Time Series Boot Camp

Introduction to Time Series Analysis in Communication Research

Jordan M. Foley

March 10, 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Overview

WELCOME!

- This is the Time Series Boot Camp hosted by the Computational Methods research group.
- We will be covering the following topics:
  - Day 1: Conceptual Overview, Univariate Models
  - Day 2: Multivariate Models (VAR, Granger causality)
  - Day 3: Software, Packages, and Coding
- Although lecturing will be necessary, I hope this evolves into more of a conversation.

What Are We Going to Cover?

1. Conceptual Overview
   - What is time series anlysis?
   - How could it be useful for me?
   - How should I think about time series relative to traditional regression techniques?
2. Univariate Time Series Models
   - ARMA(p,q) models
   - Seasonality
   - Stationarity/Integration (d)
   - Forecasting
   - Intervention Analysis
3. Examples and Code

What is Time Series Analysis?

- Set of statistical procedures used when trying to understand or model social processes in terms of change over time.
- Primary unit of analysis is a date and/or time rather than, for example, observations of individuals or groups.
- Has a long history, but underutilized in communication research. Increasingly necessary and useful.

What Makes Time Series Analysis So Different?

Time series takes what you already know about OLS regression and inverts the logic to resolve problems introduced by temporally ordered data.

- A core assumption of traditional modeling approaches is that each observation in a dataset is independent from one another.
- Time – by definition – introduces a dependency between all of our observations. The current value of a variable is, in part, a function of previous values of the same variable.
- Time series analysis prompts us to:

1. Formally *disaggregate* and *identify* these temporal properties.
2. Model their functional form and *filter* them out of your series.
3. *Inspect* the model residual and tweak specification as necessary.

What Makes Time Series Analysis So Different?

This process was popularized by Box and Jenkins (1970; 1976), who formally outlined these procedures.

- Each individual time series for a variable is thought of as a combination of trends, cycles, shocks, and stochastic processes.
- The ultimate goal is to account for these temporal dependencies *within* and *between* social processes, allowing us to more closely model the true *data-generating process* (DGP).
- Given the right data, time series techniques allow us to speak more directly to questions about agenda-setting and the evolution of frames and discourse over time.

Seems Complicated, What About Simpler Techniques

Formal time series methods are more complicated, but yield more efficient and less biased estimates, reducing Type I error (false positive).

Three other techniques are popular:

1. "Controlling" for time as an IV a model.
2. "Detrending" data by regressing linear time trend on dependent variable.
3. Generalized Least Squares (GLS) procedures like Prais-Winsten/Cochrane-Orcutt.

A *Theory-as-Equations* approach helps understand why these approaches are flawed.

Theory-as-Equations

Theoretically, these approaches treat temporal dependence as a *statistical nuisance* to be corrected rather than a *theoretically substantive* component of the model-building process.

Empirically, these approaches generate *static* models rather than *dynamic* ones by imposing deterministic assumptions about the influence time has on the social process of interest.

This requires us to eat our vegetables and talk about the mathematical foundations of time series analysis: *differential equations*.

Basic Univariate Difference Equations

- Example 1:

$$A_t = \alpha A_{t-1} + \varepsilon_t$$

- Example 2:

$$A_t = \alpha_0 + \alpha_1 A_{t-1} + \alpha_2 A_{t-2} + \beta t + \varepsilon_t$$

ARMA(p, q) Models

ARMA stands for an *autoregressive moving average model* that accounts for two temporal properties of a series:

- **AR process** - a long-term process where shocks affect all future observations at an exponentially decreasing rate. While shocks "leak" from the system, they never dissappear entirely.
- **MA process** - a short-term process where shocks affect future observations for exactly q periods before disappearing entirely from the system.

Each of these is represented by the *order* of the AR or MA process, hence p and q. While often only an AR or an MA process is present, mixed models are not uncommon.

Identification: ACF and PACF

To *identify* the order of your ARMA model, start by inspecting the plots of the *Autocorrelation function* (ACF) and *partial autocorrelation function* (PACF) of the time series.

- ACF: measure of the correlation between y_t and y_{t+k} where k is the number of lead periods into the future.
- PACF: measure of the correlation between observations that are k units apart, after the correlation at intermediate lags has been controlled for.

