PROJECT TEMPLATES FOR DAILY TEMPERATURE AND DAILY RAINFULL

Time series and regression analysis student projects can use weather data. The NEAS web site has Excel files for 1,000+ locations of daily

- high and low temperatures in degrees Fahrenheit and
- precipitation (rainfall) in hundredths of an inch.

Your student project may fit ARIMA processes to daily temperature or rainfall.

Take heed: This project template discussed daily temperature. Rainfall is more complex, since the error term is not normally distributed. We explain some rainfall items, but you must do independent work. Keep in mind two things:

- Daily temperature patterns are similar (though not identical) in many locations. Rainfall patterns differ by weather station.
- A sharp increase or decrease in the temperature is often a sign of rainfall. You can try a structural model: regress rainfall on the absolute value of the change in temperature and fit an ARIMA process to the residuals.

The project template on daily temperature is for candidates who need more guidance and have difficulty designing their own projects. We expect candidates who choose this project template to provide complete write-ups of the statistical reasoning for ARIMA modeling.

- We explain all the elements of the student project.
- We provide illustrative work-sheets and Excel data files.
- We mention several topics to consider (missing data, seasonality, warming)
- We do *not* put the pieces together in the illustrative work-sheets.
- You must understand the statistical reasoning to complete the student project.

Global warming is a much discussed topic. Many web sites analyze temperature changes and other historical weather statistics. We encourage you to explore other ideas for student projects besides ARIMA modeling of daily temperature.

Take heed: The weather service provides data bases for Fortran and SAS applications. We have placed time series data on the NEAS web site in Excel (CDS) formats. Download a compressed file, extract one time series, open it in Excel, then save it in Excel format.

Weather data are well suited to ARIMA modeling, since the patterns are seasonal with strong autoregressive properties. Several candidates have modeled daily temperature or daily rainfall.

Illustration: Ambitious candidates may do a project using hourly temperature, combining daily seasonality with yearly seasonality, strong autoregression, and cycles from warm/cold front movements. We provide data from one weather station in text fromat.

ISSUES TO CONSIDER FOR STUDENT PROJECTS ON WEATHER

Daily temperature is seasonal: high in the summer and low in the winter.

- The expected first differences are positive half the year and negative half the year.
- From mid-January to mid-July, the daily temperature increases, and vice versa in the other half-year.

De-seasonalize the data and fit an ARIMA process.

- For business sales and insurance premiums, volume for a given month in Year Z is
  - highly correlated with the volume in the same month in Year Z-1
  - less correlated for the same month in Year Z-2, and so forth.

  We use an ARIMA process with a 12 month autoregressive parameter.

- For weather, the daily temperature for a given day in Year Z is correlated with the daily temperature for the same day in all previous years. We de-seasonalize the data; we do not use a 365 day autoregressive parameter.

The volatility of daily temperature overwhelms the seasonal pattern. The first differences may be positive 55% of the time from mid-January to mid-July and negative the other 45%. The time series is a jagged curve: high stochastic variation overlaid on a smooth cycle.

To smooth the seasonal pattern, use a multi-year centered moving average.

Illustration: For the expected temperature on January 10, we might use a fifty year average of the average daily temperature on January 3 through January 17. The average of 50 × 15 days = 750 days gives a smooth annual cycle.

Take heed: The type of average depends on the length of the time series, such as

- For a 110 year time series, use a 110 year average with 7 days in each year.
- For a 20 year time series, use a 20 year average with 35 days in each year.

You may adjust for long-term trends and cycles in daily temperature.

Take heed: For rainfall, use a longer-term average, such as 100 years with 25 days in each year.

Compare your averages with those published by the weather service. Weather.com shows daily averages for every location in the U.S.; see their monthly forecasts.)

ARIMA PROCESSES

Fit an ARIMA process to seasonally adjusted daily temperature. Actual weather forecasts don't use ARIMA processes, but the ARIMA fitting makes a good student project. The ARIMA process varies by location, so each student project gives a different model.

An AR(1) or AR(2) process fits well, with perhaps an MA(1) component as well. As you write the student project, comment on what the model means.

An AR(1) process says that today's temperature depends on yesterday's temperature. Temperature changes incrementally, not suddenly, so we expect a positive ϕ_1 parameter. The daily temperature is *not* a random walk or a white noise process.

Illustration: Suppose the long-term mean for today's temperature is 65°, with the actual daily temperature ranging from 50° to 80°. If yesterday's temperature was 80° F, today's temperature is more likely between 65° to 80°, than between 50° and 65°.

The daily temperature is mean reverting, where the mean changes every day.

Illustration: Suppose the multi-year July 15th daily temperature is 90°. If the July 14, 20X8, daily temperature was 70°, we expect the July 15, 20X8, daily temperature to revert towards 90°. The AR(1) process determines the strength of the mean reversion.

An MA(1) process with a negative θ_1 parameter says that if the actual daily temperature yesterday was higher (lower) than expected, we expect today's daily temperature to also be higher (lower) than previously expected.

Illustration: Suppose the weather forecasts for Monday and Tuesday are 65°, the long-term mean is 65°, and the ARIMA process is AR(1) with $\phi_1 = 50\%$.

~ If the actual Monday daily temperature is 60°, the AR(1) forecast for Tuesday is 62.5%.
~ The –5° residual on Monday suggests a cold front is advancing. If weather patterns take an average of two days to change, the weather may stay cool (or get colder) on Tuesday, and we forecast 61.5° instead of 62.5°. This is a moving average coefficient of 20% added to the ARIMA process.

An ARMA(1,1) process has much in common with an AR(2) process. We might use an AR(2) process with the following reasoning: If yesterday's daily temperature increased from the previous day's, we expect the increase to continue today. An AR(1) process with a 50% ϕ_1 parameter plus 20% of the change from y_{t-2} to y_{t-1} gives an AR(2) process with parameters of 70% and –20%.

All three processes make sense, and the optimal process may differ by location. A student project on daily temperature shows the power of the sample autocorrelation function. A hundred years of data gives a time series with 36,525 points. Random fluctuations are smoothed, and you can use various tests to decide among AR(1), AR(2), and ARMA(1,1).

COMMON ERRORS

Several candidates have done student projects on daily temperatures. A few of them used monthly averages, relying on our recommendations for interest rates. For weather studies, use daily figures, not monthly figures.

