Lecture 9. Case Study

Reference: Chapter 3 of "An introduction to analysis of financial data with R" by R.S.Tsay

WISE&SOE, Xiamen University

2020 Spring

Outline

- Demonstrate applications of methods discussed in previous Chapters;
- To show usefulness and limitations of linear time series models;
- To gain further experience in analyzing time series data with R.

Cases study

- Two real cases:
  - (1) Forecast the weekly gasoline price;
  - (2) Forecast the global temperature.
- A main difficulty for the beginners of time series analysis is finding an adequate model for a given series. This is particularly so when the dynamic dependence of the data is complex or when many models seem to fit the data well.
- *All models are wrong, but some are useful*—George Box .
- Our goal is to find an appropriate model that is useful to the objective of data analysis. It would not be surprising that you can find alternative models from these data sets.

General guidelines for time series modeling

- First, data are only part of information available in an application. We may have some prior knowledge about the problem at hand. In this situation, it is important to make use of substantive information in model selection. Combination and cross-validation between prior knowledge and data can improve model selection.
- Second, in some cases, many models are available and the distinction between these competing models is small. The issue of model selection then becomes less important and one can comfortably use one of the models.
- Third, we may combine several competing models for pooling or combining forecasts;
- Fourth, a general principle is to start with a simple model. Another is "*Keeping it sophistically simple (KISS)*" (Arnold Zellner)

WEEKLY REGULAR GASOLINE PRICE

- The data are obtained from US Energy Information *Administration at http://www.eia.gov.*
- (i) We analyze the weekly retail regular gasoline price of US; (ii) We study the dependence of gasoline price on the crude oil price and use the latter to improve the forecast of the former.
- The crude oil price are available three days prior to the gasoline prices from Friday to Mondays.
- Because the price vary substantially, we use log prices in our analysis.

Weekly log prices of regular gasoline (dollars per gallon) and crude oil of the United States

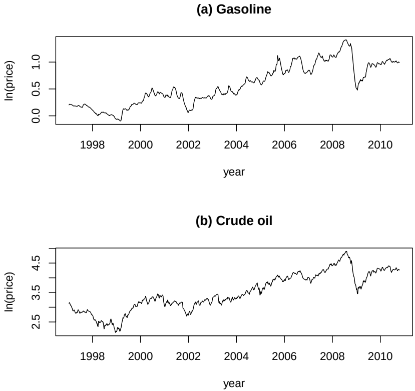


Figure 1: Weekly log prices of (a) regular gasoline (dollars per gallon) (1997/01/06-2010/09/27) and (b) crude oil (dollars per barrel) (1997/01/03-2010/09/24)

Pure Time Series Model

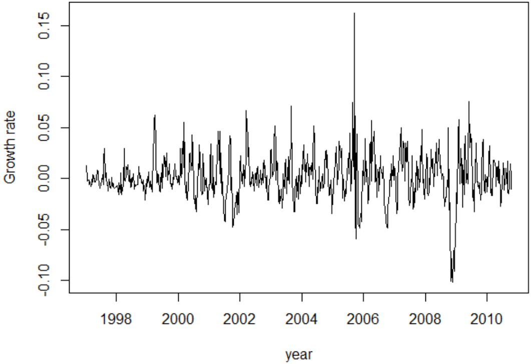


Figure 2: Weekly growth rates (log returns) of the US regular gasoline price from January 1997 to September 2010.

Pure Time Series Model

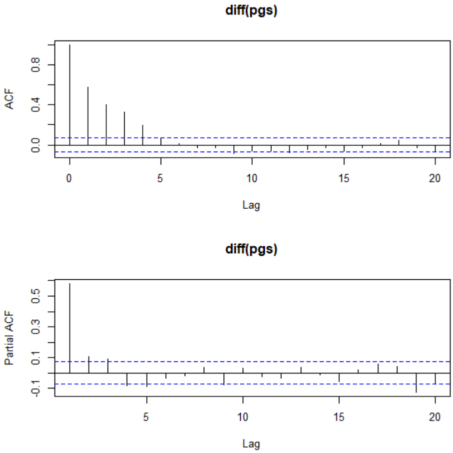


Figure 3: Sample autocorrelations and partial autocorrelations of the growth rates of weekly US regular gasoline prices from January 6, 1997 to September 27, 2010.