The patterns in these correlelograms helps us establish what kind of processes are at work, providing a starting point for testing ARMA models.

Identification: ACF and PACF

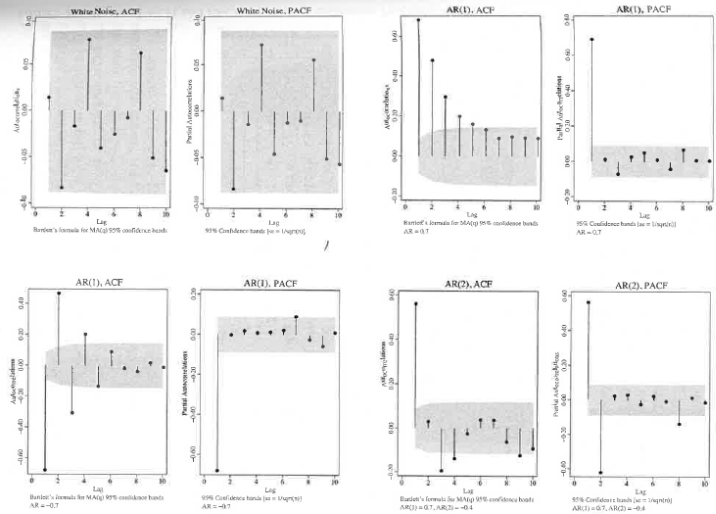


FIGURE 2.2. Realized ACFs and PACFs of Various Univariate Processes.

Figure 1: Archetypes of ACF/PACF plots, Box-Steffensmeier et al. (2014)

Identification: ACF and PACF

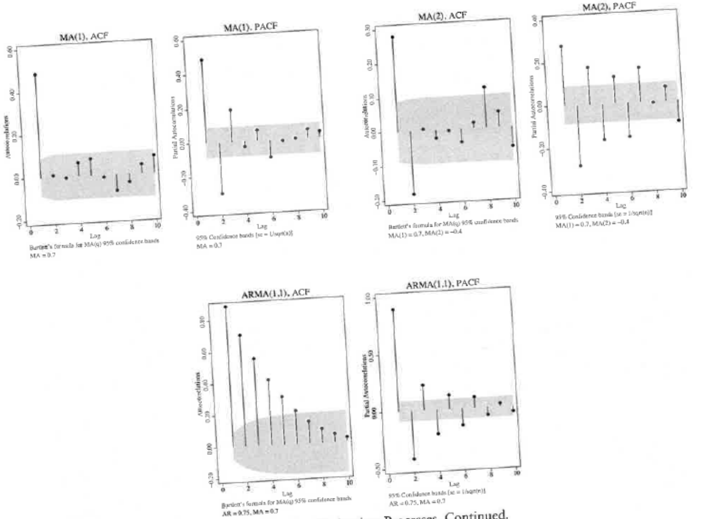


FIGURE 2.3. Realized ACFs and PACFs of Various Univariate Processes, Continued.

Figure 2: Archetypes of ACF/PACF plots, Box-Steffensmeier et al. (2014)

Seasonality

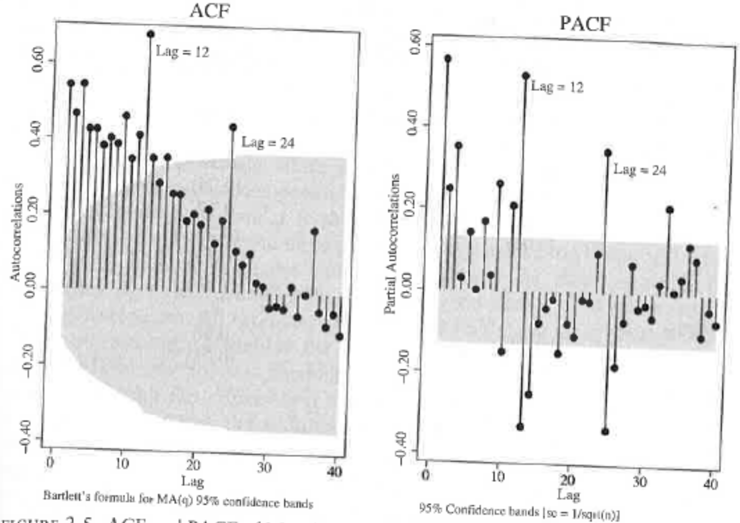
In addition to AR and MA processes, time series data may also have a seasonal component.