- If Monday is colder than usual, we expect Tuesday to be colder as well.
- If April is colder than usual, we don't expect May to be colder as well.

For good ARIMA models, we work with many years, giving thousands of data points. Excel handles all the data crunching for you.

COMMON ERRORS

De-seasonalizing the daily temperature for a single year removes the trend as well (if any). If the time series has long-term trends or cycles, use separate periods or detrend the data.

PROJECT TEMPLATE ON DAILY RAINFALL

Daily rainfall (precipitation) is an interesting time series project template. Candidates who live in rainy climes or who want a project with different properties may use daily rainfall.

(1) The daily rainfall time series is not a differentiable function, since many days have zero rainfall. We can model the time series with an ARIMA process, but the error terms are not normally distributed. Tests of significance assume a normal distribution, so be careful in your interpretation of the results. Bartlett's test is not appropriate for this time series.

Illustration: Suppose we fit an AR(1) process to daily rainfall, with $\delta = 1$ and $\phi_1 = 50\%$, in a location where rain falls half the days. The mean rainfall is $1 / (1 - 50\%) = 2$. It rains half the days, so the mean rainfall when it rains is 4.

(The units are integers, for simplicity. Think of them as tenths of an inch. Average rainfall on rainy days is 0.400 inches. Average rainfall for all days is 0.200 inches.

The residuals are not normally distributed. Since rainfall is zero half the days, the residual is a discrete figure on days of no rainfall.

~ If it doesn't rain on Monday, the forecast for Tuesday is $1 + 50\% \times 0 = 1$. If it doesn't rain on Tuesday, the residual is –1. If it does rain on Tuesday, the average residual is $4 - 1 = +3$.
~ If it rains on Monday, the *average* forecast for Tuesday is $1 + 50\% \times 4 = 3$. If it doesn't rain on Tuesday, the average residual is –3. If it does rain on Tuesday, the average residual is $4 - 3 = +1$.

Even on rainy days, the amount of rain is not a normal distribution. In rainy climates, days with just a few light showers are not common. Rainfall is either zero (no rain), or several units, such as three to five tenths of an inch.

Recommendation: We can use time series models, but hypothesis testing is harder. An ARIMA process does not model the time series adequately. Your student project may form an ARIMA process and show that the residuals do not have a normal distribution.

If rain is infrequent in a location, do not use daily rainfall. You can not properly evaluate the ARIMA models, and no model fits well.

(2) Rainfall is seasonal, but the seasonality is complex and varies by location. Use a longer time period and more smoothing to get the mean daily rainfall.

The number of days for the long-term average depends on the percentage of days with rainfall. A hundred days with rainfall is sufficient for the student project. If rain occurs one day in four, use 400 days. For 25 years of data, use a 17 day centered moving average.

If the average daily rainfall is not smooth, use a longer centered moving average, such as 25 days or 31 days. But be careful: a wider moving average loses information. If average rainfall differ between May and June and we use a 90 day moving average, we get smooth curves that have less information. If we use a 365 day moving average, we get even smoother curves but we lose all the seasonality.

Take heed: Examine the graphs of different moving averages. Choose an average that smooths stochasticity but retains seasonality.

The weather service provides mean daily rainfall for the month. Do not use this monthly mean for each day of the month, as it distorts the autocorrelations.

Illustration: Suppose the monthly means of daily rainfall are 4 for May and 2 for June. May 31 and June 1 might have the same rainfall of 3. But the seasonally adjusted figures are –1 for May 31 and +1 for June 1. This gives a negative sample autocorrelation for consecutive days with the same rainfall – an illogical result.

Determine the average daily rainfall for each day. Verify your averages with the monthly means of the weather service.

Illustration: Suppose the monthly means of daily rainfall are 4 for May and 2 for June. Your analysis gives an overall average of 6 for May and 8 for June. You have probably made an error. But if your overall averages are 3.6 for May and 1.8 for June, your results are close enough to proceed with the analysis.

(3) Rainfall and daily temperature are related. We can construct a simple regression model using daily temperature and daily rainfall and fit an ARIMA process to the residuals.

A change in the daily temperature often heralds rainfall, as a warm front pushes against a cold front. Your student project can regress daily rainfall on the first differences of the daily temperature and model the residuals with an ARIMA process.

The student project does not construct a real weather forecast. We don't actually forecast daily rainfall by an ARIMA process on the residuals of a regression on daily temperature. We encourage structural models (residuals from a regression analysis) for the student project because they allow us to apply more techniques from the on-line courses and they often produce better fitting ARIMA models.

Daily Temperature Files

The NEAS web site has files with high and low daily temperatures in degrees Fahrenheit and daily rainfall in hundredths of an inch for 1,221 national weather service locations in the continental U.S. The weather service collects daily temperatures from over 3,000 locations, and it publishes long-term historical file for these 1,221 locations.

The starting and ending dates vary by location. Most files begin before 1900 and end either at 2000 or at 2005.

- You don't have to use all the data; see the comments below on missing records.
- Use a long enough series that your results are robust. Test your model by comparing two periods, such as 1901-1950 and 1951-2000. If you obtain the same ARIMA process for both periods, your model fits well.

Location

To choose a location, start with the file *Station Names*.

To pick a particular station, highlight Column D (the state abbreviations) and find (control F) the state you want, such as "NY." Select one of the stations, such as *New York Central Park* or *Rochester Airport*. Each station has a code: 305801 = *New York Central Park* and 307167 = *Rochester Airport*.

The weather station are not in numerical order. Check which WinZip file has the weather station you want. The table showing the zip file for each weather station may not yet be posted on the discussion forum.

If you are not choosy, use a station with many years of data and few missing records.

DATA FILE

The data files are in 14 WinZip files, each of which contains up 120 station files.

- Download a WinZip file from the NEAS web site.
- Extract the file you want. The file shows the station code.
- The file is in CSV format (comma separated values). Open the file in Excel. Double clicking on the file should open it in Excel. With some Excel versions, you may have to start Excel, click on *OPEN*, specify the CSV file, and tell Excel that it is in CSV format.

Save the file in Excel format, so that your formulas, charts, and graphs are saved. CSV is a compressed format, so we can store records for 1,221 stations. For a student project with one or two stations, use standard Excel format.