Pure Time Series Model

- Both ACFs and PACFs decay quickly, confirming that the series is weakly stationary.
- The plot shows that the first five lags of PACF are significantly different from zero, suggesting that an AR(5) model might be appropriate for X_t.
- Is the mean equal to 0 ?
  - One-sample t-test gives a t-ratio of 1.306 with p-value is 0.192.
- Remove the insignificant coefficient.
- The fitted model is:

$$(1 - 0.504L - 0.079L^2 - 0.122L^3 + 0.101L^5)X_t = \varepsilon_t \qquad (1)$$

$$\sigma_\varepsilon^2 = 3.265 \times 10^{-4}, \qquad \text{AIC} = -3704.96$$

The standard errors of the coefficients are $0.037, 0.042, 0.039$, and 0.033, respectively.

Pure Time Series Model

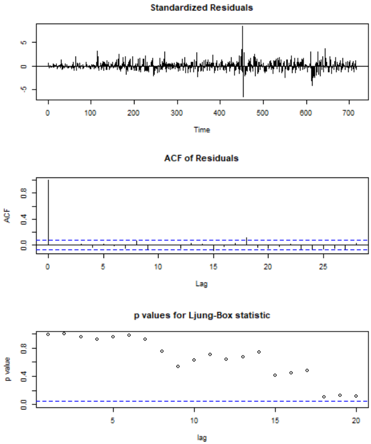


Figure 4: Model checking of AR(5) model in Equation (1) for the growth rate of weekly US regular gasoline prices from January 6, 1997 to September 27, 2010. (a) Standardized residuals, (b) ACF of residuals, (c) p-values for Ljung-Box statistic.

Pure Time Series Model

- The sample PACF of X_t shows a dominating correlation at lag 1. The ACF of X_t indicates that the autocorrelations decay exponentially. These two features suggest that $p = 1$.
- The significance of ACFs and PACFs at higher order lags indicates $p = 1$ is not sufficient. Thus, another possibility is to entertain an ARMA model.
- To this end, we consider an ARMA(1,3) model. After removing an insignificant coefficient, we obtain the model

$$(1 - 0.633L_{(0.051)})X_t = (1 - 0.127_{(0.061)}L + 0.141_{(0.041)}L^3)\varepsilon_t \quad (2)$$

$$\sigma_\varepsilon^2 = 3.276 \times 10^{-4}, \qquad \text{AIC} = -3704.6.$$

- The AIC is larger than that of an AR(5) model in Equation (1). Consequently, we select the AR(5) model in Equation (1) as the pure time series model for x_t.

Use of Crude Oil Prices

- As gasoline prices depend heavily on spot prices of crude oil, we employ a regression model with time series errors to improve the accuracy in forecasting weekly gasoline price. Let Z_t be the weekly growth rates of the US crude oil price.
- A simple linear regression model gives

$$X_t = 0.287_{(0.015)} Z_t + \varepsilon_t, \quad \text{Adjusted } R^2 : 0.3357. \tag{3}$$

Use of Crude Oil Prices

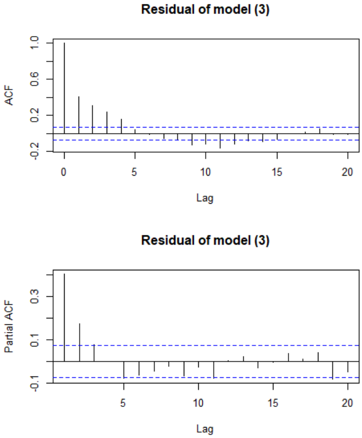


Figure 5: Sample ACF of residual series of linear regression for two series of oil price