Seasonality refers to any cyclical fluctuation in a series that recurs at the same phase of the period of time.

Examples:

- Ice cream sales always spike during the summer and fall during the winter.
- Violent crimes reach their peaks during summer months.
- Twitter posts may have a daily cycle, peaking in the afternoon/early evening and falling late at night.
- News cycles may follow a weekly pattern with news dumps on Friday and fewer stories published on weekends.

Seasonality



Bartlett's formula for MA(q) 95% confidence bands

95% Confidence bands [se = 1/sqrt(n)]

FIGURE 2.5. ACF and PACF of Monthly Homicide Data.

Figure 3: Seasonality in ACF/PACF plots, Box-Steffensmeier et al. (2014)

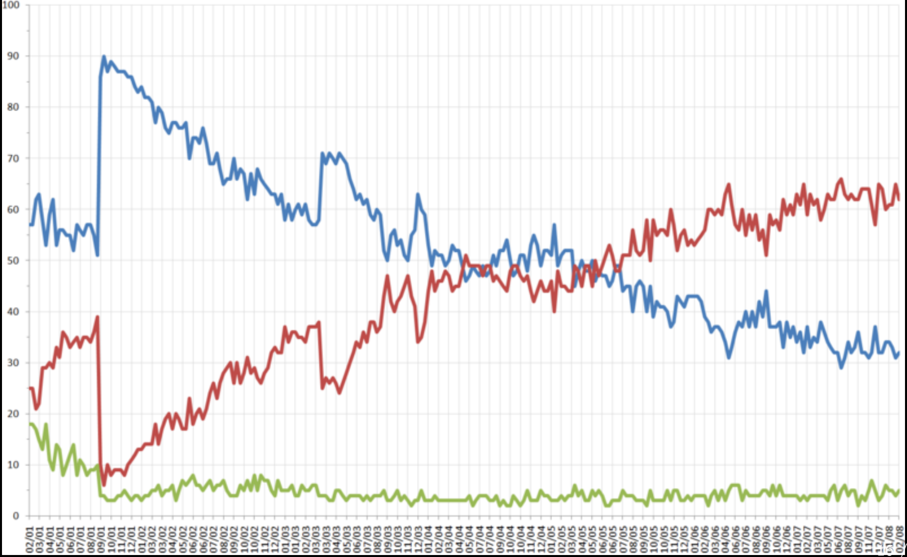
Stationarity/Integration (d)

A core assumption in most time series models is that the data in question is *stationary*.

- Also referred to as whether or not a series is *integrated* or if it contains a *unit root*.
- A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. When a shock enters the system, returns to equilibrium.
- Rather than "leaking" out of a system over time, shocks are integrated into the series and accumulate over time. In other words, the series has a *permanent memory* of shocks to the system.
- If the assumption of stationarity is violated, spurious results and model misspecification are inevitable.

Testing for Integration

Sometimes a visual inspection is all you need:



Testing for Integration

There are lots of formal unit root test statistics to choose from but the two most popular are:

1. Augmented Dickey-Fuller test - null hypothesis is that the series is non-stationary.
2. KPSS test - null hypothesis is that the series is stationary.
3. Variance Ratio Test - the variance of a series with a unit root grows linearly over time. Thus, the ratio of the variances, separated by k periods, should equal one. If they deviate from one, it is evidence the series does not contain a unit root.

The results of these tests determine whether the I(d) term in an ARIMA model takes on a value of 0 or 1. Double unit roots and multiple structural breaks are possible, are rare.

So What Do I Do If I Have Unit Root?

If you identify a integrated process, the recommended solution is transform the serie by *first differencing* the data.

This is done by subtracting the value of y_{t-1} from the current value, y_t.

Think of it as similar to taking a derivative over time or like applying a logarithmic transformation to stabalize the variance of a variable.

But, integration is thought of as a binary process, a series is either nonstationary or stationary. But we know that there is a huge gulf between those two things.

Fractional Integration

Not going to get into this much other than to say that I(d) *can* take on values between 0 and 1, which may be appropriate if you have a *long memory* process that is not fully-integrated, but clearly have non-stationary properties at work.