Take heed: If you can't open a CSV file or save it in Excel format, post a question on the discussion forum. Most candidates can un-zip compressed files, but some machines lack the required software. We will place several plain Excel files on the discussion forum.

DATA INTEGRITY

Some observations are missing in almost every file. These records show a high or low daily temperature of –999 or 0. Correct the data one of two ways:

- If a single observation (or a short sequence of observations) is missing, interpolate with the surrounding values.
- If a long series of observations is missing, eliminate the rows.

Both corrections slightly distort the ARIMA model. With 100 years of data, a few missing values does not have a material distortion.

Take heed: Inspect the data to ensure its integrity. If you overlook missing values, your time series analysis may be invalid, and you will have wasted several hours.

Illustration: A candidate fits an ARIMA model to a 20 year time series of daily temperature. Three values are missing, with –999 as the reported high temperature. The average high temperature is 60° Fahrenheit. If the missing values for these days are not corrected, the averages for these days becomes about 10° Fahrenheit. The residuals for these three days are about 50° Fahrenheit, and the fitted regression line is distorted.

Illustration: Station 280325

Station 280325 (Atlantic City State Marina, New Jersey) has 48,212 records from January 1, 1874, to December 31, 2005. December 31, 1894, is missing data:

12/30/1894	33	17
12/31/1894	-999	-999
01/01/1895	29	16

Interpolate the high daily temperature as 31 and the low daily temperature as 16.5.

Take heed: The illustrative work-sheet (station 280325) shows the interpolated high daily temperatures in red.

The interpolations of daily high temperatures in this file are for 12/31/1894, 11/23/1980, 5/1/1987-5/4/1987, 7/18/1995, 7/20/1995, 9/18/1996,

Most missing values are at a single date. Some values are missing for several days.

Illustration: Station 280325 is missing data for 5/1/1987-5/4/1987. We interpolate for the missing values as

4/30/1987	65.0
5/1/1987	63.2
5/2/1987	61.4
5/3/1987	59.6
5/4/1987	57.8
5/5/1987	56.0

If values are missing for a long sequence of days, delete the rows.

Values are missing for 10/1/1998 – 6/30/1990. Delete these rows from your work-sheet. The one day-to-day change from 9/30/1998 to 7/1/1990 slightly distorts the ARIMA model, but it does not have a material effect on a time series of 48,000 values.

Take heed: The denominator for the daily average over all years may depend on the day. For Station 280325, the denominator is either 130 or 131. We use the *COUNTIF* and *SUMIF* built-in functions to form averages.

Take heed: If you delete rows, a 365 day seasonal autoregressive parameter is distorted. For daily temperature, we de-seasonalize the data; we do not use seasonal autoregressive parameters. See the discussion forum posting on seasonality.

DATES

This step is geared to candidates with no knowledge of Excel. If you use dates in Excel, you can perform the items here more efficiently.

The observation day is in column A in MM/DD/YYYY format (month/day/year format). You must extract the month and day to compute averages for seasonal adjustments.

- For dates after January 1, 1900, use the Excel month, day, and year built-in functions.
- For dates before January 1, 1900, use the Excel string built-in functions. The illustrative work-sheet uses the Excel *VALUE* function to convert strings to numbers.

The illustrative worksheet determines a "day of the year index" as month × 100 + day.

- January 13 becomes 113.
- November 26 becomes 1126.

This MMDD format is easy to understand. We use this number for the *COUNTIF* and *SUMIF* built-in functions.

Excel has a built-in function to convert dates to an index starting at January 1, 1900. If you are familiar with Excel's date indices, they may simplify your code. The instructions here are geared to new Excel users; they are simple, but not efficient.

Caution: Take care to form monthly averages correctly. Not all days of the year have the same number of observations if you eliminate some records.

Alternatively, use data from 1/1/1874 to 12/31/1997 for your student project. You must still correct a few missing numbers, but the interpolations should take a few minutes.

FORM DATES

To analyze seasonality, form the month, day, and year of each observation.

- For dates of 1/1/1900 and later, use the Excel built-in functions *MONTH*, *DAY*, and *YEAR*. If you have never used these functions, see the Excel on-line *HELP* and the examples in the file for Station 280325.
- For dates before 1/1/1990, use the Excel string functions to extract the month, day, and year. The file for Station 280325 has examples. Copy this code or write your own.

Illustration: For the date 3/18/1926 in cell A19070:

- The cell formula =*MONTH*(A19070) returns 3.
- The cell formula =*DAY*(A19070) returns 18.
- The cell formula =*YEAR*(A19070) returns 1926.

For the date 3/18/1874 in cell A78, the date functions return *VALUE!*. We manually extract the month, day, and year.

In the daily temperature files, dates before 1/1/1990, have the format MM/DD/YYYY.

- The cell function "=*LEFT*(A78,2)" returns "03."
- The cell function "=*MID*(A78,4,2)" returns "18."
- The cell function "=*RIGHT*(A78,4)" returns "1874."

The Excel string function return strings, not numbers. To convert them to numbers, use the Excel *VALUE* built-in function.

- The cell function "=*VALUE(LEFT*(A78,2))" returns 3
- The cell function "=*VALUE(MID*(A78,4,2))" returns 18
- The cell function "=*VALUE(RIGHT*(A78,4))" returns 1874

To compute average daily temperature for a given day, use the Excel *SUMIF* and *COUNTIF* built-in functions.

Take heed: If the date strings in Column A did not have ten characters each, we would search for the slashes in the string: =*SEARCH*("/",A78,1).

DATE INDEX

To de-seasonalize the daily temperatures, we assign each day of the year a unique value. For simplicity, we use the value $MONTH \times 100 + DAY$

We use this value for the countif and sumif functions.

DE-SEASONALIZE THE DATA

Seasonality has several forms. The discussion forum posting on *seasonality* explains the types of seasonality and the methods used to adjust for them.

The daily temperature depends on the time of year, not on the value at the same date one year ago. We de-seasonalize the data; we do not use a 365 day autoregressive parameter.

Illustration: The expected daily temperature depends on the time of the year. It may be 25° on February 15 and 95° on August 15. It does not depend on the daily temperature one year ago. If the daily temperature is 45° on February 15, 20X8, we expect it to be 25° on February 15, 20X9, not 45°.