Use of Crude Oil Prices

- The sample ACF and PACF of the residual ε_t of Equation (3) are similar to those of X_t. Thus, the use of Z_t does not alter the model specification of X_t.
- As a matter of fact, the ar command in R specifies an AR(6) model for the residual ϵ_t.
- Removing the insignificant coefficients, the fitted model becomes

$$(1 - 0.404L - 0.164L^2 - 0.096L^3 + 0.101L^5)(X_t - 0.191Z_t) = \eta_t, \qquad (4)$$

$$\sigma_\eta^2 = 2.53 \times 10^{-4}, \qquad \text{AIC} = -3884.95.$$

The standard errors of the coefficient estimates are $0.039, 0.040, 0.039, 0.035$, and 0.014, respectively.

Use of Crude Oil Prices

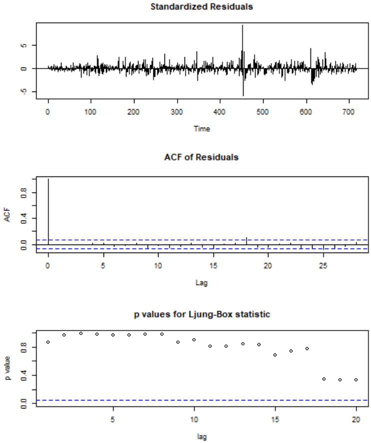


Figure 6: Model checking for regression model with AR(5) innovations in Equation (4) (a) Standardized residuals, (b) ACF of residuals, (c) p-values for Ljung-Box statistic.

Use of Crude Oil Prices

- Given the high dependence between the two oil prices, one would expect the regression model with time series errors in Equation (4) to be a better model. This is indeed the case.
- The information on crude oil prices reduces the residual variance from 3.265×10^4 to 2.532×10^4, a 22.5% reduction.
- Similarly, the AIC drops from -3704.6 to -3884.95.
- Up to now, the regression model with additional information of the growth rates of crude oil prices is a better model.
- As it is seen later, the regression model also produces more accurate out-of-sample forecasts.

Use of Lagged Crude Oil Prices

- The usefulness of the improved model in Equation (4) is limited to 3 days, because it uses the growth rate of crude oil price 3 days earlier. To increase the lead time in forecasting, one can use the lagged growth rate of crude oil price.
- For instance, the model

$$X_t = \beta Z_{t-1} + \varepsilon_t$$

would give the analysts 10 days in advance to predict the weekly gasoline price. The fitted model is

$$X_t = 0.186_{(0.0172)} Z_{t-1} + \varepsilon_t, \qquad \text{Adjusted } R^2 = 0.1410 \tag{5}$$

- The adjusted R^2 as expected, is much lower than that of the regression model in Equation (3). This is understandable because the correlation between X_t and Z_{t-1} is smaller than that between X_t and Z_t.

Use of Lagged Crude Oil Prices

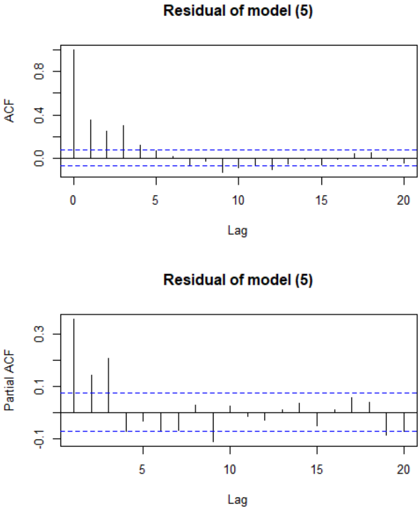


Figure 7: Sample ACF of residual series of linear regression model (5)