This is referred to as an ARFIMA model.

However, estimating d is complicated and requires a much higher sample size to generate reliable and valid estimates.

Fractional Integration

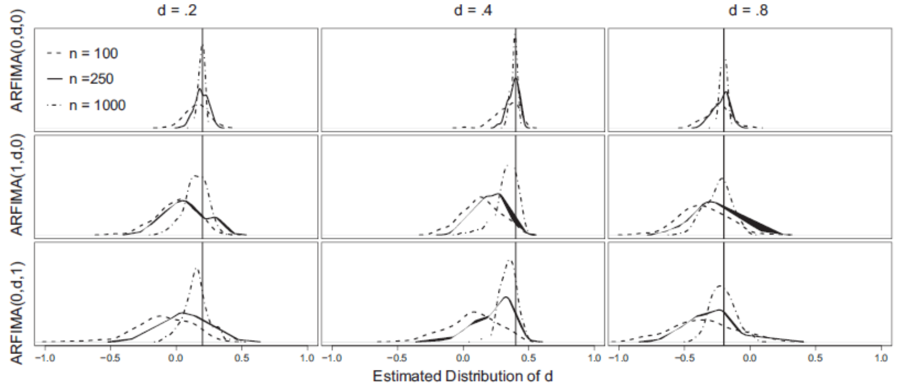


Fig. 1 Distributions of Estimates for *d*.
Each panel shows the distribution of the exact maximum likelihood estimates of *d* from the simulations for samples of size $t = 100$ (dashed line), $t = 250$ (solid line), and $t = 1000$ (dotted line). The solid vertical line in each plot represents the true value of *d*. Details of the simulations are given in the text.

Figure 5: ARFIMA simulations from Keele, Linn, & Webb (2016)

Intervention Analysis

Once armed with what we now know about the data-generating process of a time series, we can perform tests about how events may impact the process.

By modeling the dynamics of a proposed intervention in the series, we can put together quasi-experimental designs that test (a) whether events have an effect and (b) the functional form of that impact.

These can be used to understand whether the shock to the system altered the underlying data-generating process (*structural break*) or how an event ripples throughout the series over time.

Intervention Analysis

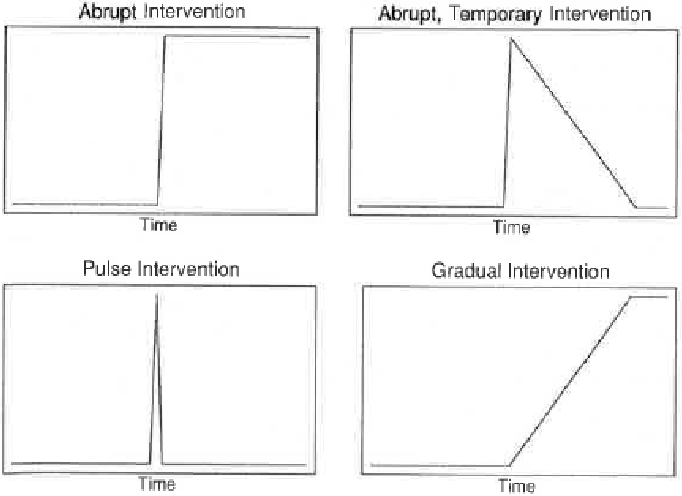


FIGURE 2.10. Types of Interventions.

Figure 6: Types of Interventions, Box-Steffensmeier et al., 2014

Forecasting

One of the biggest benefits of time series analysis is the ability to dynamically predict future values based on the temporal properties of your data.

Forecasts provide information about the direction, movement, and timing of a series into the future. Think of it as a form of curve fitting.

Depending on the nature of your data, these can yield practical information about how well a model performs and what one might expect to see.

Examples and Code

For these examples I'll be using the `AirPassengers` dataset in `R`, which contains the total number of airline passengers from 1949 to 1960 by month.

We will use this data to test our ARIMA modeling skills. First, lets just look at the plot of observed values.