Determine the average daily temperature by day of the year in three steps. The daily temperature records are in rows 12 through 47575.

Excel's *COUNTIF* built-in function counts the number of occurrences of a value. Place the formula "=COUNTIF(L12:L47585,L12)" in Cell M12. The formula says:

> Count the number of times that the value in Cell L12 occurs in the range L12:L47585.

This is the number of times the value 101 (January 1) appears in Column L.

To copy this cell formula to other cells, first make the range absolute. Change the formula to "=COUNTIF(L$12:L$47585,L12)"

Excel's *SUMIF* built-in function adds values in cells that meet specified criteria. Place the formula "=SUMIF(L12:L47585,L12,B12:B47585)" in Cell N12. The formula says:

> Add the values in the range B12:B47585 (the high daily temperature) for cells whose value in the range L12:L47585 equals the value in L12 (i.e., January 1).

For absolute ranges, change the formula to "=SUMIF(L$12:L$47585,L12,B$12:B$47585)"

The average daily temperature is Column N divided by Column M. Place the formula "=N12/M12" in Cell O12.

Copy the formulas in Cells M12:O12 to Cells M13:O376 (the row for December 31, 1874). Excel computes the average daily temperature for each day of the year for the first year.

Take heed: The simplest way to code the average daily temperature for each record in the file is to copy these formulas to Cells M13:O47585. For each row, we compute the number of times the day occurs in the time series, the sum of the daily high temperatures for that day, and the average daily high temperature for that day.

But *COUNTIF* and *SUMIF* are memory intensive functions. Excel searches through 47,574 records for the day and then adds the high temperature to the sum. Excel does $2 \times 47{,}574^2$ = 4,526,570,952 operations, which can bring your machine to a halt. We use a less memory intensive procedure to compute the average daily temperatures.

V-LOOK-UP TABLE

We create a *V-Look-Up Table* for the day index and the average daily temperature.

First we add a row for February 29, which does not occur in 1874. Enter the date 229 in Cell M11, and copy Cells M12:O12 to Cells M11:O11. Note that February 29 occurs 32 times in this file, not 130 times.

Select the cells L11:O376. We need columns L and O (not Columns M and N), but the extra columns in the *V-Look-Up Table* don't cause a problem.

- The first column in this table (Column L) is the date index.
- The fourth column in this table (Column O) is the average daily temperature.

Choose *INSERT* from the menu bar; choose *NAMES* > *CREATE*. Enter the name VlookupTable for the cells L11:O376.

Place the formula *VLOOKUP*(L377,VlookupTable,4) in Cell O377. The formula says:

Look up the value in Cell L377 in the first column of the VlookupTable (Cells L11:L376). From the matching row of the table, take the value in the fourth column (Column O), which is the average daily temperature.

Copy the formula in Cell O377 to Cells O378:O47585. Each row has the average daily temperature for its day.

Take heed: The instructions here mix simplicity and efficiency. The ARIMA modeling can be done more efficiently in VBA, in "R," in SAS, or in MINITAB. If you use VBA for the average temperatures, your spread-sheet will run faster. For daily correlograms of a 100+ year time series, use the VBA macro in the illustrative work-sheet.

CENTERED MOVING AVERAGES

We want the expected temperature to determine the seasonally adjusted temperature.

Illustration: Suppose the average daily temperature is 60° and the average temperature on February 15 is 30°. A temperature of 40° on February 15, 20X8, is a seasonally adjusted temperature of 40° + (60° – 30°) = 70°.

A 130 year average eliminates most random fluctuations in the temperature, but some remains.

Illustration: If the daily temperature has a standard deviation of 10°, the 130 year average has a standard deviation of 10° / $\sqrt{130}$ = 0.877°.

Recommendation: Verify this relation. Compute the standard deviation of daily temperature in January (using Excel's built-in function) and the standard deviation of the average daily temperature in January. The relation will not be precise, because the expected temperature is not the same on each day of January.

The average daily temperature should form a smooth sine-shaped curve, with a peak in July and a nadir (low point) in January.

Form a line chart using the chart wizard.

Take heed: The illustrative worksheets do not graph the data. Graphs are formed easily with the chart wizard. If you use line charts, set the chart wizard default to line charts, not bar charts, to save a step. Be sure to label the axes and add titles and legends before copying the charts to Word (or whatever text file you use).

The averages are not perfectly smooth, which distorts the fitted ARIMA model.

Illustration: The average daily high temperatures are 40.015° for February 13 and 42.085° for February 14. A year with high temperature of 41.05° on February 13 and 14 would have seasonally adjusted figures of +1.035° and –1.035°. The residuals of the ARIMA process are over-stated because of the random fluctuations in the averages.

Take heed: The strong seasonality, the remaining random fluctuation in the averages, and the high number of observations cause the fitted ARIMA model to fail the Box-Pierce Q statistic. With 47,574 observations in this time series, the standard deviation of the sample autocorrelations of the residuals should be about $1/\sqrt{47{,}574} \approx 0.46\%$. Random fluctuation in the averages causes the sample autocorrelation of the residuals to be too high.

We use centered moving averages to smooth the sequence of averages.

Take heed: Several types of centered moving averages can be used. The simplest method is an N day centered moving average. Pick an N that smooths the random fluctuations but keeps the true differences among days.

- A 3 day centered moving average keeps too much of the random fluctuations.
- A 365 day centered moving average leaves no differences among the days.

Try 7, 15, and 31 day centered moving averages. Graph the results for one year. Each year is the same, so your graph needs just 365 days.

- If the graph is jagged, with random fluctuations from day to day, use a longer centered moving average.
- If the centered moving average is too long, the graph is squeezed. The top and bottom of the graph seem flattened.

    - With a 1 day average, the graph looks like a sine curve. The apex looks like a hill and the nadir looks like a crevice. They show a maximum and minimum.
    - With longer averages, the apex turns into a plateau and the nadir turns into a shallow basin. The average residual is positive at the apex, and negative at the nadir. The positive serial correlation distorts the sample autocorrelations.

Some statisticians use weighted centered moving averages. For five points, you might use weights of 15%, 20%, 30%, 20%, 15%. Some statisticians use complex weighting systems.