Use of Lagged Crude Oil Prices

- The sample ACFs and PACFs of the residuals are similar to those of the residuals of Equation (3), albeit a little bit more complicated.
- The **ar** command of **R** identifies an AR(9) model for the residuals. After removing all insignificant estimates, we obtain the model

$$(1 - 0.454L - 0.088L^2 - 0.142L^3 + 0.083L^5 + 0.064L^9)(X_t - 0.04Z_{t-1}) = \eta_t, \qquad (6)$$

$$\sigma_\eta^2 = 3.23 \times 10^{-4}, \quad \text{AIC} = -3703.4.$$

  The standard errors of the estimates are, in order, 0.043, 0.041, 0.039, 0.035, 0.032, and 0.018, respectively.
- The AIC is close to that of the pure time series model in Equation (2), but much higher than that of the model in Equation (4).
- Diagnostic checking statistics of model (6) are similar to those of model (4) and, hence, are omitted. The model is also adequate.

Out-of-sample prediction

- We divides the data into modeling and forecasting subsamples and uses an iterative procedure to compute prediction. Specifically, the iterative procedure consists of an estimation-prediction cycle, and starts with the last data point of the modeling subsample as the first forecast origin. Once a forecast is produced, the procedure advances the forecast origin by 1 and repeats the estimation-forecasting cycle.
- The recursive 1-step ahead forecast errors in the forecasting subsample are then used to measure the accuracy of prediction.
- Two most widely used measures of forecasting accuracy are the root mean square of forecast errors (RMSFE) and the mean absolute forecast errors (MAFE).

Out-of-sample prediction

- For the growth rates of weekly gasoline price, we divide the data into modeling and forecasting subsamples with the latter consisting of the last 400 data points. That is we start the forecast origin on January 24, 2003. The 400 observations in the forecasting subsample should provide reliable measures of RMSFE and MAFE.
- The results are given below:

Table 1:

Model	RMSFE	MAFE
AR(5) model in Equation (2)	0.02171	0.01538
Regression model in Equation (4)	0.01926	0.01285
Regression model in Equation (6)	0.02166	0.01548

Out-of-sample prediction

From the table, we make the following observations.

- The regression model with crude oil price 3-day earlier performs best. This is consistent with the in-sample comparison and understandable. It shows that the gasoline price reflects the crude oil price quickly.
- The other two models fare similarly in out-of-sample prediction. The contribution of crude oil price 10-day earlier is small, if any.
- If one wants to predict the gasoline price more than 10 days in advance, one can simply use the pure time series model in Equation (2). On the other hand, one should use the regression model in Equation (4) if the forecast horizon is less than 3 days.

Case 2: GLOBAL TEMPERATURE ANOMALIES

- Global warming is a topic of considerable importance and has attracted much more attentions in recent years, ranging from environmental engineers to scientists to economists. If the rise in global temperature continues, it will have a major impact on the global economy.
- In this section, we will analyze the monthly global temperature anomalies from January 1880 to August 2010.
- Our goals:
  - To illustrate the methods discussed in the previous chapters about time series modeling and forecasting;
  - To compare different models;
  - To see the limitation of time series models in long-term prediction;
  - To show the difficulty in distinguishing trend-stationarity from unit-root stationarity based purely on the data

Case 2: GLOBAL TEMPERATURE ANOMALIES

There are several data sets available for global temperature anomalies:

- Goddard Institute for Space Studies (GISS), National Aeronautics and Space Administration(NASA): https://data.www.giss.nasa.gov/gistemp/
- National Climatic Data Center (NCDC), National Oceanic and Atmospheric Adminstration(NOAA): https://www.giss.nasa.gov/
- We employ the series of monthly means based on land-surface air temperature anomalies of GISS, NASA.
- However, we obtained similar results from the data of NOAA. The same models apply to both series.