Plot Observed Values

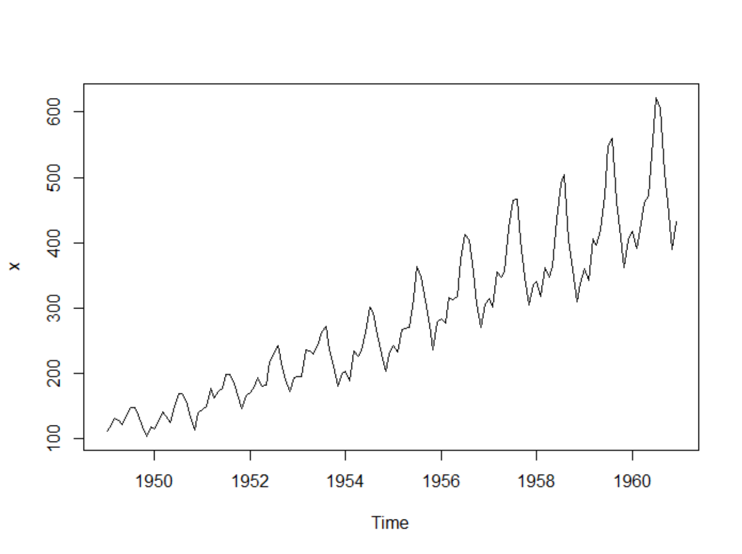


Figure 7: Plot of Observed Values

Inspect ACF/PACF Plots

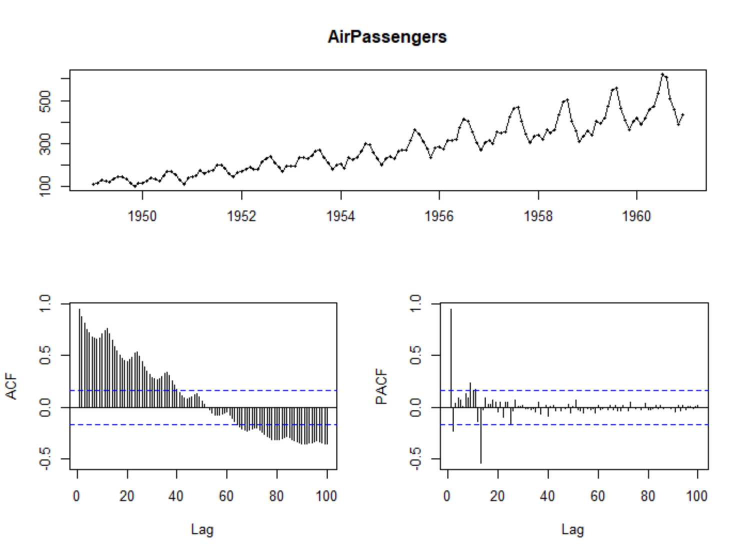


Figure 8: ACF and PACF Plots

Decompose Series

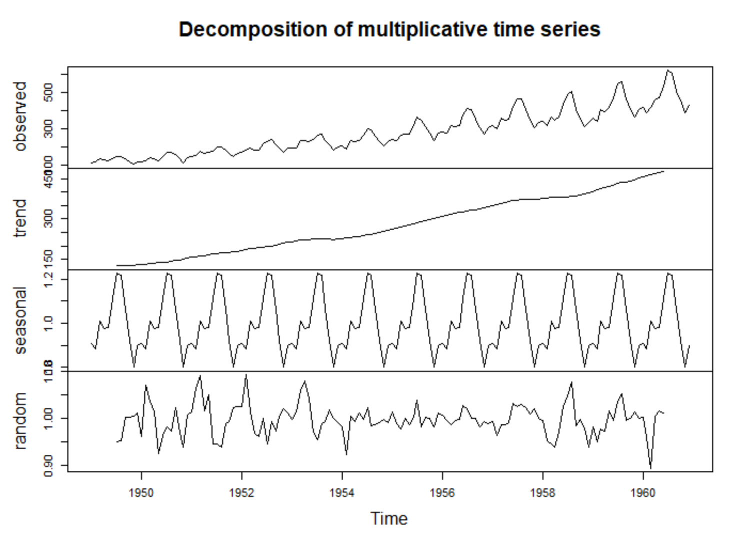


Figure 9: Decomposed Time Series

Lets Jump in and Build Our Own Model!

We are going to move to RStudio and walk through a few basic commands to build an ARIMA model for this data and forecast it!