You may vary the number of points over the year.

- At the apex (early July) and nadir (early January), you might use 7 points.
- At Spring and Fall (April and October), you might use 31 points.

Recommendation: No specific average is correct. Your student project should explain what you did, such as “A 50 year average leaves a somewhat jagged curve. I used centered moving averages of 3, 5, 7, … days. The curve becomes smoother until about 21 days, so I used a 21 day centered moving average.” We check if you understand the logic of moving averages, not if your choice is the same as the course instructor’s.

The illustrative worksheet shows a 31 day centered moving average.

- Place the formula =AVERAGE(O12:O42) in Cell P27 (January 16, 1874) – an average from January 1 to January 31.
- Copy this formula to the next 364 cells, ending on January 15, 1875.

Extend the centered moving average to the rest of the rows. Copy the formula to all cells in Column P except the first 15 cells and the last 15 cells. Excel’s *AVERAGE* built-in function is efficient, so we do not use the *VLOOKUP* function. We make several manual adjustments.

- The first 15 rows have no centered moving average. Copy the centered moving averages for January 1 – 15, 1875, to January 1 – 15, 1874.

- The last 15 rows have no centered moving average. Copy the centered moving averages for December 17 – 31, 2004, to December 17 – 31, 2005.

If you deleted rows from your time series because of missing values, replace the 15 centered moving averages right before and after the missing values.

Illustration: The illustrative work-sheet has missing values for 10/1/1998 – 6/30/1990. We replace the centered moving averages for 9/16/1998–9/30/1993 and 7/1/1990– 7/15/1990 with the correct centered moving averages for those days.

The centered moving average works well except at the apex and nadir of the curve.

- A 31 day centered moving average under-states the average daily temperature in July and over-states the average daily temperature in January.
- The seasonally adjusted daily temperatures are too high in July and too low in January.

The graph of the smoothed average daily temperature is too flat at the apex and nadir. You may fit the graph to a smooth curve or judgmentally adjust the averages.

Take heed: The illustrative work-sheet does not make these corrections. Examine the graphs of your time series and judge if a correction is needed. Do not get bogged down in second decimal place accuracy if the effect is not material. You might find that a 25 day weighted and centered moving average works fine, with no correction needed.

Take heed: If you have deleted many rows with missing values, it may be easier to use the *VLOOKUP* built-in function as we did for each day's 130 year average.

ADDITIVE VS MULTIPLICATIVE MODEL

Seasonal adjustments can be additive or multiplicative. Multiplicative models are common for other time series (not daily temperature).

Illustration: If sales in December 20X8 are $100,000, and December sales are 25% higher than those of the average month, the seasonally adjusted sales for December 20X8 are $100,000 / 1.25 = $80,000.

For daily temperatures, multiplicative vs additive models mean:

- *Additive model:* The seasonally adjusted daily temperature is the reported temperature *minus* the average daily temperature for that day of the year.
- *Multiplicative model:* The seasonally adjusted daily temperature is the reported daily temperature *divided by* the average daily temperature for that day of the year.

The illustrative work-sheet uses an additive model.

- A multiplicative model assumes the variance of the error terms is proportional to the long-term average.
- An additive model assumes the variance of the error terms is constant

Take heed: If the temperature scale is arbitrary, the variance of the error term is unrelated to the mean temperature. The Fahrenheit scale for daily temperature is arbitrary.

Illustration: Suppose the average daily temperature is 25° in January and 100° in August.

- We don't say the daily temperature is four times greater in August than in January.
- We don't assume the variance of the daily temperature is four times greater in August than in January.

The problem with a multiplicative model is even clearer if the average daily temperature in January is –1°, 0°, or +1°.

- If the average daily temperature in January is –1°, a temperature of +10° is a relativity of –1,000%. The same relativity in August is a daily temperature of –1,000°.
- If the average daily temperature in January is +1°, a temperature of +10° is a relativity of +1,000%. The same relativity in August is a daily temperature of +1,000°.
- If the average daily temperature in January is 0°, any temperature is a relativity of infinity.

Most seasonally adjusted daily temperatures (using the additive model) lie in the range (–20°, +20°). Examine the time series for two patterns before fitting an ARIMA model.

White noise: If the time series is a white noise process,

- A positive residual is just as likely to be followed by a negative residual as by another positive residual.
- A negative residual is just as likely to be followed by a positive residual as by another negative residual.

The mean seasonally adjusted daily temperature for each day is zero. The observations are also the residuals if this is a white noise process.

The pattern of residuals for daily temperature is not a white noise process. Positive and negative residuals come in streams, such as 8 positive residuals followed by 6 negative residuals. This pattern indicates an autoregressive process with $\phi_1 > 0$. The process may have more autoregressive parameters and may have moving average parameters as well.

Random walk: If the time series is a random walk, it should show no mean reversion. If a residual is +10°, the following residuals should also be about +10°.

The pattern of residuals here is clearly not a random walk. The mean reversion is strong.

- If a residual is large and positive, such as +10°, the next residual is generally lower.
- If a residual is large and negative, such as –10°, the next residual is generally higher.

Graph the time series to see these patterns.

Take heed: If you graph all 47,574 days with one line chart, you won't see the patterns. Select half a year (182 days) and examine the chart.

Take heed: The correlogram for the first year on the illustrative work-sheet suggests an ARMA(1,1) process. The daily temperatures for 1874 alone do not give the optimal model for the next 130 years. Be sure that your model is *robust*.

A robust model does not depend on the period. *Suppose that*

- A model fitted from 1874 figures alone differs from a model fitted from 1875 figures.
- A model fitted with 1874-1939 figures is similar to a model fitted for 1940-2005.

We infer that one year does not give a robust model, but 65 years gives a robust model.

Take heed: Every weather station differs. The time series for another weather station may differ in two time periods.

CORRELOGRAM

Take heed: Forming a correlogram from 47,574 observations is memory intensive. The response time varies with the memory and power of your computer.

- An old computer may not be able to handle the computations, or it may take half an hour to show the results.
- A high-powered computer with a fast processor will do the calculations in half a minute.

If you code the *SUBPRODUCT* built-in function for the first cell with the *OFFSET* feature and copy it to the other cells on a slow machine, you may wait an hour for the response.