Case 2: GLOBAL TEMPERATURE ANOMALIES

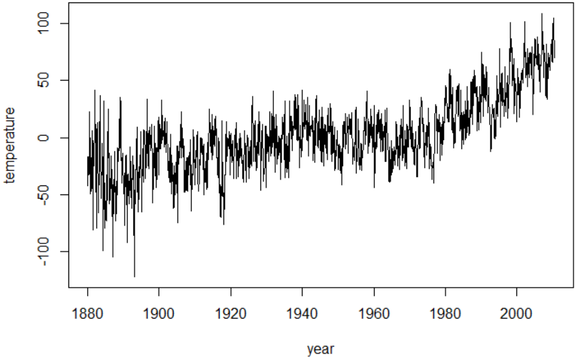


Figure 8: Monthly global temperature anomalies from January 1880 to August 2010, sample size=1568.

GLOBAL TEMPERATURE ANOMALIES

- An upward trend is clearly seen from the plot. In particular, the slope of the trend seems to increase in the early 1980s. On the other hand, the variability of the temperature is relatively stable over the 131 years.
- Let G_t denote the monthly global temperature anomalies. To specify a model for G_t, we start by examining the dynamic dependence of the series.

Unit-Root Stationarity

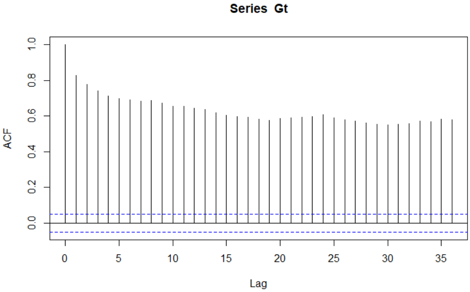


Figure 9: Sample autocorrelation function of the monthly global temperature anomalies.

- As expected, the ACFs are high and decay slowly. A careful inspection also shows that the ACFs exhibit a cyclic pattern with peaks occurring around lags 24 and 36. This latter feature is not surprising because temperature often has a seasonal pattern.

Unit-Root Stationarity

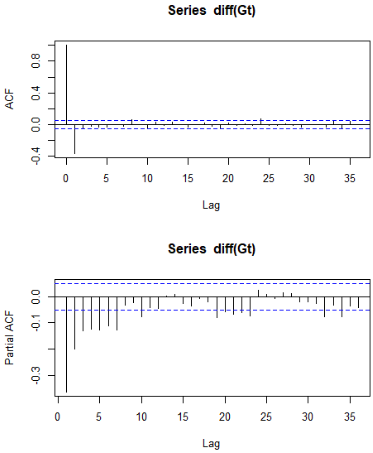


Figure 10: Sample autocorrelation and partial autocorrelation functions of the differenced G_t.

Unit-Root Stationarity

- We start with the simple ARIMA(1,1,2) model. Here, we use $p = 1$ because the differenced series X_t has a large lag-1 PACF and $q = 2$ because the first two ACFs of X_t are significant. As the specified MA(2) model can have significant PACFs at lower order lags, we decide to keep $p = 1$. The high order ACFs are temporarily ignored because we like to keep the model simple. The fitted model is

$$(1 - 0.739_{(0.0406)} L)(1 - L) G_t = (1 - 0.297_{(0.0533)} L + 0.318_{(0.0492)} L^2)\varepsilon_t, \tag{7}$$

$$\sigma_\varepsilon^2 = 272.1, \quad \text{AIC} = 13,241.1.$$

Unit-Root Stationarity

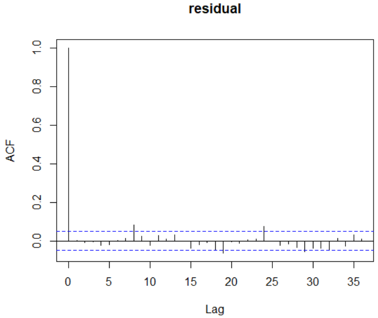


Figure 11: Sample autocorrelation function of the residuals of ARIMA(1,1,2) model for global temperature anomalies.

