.01e-06 -10.779 0.000 -8.93e-05 -6.18e-05

ar.L1 0.8389 0.033 25.719 0.000 0.775 0.903

ar.S.L12 0.6034 0.049 12.234 0.000 0.507 0.700

sigma2 0.0348 0.003 13.133 0.000 0.030 0.040

===================================================================================

Ljung-Box (Q): 143.72 Jarque-Bera (JB): 4.11

Prob(Q): 0.00 Prob(JB): 0.13

Heteroskedasticity (H): 0.75 Skew: -0.22

Prob(H) (two-sided): 0.07 Kurtosis: 3.13

===================================================================================

======================================================

RMSE of Forecast is 0.7377732104085342

======================================================

*Model 2B: Seasonal ARIMA(1,0,0)(1,0,0)12 with Quadratic Trend and Monthly Dummy Variables*

==========================================================================================

Dep. Variable: Black\_LPR No. Observations: 473

Model: SARIMAX(1, 0, 0)x(1, 0, 0, 12) Log Likelihood -323.998

Sample: 01-01-1972 AIC 681.997

- 05-01-2011 BIC 752.702

Covariance Type: opg HQIC 709.807

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

month\_1 57.3001 1.116 51.346 0.000 55.113 59.487

month\_2 57.3073 1.102 52.025 0.000 55.148 59.466

month\_3 57.5721 1.108 51.949 0.000 55.400 59.744

month\_4 57.4695 1.118 51.403 0.000 55.278 59.661

month\_5 57.7245 1.114 51.806 0.000 55.541 59.908

month\_6 59.2232 1.081 54.782 0.000 57.104 61.342

month\_7 59.9922 1.080 55.565 0.000 57.876 62.108

month\_8 59.1152 1.093 54.064 0.000 56.972 61.258

month\_9 57.9716 1.094 53.009 0.000 55.828 60.115

month\_10 58.2102 1.110 52.457 0.000 56.035 60.385

month\_11 58.1541 1.117 52.058 0.000 55.965 60.344

month\_12 57.8391 1.125 51.413 0.000 55.634 60.044

t1 0.0388 0.010 3.844 0.000 0.019 0.059

t2 -6.273e-05 2e-05 -3.136 0.002 -0.000 -2.35e-05

ar.L1 0.8583 0.032 26.532 0.000 0.795 0.922

ar.S.L12 0.6078 0.042 14.315 0.000 0.525 0.691

sigma2 0.2202 0.016 13.527 0.000 0.188 0.252

===================================================================================

Ljung-Box (Q): 129.27 Jarque-Bera (JB): 2.20

Prob(Q): 0.00 Prob(JB): 0.33

Heteroskedasticity (H): 1.00 Skew: 0.08

Prob(H) (two-sided): 1.00 Kurtosis: 2.71

===================================================================================

======================================================

RMSE of Forecast is 1.4830088452382215

======================================================

*Model 3A: Seasonal ARIMA(1,1,1)(1,0,1)12*

==========================================================================================

Dep. Variable: White\_LPR No. Observations: 473

Model: SARIMAX(1, 1, 1)x(1, 0, 1, 12) Log Likelihood 156.236

Sample: 01-01-1972 AIC -302.471

- 05-01-2011 BIC -281.686

Covariance Type: opg HQIC -294.295

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.4131 0.119 3.462 0.001 0.179 0.647

ma.L1 -0.6944 0.090 -7.704 0.000 -0.871 -0.518

ar.S.L12 0.9949 0.001 711.229 0.000 0.992 0.998

ma.S.L12 -0.7395 0.030 -24.336 0.000 -0.799 -0.680

sigma2 0.0282 0.002 15.863 0.000 0.025 0.032

===================================================================================

Ljung-Box (Q): 82.98 Jarque-Bera (JB): 6.21

Prob(Q): 0.00 Prob(JB): 0.04

Heteroskedasticity (H): 0.71 Skew: -0.21

Prob(H) (two-sided): 0.03 Kurtosis: 3.37

===================================================================================

======================================================

RMSE of Forecast is 0.6582054485406676

======================================================

*Model 3B: Seasonal ARIMA(1,1,1)(1,0,1)12*

==========================================================================================

Dep. Variable: Black\_LPR No. Observations: 473

Model: SARIMAX(1, 1, 1)x(1, 0, 1, 12) Log Likelihood -293.889

Sample: 01-01-1972 AIC 597.777

- 05-01-2011 BIC 618.562

Covariance Type: opg HQIC 605.953

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.4584 0.086 5.355 0.000 0.291 0.626

ma.L1 -0.7789 0.060 -12.973 0.000 -0.897 -0.661

ar.S.L12 0.9937 0.002 522.302 0.000 0.990 0.997

ma.S.L12 -0.8021 0.027 -29.575 0.000 -0.855 -0.749

sigma2 0.1921 0.013 14.358 0.000 0.166 0.218

===================================================================================

Ljung-Box (Q): 29.34 Jarque-Bera (JB): 1.49

Prob(Q): 0.89 Prob(JB): 0.47

Heteroskedasticity (H): 0.81 Skew: 0.13

Prob(H) (two-sided): 0.18 Kurtosis: 3.08

===================================================================================

======================================================

RMSE of Forecast is 2.706556168062298

======================================================

1. Data can be queried at fred.stlouis.org, series codes LNU01300003 and LNU01300006 for White and Black LPR, respectively [↑](#footnote-ref-1)