The illustrative work-sheets show several sample autocorrelation functions. For your student project, use the chart wizard to form correlograms.

The first sample autocorrelation function uses one year of daily temperatures, to verify that the cell formulas are correct. Examine the following items:

The annual seasonality and stochasticity are obvious in the data. Long-term trends and cycles are not easy to see. We expect a stationary time series, with rapidly declining sample autocorrelations. From the tenth lag onwards, the sample autocorrelations should have about equal numbers of positive and negative figures.

Recommendation: The sample autocorrelation of lag 1 may range from 20% to 70%, depending on the location. Your student project may compare two weather stations:

- A coastal city with rapidly shifting temperature and a low sample autocorrelation.
- An inland city with slowly shifting temperature and a high sample autocorrelation.

Form a correlogram from two or three years of observations, to see how quickly your machine computes the sample autocorrelations. The response time depends on your machine's memory and CPU speed.

As you add years, the response time increases exponentially. If a correlogram for two years takes a minute, a correlogram for 50 years may take several hours.

Take heed: These warnings about response time are for older machines. A high powered laptop bought after 2006 will handle the computations.

Important: Unless your machine is powerful, set *CALCULATION* to *MANUAL* and remove the check-mark from *RE-CALCULATE WHEN SAVING.*

Even if you form a correlogram for all observations, compute the sample autocorrelations for a limited set of lags. If daily temperature is a stationary AR(1) time series with $\phi_1 \approx$ 50%, the expected sample autocorrelation should decline to 0.1% by the tenth lag.

Take heed: The illustrative work-sheet has sample autocorrelations that decline rapidly to zero for the 365 day correlogram and sample autocorrelations that stay above zero for about thousands of lags for the 130 year correlogram. You can distinguish a stationary from a non-stationary time series with 40 or 50 lags.

For the Box-Pierce Q statistic, use 40 or 50 lags, unless your computer is powerful. Use more lags if the result is indeterminate.

If you have daily temperature for 120 years, use three periods of 40 years each. If the three periods gives the same (or similar) ARIMA processes, your model is robust.

Take heed: We don't expect a ARIMA process more complex than ARMA(1,1) for daily temperature, and we presume the ARIMA process is the same for all the years. Long-term trends and cycles may causes non-stationary processes and different ARIMA parameters. Compute just the statistics you need for the student project. The illustrative worksheet uses the Atlantic City data to show the complications from non-stationary time series.

The illustrative work-sheet shows sample autocorrelations for one year of data: January 16, 1874, to January 15, 1875. We use temperature deviations, having an average of zero for all years. This simplifies your computations.

- Column Q is the seasonally adjusted daily high temperature. The average is zero for all 130 years. For the 365 days in this correlogram, the average is –2.006° (Cell Q26). [*Note:* The average is not exactly zero even for all 130 years; see below.]
- The –2° deviation for 1874 may be random fluctuation or a true trend or cycle. The 130 year time series shows a change of about +4° in total, giving a deviation of –2° for the first year. But daily temperature has a high standard deviation. Average temperature varies greatly from year to year. You analyze the pattern for your weather station.
- Column R is Column Q minus –2.006°, so the average is zero. We need a zero average to compute the sample autocorrelations. (The VBA macro automatically computes the deviations.)
- Column S has an index for the day. January 16, 1874, is day 1, and January 15, 1875, is day 365. This is *not* the date index used for the average daily temperatures, of 101 for January 1 and 1231 for December 31. This is an index used for the *OFFSET* function.
- Cell R27 has =SUMPRODUCT(OFFSET(R$27,0,0,365-S27,1),R28:R$391). This is the *SUMPRODUCT* of two series of 364 days each, with a one day lag between them.
- Copy this formula to Cells R28:R388. We don't form the sample autocorrelation for the last three cells. See the discussion forum posting on time series techniques.
- With a one year correlogram, we can copy the formula to 361 cells without causing Excel to slow down. With a 40 year or 130 year correlogram, we copy the formula to just ten or twenty cells to ensure that Excel can handle the computations.

We explain the cell formula for the *OFFSET* and *SUMPRODUCT* built-in functions.

OFFSET(R$27,0,0,365-S27,1) means an array of cells:

- The first cell (upper-left corner) is 0 rows down and 0 columns across from Cell R27 ⇛ this is Cell R27. We can copy this formula to any other cell in Column Q, and the array still begins in Cell R27.
- The array has a height of 365 – S27 rows. Column S is a row index, starting at 1 for row 27. The height is 365 – 1 = 364 rows. When we copy the formula to Cell Q28, the formula changes to 365 – S28 = 365 – 2 = 363.

The SUMPRODUCT(OFFSET(R$27,0,0,365-S27,1),R28:R$391) formula specifies the second array of figures as ending in R391 and beginning 364 rows earlier. When we copy the formula to Cell Q28, the second array of figures becomes R29:R$391.

Take heed: Once you grasp the logic, use the VBA macro for sample autocorrelations.

FULL CORRELOGRAM

We compute the full sample autocorrelation function for all 130 years in Columns T, U, and V. We note the adjustments made to various rows. Your student project will have other adjustments or corrections, but the logic will be similar.

Take heed: This sample autocorrelation function uses cell formulas, so that you can copy the formulas to your own student project. We recommend that you use the VBA macro instead, which is more efficient, easier to use, and less likely to freeze your worksheet. But some candidates do not want to use VBA, so we show a cell formula solution as well.

Column P has 31 day centered moving averages of the daily average high temperatures.

- For most days, we use a 31 day centered moving average.
- For the first 15 days, the centered moving average is not exact. Instead, copy the centered moving averages for 1/1/1875 – 1/15/1875 to 1/1/1874 to 1/15/1874.
- For the last 15 days, the centered moving average is not exact. Instead, copy the centered moving averages for 12/17/2004 – 12/321/2004 to 12/17/2005 – 12/321/2005.

A year and a half have missing values in this time series. The centered moving averages for the 15 days right before and right after the missing data are not correct. Instead, copy the centered moving averages from a year with no missing data to these rows.