- Based on the residual ACFs, the model is inadequate because the ACFs are significant at lags 8 and 24.

Unit-Root Stationarity

- The significance of ACF at lag 24 is understandable because of the seasonal nature of temperature. On the other hand, it is not easy to explain the serial correlation at lag 8.
- Consequently, we refine the model:

$$(1-\phi L)(1-L)G_t = (1-\theta_1 L - \theta_2 L^2)(1-\theta_{24}L^{24})\varepsilon_t$$

- The fitted model is

$$(1-0.761_{(0.038)}L)(1-L)G_t = (1-1.324_{(0.052)}L+0.342_{(0.049)}L^2)(1-0.072_{(0.024)}L^{24})\varepsilon_t, \tag{8}$$

$$\sigma_\varepsilon^2 = 270.6, \quad \text{AIC} = 13,234.4.$$

Unit-Root Stationarity

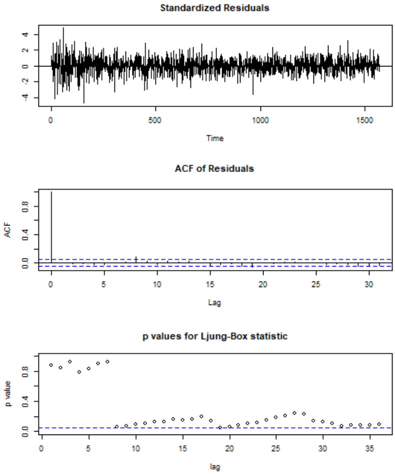


Figure 12: Diagnostic checking plots of the model in Equation (8). (a) Standardized residuals, (b) ACF of residuals, (c) p-values for Ljung-Box statistic

Unit-Root Stationarity

- The residual plot looks reasonable and the p values of Ljung-Box statistics are above 0.05 except for Q(8) and Q(19). As expected, the residual ACFs show marginally significant values at lags 8 and 19.
- As mentioned earlier, it is hard to explain the lag-8 serial correlation and the magnitude of the ACF is small, we terminate the modeling process and treat the model in Equation (8) as an adequate model. The AIC of model (8) is 13,234.4, which is smaller than 13,241.1 of model (7).

Trend-Stationarity

In the literature, some analysts and scientists use time trend to model the global temperature anomalies. By time trend, we mean using time index as an explanatory variable. Consider the model

$$G_t = \beta_0 + \beta_1 t + Z_t$$

where Z_t is an innovation series denoting the deviation of the global temperature anomalies from a time trend. For the global temperature anomalies, the fitted linear regression model is

$$G_t = -38.04_{(1.135)} + 0.05156_{(0.0013)} t + Z_t, \tag{9}$$

Trend-Stationarity

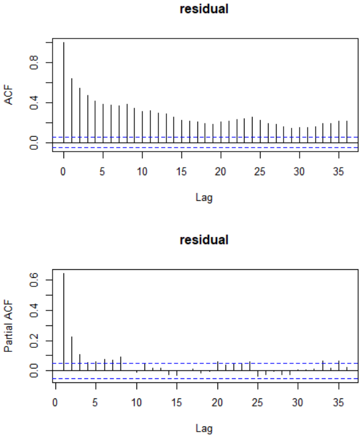


Figure 13: Sample autocorrelations and partial autocorrelations of the innovation series Z_t in Equation (9) for monthly global temperature anomalies.

Trend-Stationarity

- The PACFs decay quickly and the ACFs do not show any high value. Therefore, it is reasonable to assume that zt does not have a unit root. That is, Z_t is stationary and, hence, G_t is trend-stationary.
- Next, we specify a model for the innovation series Z_t. Because its ACFs of Z_t do not cut-off at any finite lag, Z_t does not follow a simple MA model. In other words, some AR component is needed. The PACFs of Z_t have two discernible features. First, the first eight lags of PACFs are significant, indicating that Z_t does not follow a low order AR model. This implies that $q > 0$. Second, the PACFs do not follow a simple exponentially decaying pattern. This means that $p > 1$.