- Place the formula =B12-P12 in Cell Q12. This is the deviation of the January 1, 1874, high temperature from the centered moving average high temperature for that day.
- Copy the formula from Cell Q12 to Cells Q13:Q47585.
- Compute the average of Cells Q12:X47585 and place it in Cell Q11. If the time series is infinite and no rows are missing from your time series, this average should be zero. If some rows are missing, the average may not be exactly zero.
- We copies the centered moving averages to 4 × 15 = 60 cells. The 15 days before the first row and the 15 days after the last row are like missing values.
- The average in the illustrative worksheet is -0.00027714°.

Take heed: The VBA macro adjusts for the non-zero average. The computations here are needed only for the cell formulas.

We adjust for the non-zero average.

- Column Y adjusts for the non-zero average. Place the formula =Q12-Q$11 in Cell R12. Copy this formula to Cells R13:R47585.
- Check that your adjusted figures have an average of zero. Copy cell Q11 to Cell R11. The new average is now zero.

Autofill column S with lags ranging from 1 in row 12 to 47574 in row 47585. A simple way to autofill is to place a 1 in Cell S12 and a 2 in Cell S13. Select both cells and drag the autofill handle on the lower right corner of cell S13 to Cell S47585.

Place the formula “=SUMPRODUCT(OFFSET(R$12,0,0,47574-S12,1),R13:R$47585)” in Cell T12. Be sure to use the correct figures for your time series.

- 47,574 is the number of observations in the time series.
- R$12 is the first seasonally adjusted daily high temperature.
- R$47585 is the last seasonally adjusted daily high temperature.

See how much time Excel needs to compute the *SUMPRODUCT*. Depending on the response time, copy the formula to the next ten cells or the next 100 cells.

When Calculation is Manual, press F9 to re-calculate.

The sample autocorrelations for Atlantic City are positive for the first 6,345 lags. They are negative for the next few lags, and then turn positive again. After 9,000 lags, the sample autocorrelations are equally likely to be positive as negative.

NON-STATIONARY

The time series of daily temperatures is not stationary over the full 130½ year period: 132 years from 1/1/1874 to 12/31/2005 minus 1½ years of missing data.

Your student project should identify why the time series is not stationary and make the needed adjustments to fit an ARIMA process.

A time series may be non-stationary for several reasons: random walks, trends, cycles.

A common reason a time series is not stationary is that it is a random walk. Stock prices, price levels, average claim severities, GDP, foreign currency exchange rates, and many other financial and actuarial time series are random walks. The time series is an autoregressive process with $\phi_1 \approx 1$.

Intuitively, daily temperature is a mean reverting stationary process. In any single year, your time series analysis shows a stationary process with a sample autocorrelation of 20% to 70% for the first lag. It is not a random walk.

A second common reason is that the time series has a trend. If the trend is stable (i.e., the same trend over the full time series), we take first differences (or logarithms and first differences) to create a stationary time series.

A third common reason is a change in measurements. Suppose the weather service thermometers used in the last 30 years gave readings 2° higher than average, and the thermometers used in the first 54 years gave readings 2° lower than average. The trend may be zero in each of three periods (first, middle, and last set of years), but the correlogram will have the pattern shown here.

A related reason is temperature cycles. Meteorologists find medium term and long-term cycles that may cause the patterns here.

The daily temperature is stochastic. It is hard to distinguish

- an actual change in the daily temperature from
- a change in the measurement of daily temperature.

Actual changes in the daily temperature may occur for several reasons:

- The daily temperature has long-term cycles. The earth is warming and cooling over long periods, as shown by ice ages that come and go.
- Daily temperature along certain coastal areas has short cycles from ocean currents.
- Daily temperature in large cities may reflect urban activity. Smog from autos and more heating in winters may raise urban temperatures even if rural temperatures don't rise.
- Human activities may cause a greenhouse effect that raises daily temperatures.

These items are disputed, and we do not know the causes of daily temperature changes over time. Your student project will not decide the disputes, but it shows the sensitivity of ARIMA modeling to slight changes in trends or means. Use the following steps.

Step #1: Graph the seasonally adjusted daily temperature. The random fluctuations overwhelm any trend or change in the mean. The figures range from –30° to +30°, and you see fluctuations of 30 to 40 degrees in short periods.

Step #2: Form 365 centered moving averages and form a new graph. Your line chart is now smooth. The figures range from –4° to +4°, and changes occur gradually. You can see changes from one year to the next, but these still seem like random fluctuation, not trend.

Step #3: Distinguish between a trend over the entire time series and separate time series with different means. Perhaps Atlantic City has lower than average temperatures for the first 54 years, higher than average temperatures for the last 30 years, and middle temperatures for the middle 45 years. This may be a slow trend over all 130 years that is obscured by random fluctuations or a change in the mean of the series but not a trend.

Your student project may proceed several ways.

- Detrend the time series: regress the seasonally adjusted daily temperature on the day, running from 1 (January 1, 1874) to 365.25 × 132 = 48,213 (December 31, 2005). The day index includes the missing values in the time series. If $\beta = 0.0001$, subtract 0.0001 × day index from each seasonally adjusted value. Re-analyze the time series: see if it is stationary and re-compute ϕ_1.
- Take first differences. The textbook often uses first differences. But if the time series is stationary, taking first differences gives a worse fit. Examine the first differences to understand the pattern in the time series.
- Divide the time series into periods. For discrete changes, separate periods are ideal. A change in the measurement tool is best modeled by separate periods. If the ARIMA models are similar for both periods, we don't need separate periods.

Take heed: No method works always. The trend in any period is unclear. Over the full 130 years, the daily high temperature rises a few degrees. But this is not a consistent rise of 0.03° each year. Some years are warmer and some are cooler. The daily temperature pattern for Atlantic City is the not the same as the pattern for another weather station.

SUB-PERIODS

The illustrative work-sheet shows four sample autocorrelation functions.

One correlogram uses all 47,574 observations and cell formulas, not the VBA macro. The next three correlograms uses the VBA macro.

- Candidates who have never used VBA and do not want to start now may copy the cell formulas, changing the parameters as needed.
- The VBA macro is more versatile, quicker, and less likely to freeze your work-sheet. You spend a bit more time at first if you are not acquainted with VBA, but you save much time if you form several correlograms.