Trend-Stationarity

- Putting information together and keeping the order simple, we start with an ARMA(2,1) model for Z_t, Then, the model for G_t becomes

$$(1-\phi_1 L-\phi_2 L^2)(G_t-\beta_0-\beta_1 t)=(1-\theta_1 L)a_t$$

- The fitted model is

$$(1-1.239_{(0.0567)}L+0.272_{(0.0477)}L^2)(G_t+38.72_{(5.35)}-0.053_{(0.0059)}t)=(1-0.78_{(0.0460)}L)\epsilon_t, \tag{10}$$

$\sigma_\epsilon^2=272.9$

Trend-Stationarity

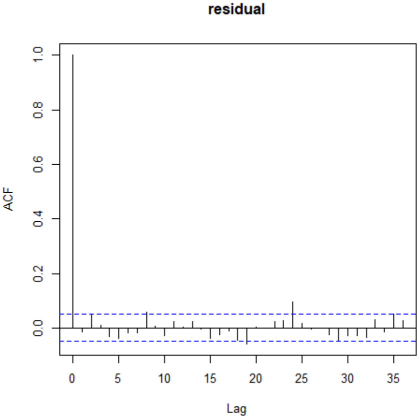


Figure 14: ACF of residual

Trend-Stationarity

- However, the residual ACFs of the model (10) show a significant value at lag 24. Consequently, we further refine the model and obtain

$$(1-1.196L+0.239L^2)(G_t+38.72-0.0529t)=(1-0.745L)(1-0.0856L^{24})\varepsilon_t \quad (11)$$

$$\sigma_\epsilon^2 = 270.8, \quad \text{AIC} = 13,247.5.$$

The standard errors of the coefficient estimates are, in order, 0.059, 0.048,5.18, 0.006, 0.049, and 0.024, respectively.

Trend-Stationarity

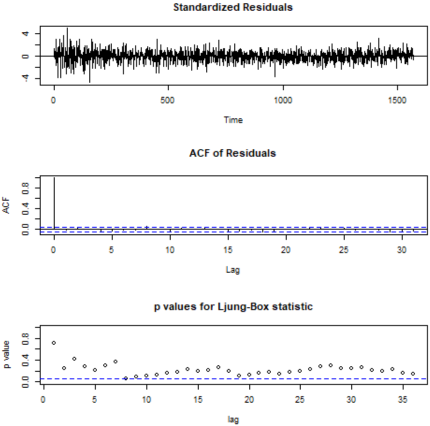


Figure 15: Diagnostic checking of the fitted model in Equation (11). (a) Standardized residuals, (b) ACF of residuals, (c) p-values for Ljung-Box statistic.

Trend-Stationarity

- On the basis of model in Eq.(11), the global temperature increases on an average $0.0529/100^{o}$ C per month. That is, the global temperature increases $0.00635^{o}C$ per year.
- This is very significant because it implies that global temperature on an average will increase $1^{o}C$ every 157 year.

Out-of-Sample Comparison

- We divide the sample into modeling and forecasting subsamples with the latter consisting of the last 200 observations. We then apply the same method as case 1 to compute the 1-step ahead prediction of the two competing models in Equations (8) and (11). For the global temperature data with 200 1-step ahead out-of-sample predictions, we obtain the following results:

Table 2:

Model	RMSFE	MAFE
Difference-stationary model in Equation (8)	14.526	11.167
Trend-stationary model in Equation (11)	15.341	11.966

- The difference-stationary model is preferred based on the 1-step ahead prediction. The drop in RMSFE is about $(15.341 - 14.526)/14.526 = 5.6\%$.

Long-Term Prediction

- Global warming is concerned with long-term prediction. We consider and compare the performance of the two competing models in Equations (8) and (11) using long-term prediction.
- Specifically, using August 2010, which gives the last data point, as the forecast origin, we compute 1-step to 1200-step ahead predictions of the monthly global temperature anomalies. In other words, we use the models built based on data of the past 131 years to predict the global temperatures for the next 100 years.

Long-Term Prediction Using Difference-Stationary model in Eq. (8)

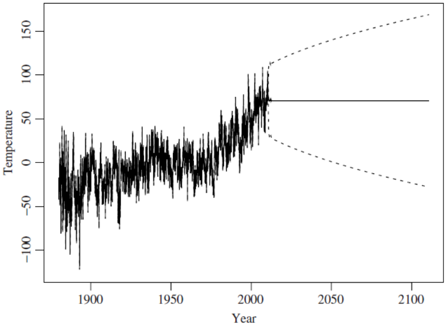


Figure 16: Long-term point and interval forecasts of the monthly global temperature anomalies based on the difference-stationary model in Equation (8). The forecast origin is August 2010 and the forecast horizon is 100 years.

Long-Term Prediction Using Difference-Stationary model in Eq. (8)

- First, similar to other unit-root models, the long-term forecasts converge to a constant represented by a horizontal line in the plot. The level of this horizontal line depends on the forecast origin. Second, the length of the 95% interval forecasts continues to grow with the forecast horizon. In fact, the length of the interval diverges to infinity eventually.
- These two features have important implications in forecasting. First, they indicate that the long-term forecasts are rather uncertain. This makes intuitive sense because long-term predictions of the model are dominated by its random walk component and for a random walk the current value contains little information about the future. Second, they demonstrate clearly that the model is only informative in short-term prediction.

Long-Term Prediction Using Trend-Stationary Model in Eq. (11)

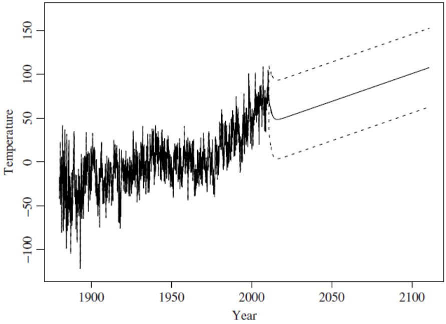
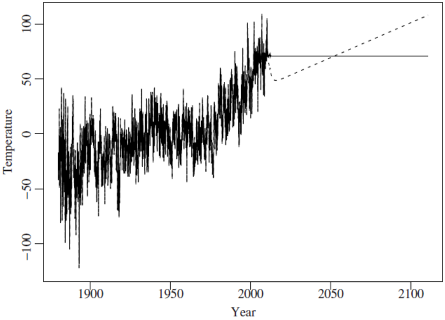


Figure 17: Long-term point and interval forecasts of the monthly global temperature anomalies based on the trend-stationary model in Equation (11). The forecast origin is August 2010 and the forecast horizon is 100 years.

Long-Term Prediction Using Trend-Stationary Model in Eq. (11)

- First, because of the positive time slope, the predictions grow with the forecast horizon.
- Second, the lengths of the interval forecasts are stable over time. In fact, the lengths quickly converge to a constant with the constant being approximately $4\sigma_z$, where σ_z is the sample standard error of the innovation series z_t .
- The innovation series z_t in Eq. (9) is stationary. As such the variances of the forecasts of z_t converge to its variance, σ_z^2 , when the forecast horizon increases.
- For the trend-stationary model, in Equation (11), the prediction of the time trend is certain conditioned on the coefficients β_0 and β_1 being fixed. The uncertainty in forecasts is determined by that of z_t . Consequently, the variances of forecast errors of the model in Equation (11) converge to that of z_t.

Point forecasts of the monthly global temperature anomalies based on two competing models in Equations (8) and (11). The forecast origin is August 2010 and the forecast horizon is 100 years.


Temperature
Year