The sample autocorrelations are 0.51442 for lag 1 and 0.26526 for lag 2:

- $0.51442^2 = 0.26463$, and $(0.26526 – 0.26463) / 0.26526 = 0.24\%$.
- These figures are consistent with an AR(1) process.
- The difference of a quarter of a percent is random fluctuation.

For the first group of 10,000 observations, the sample autocorrelations are 0.50885 for lag 1 and 0.26648 for lag 2.

- $0.50885^2 = 0.25893$, and $(0.26648 – 0.25893) / 0.26648 = 2.83\%$.
- These figures are consistent with an AR(1) process.
- The difference of less than 3% is random fluctuation.

For a middle group of 10,000 observations, the sample autocorrelations are 0.49628 for lag 1 and 0.24502 for lag 2.

- $0.49628^2 = 0.24629$, and $(0.24502 – 0.24629) / 0.24502 = -0.52\%$.
- These figures are consistent with an AR(1) process.
- The difference of half a percent is random fluctuation.

For the last group of 10,000 observations, the sample autocorrelations are 0.54844 for lag 1 and 0.27875 for lag 2.

- $0.54844^2 = 0.30079$, and $(0.27875 – 0.30079) / 0.27875 = -7.91\%$.
- The 8% difference is large for a time series of 10,000 observations.
- We infer either an AR(2) process (with $\phi_2 \approx 0.02$) or a non-stationary process.

ANALYSIS

Determining if the time series is stationary is not easy.

- For the full period of 130 years, the time series is not stationary, with positive sample autocorrelations for thousands of lags.

- For each subperiod of 10,000 observations, the correlogram approaches zero faster, but still too slow for a stationary series. The first hundred sample autocorrelations remain positive in each correlogram.

But the time series is not a random walk and the high volatility of the daily temperature makes it hard to identify a trend. The AR(1) model is robust:

- For each sub-period, the first two sample autocorrelations indicate an AR(1) process with $\phi_1 \approx 50\%$.
- We chose the sub-periods at random. Your student project should graph the average daily temperature for each year and select sub-periods with different means or trends.

We infer that daily temperature has an autoregressive process, with $\phi_1 \approx 50\%$. But daily temperature has medium-term cycles and trends. A correlogram of 10,000 observations shows a non-stationary time series.

FIRST DIFFERENCES

We examine the first differences when the time series is a random walk or has a trend. Taking first differences of a stationary autoregressive process does not improve the model.

Illustration: For an AR(1) process with $\phi_1 = 50\%$, the autocorrelation of lag 1 for the first differences is –25%.

The sample autocorrelations of the first differences of the daily temperature are –0.24345 for lag 1 and –0.17510 for lag 2. We do not use the time series of first differences.

An AR(1) process for daily temperature is reasonable, and the autoregressive parameter is stable through all sub-periods. We infer that the time series is an autoregressive process with $\phi_1 = 50\%$. But the time series has long-term cycles and possibly trends.

- In a 1 year time series, the cycle or trend is eliminated by the seasonality adjustment.
- In a time series of moderate length (20 years to several hundred years), the cycle makes the time series non-stationary.

Over a million years, long-term cycles could perhaps be modeled. Meteorologists believe the earth's weather fluctuates in a regular pattern. For the 130 year time series, we try to identify and offset cycles, not to model them.

EXCEL EFFICIENCY

Excel is ideal for medium size operations, such as fitting exponential trends to average claim severities. Many users set *CALCULATION* to automatic: Excel recalculates the cell formulas whenever an input cell changes or when the work-sheet is saved.

A daily temperature time series of 135 years has 136 × 365.25 = 49,674 observations. If your student project examines the ARIMA processes for high vs low temperatures, you may have two time series of 50,000 observations each.

The *SUMPRODUCT* formula for the first lag of a time series with 50,000 values adds 50,000 products, requiring 100,000 operations. The column of *SUMPRODUCTS* needs ½ × 50,000 × 100,000 = 2,500,000,000 operations. Two columns use 5 billion operations. Calculating these columns takes much time: between ten minutes and several hours, depending on the power of your computer. To work efficiently

- Set Calculation to Automatic. Use the F9 key to calculate cells. If you copy a formula from one cell to another, the value is not re-calculated until you press F9.
- Remove the checkmark from "Recalculate upon saving." If your worksheet takes half an hour to recalculate, don't recalculate every time you save.

You can work on your spread-sheet and save frequently without freezing your computer.

The time series analysis is a series of steps. You compute sample autocorrelations from raw temperature readings and from seasonally adjusted figures, using several ways to adjust for seasonality. You compute correlograms for the entire period or for sub-periods. You may use high or low temperatures.

Excel is so powerful that one tends to leave all the computations in a single work-sheet. But a work-sheet with billions of computations is likely to freeze, and you may lose hours of work. The recommendations below should speed your work and prevent loss of material.

We give a VBA macro to compute the sample autocorrelations. The macro runs quickly and does not leave cell formulas in the worksheet.

If you use cell formulas for the sample autocorrelations, use separate workbooks or worksheets for each part of your student project. Complete each part and save it, so you may come back to your saved version if the next part of your student project is in error.

Dates: convert Excel dates or string dates to a date index and the year. Correct errors and missing values in the temperature readings. Create a worksheet or workbook with the clean data and a date index for each row.

Illustration: Suppose the original weather data runs from 1/1/1870 to 12/31/2000. Convert a date of November 15, 1878, to a date index of 1115 and a year of 1878. Inspect the data

for errors and missing values. Interpolate or delete rows, as appropriate. Keep a record of the data you change.

Seasonal adjustments: convert raw temperature readings to seasonally adjusted figures. Use separate worksheets for high vs low temperatures. Examine the adjusted time series for trend and cycles with graphs and moving averages. Form time series of detrended, seasonally adjusted values.

The write-up explains the seasonal adjustment and any observed trends or cycles.

Correlograms: Form correlograms separately for each time series and period you examine. Try various adjustments for trend or cycles, and examine each period whose mean differs.

ARIMA fitting: Use linear regression or Yule-Walker equations for the autoregressive and moving average parameters. Form residuals for each model and test them using the Box-Pierce Q statistic and Bartlett's test.

INTERPRET THE CORRELOGRAM

The student project shows if you can interpret statistical techniques. Correlograms are hard to interpret if a time series is not stationary or has few observations

- The time series of daily temperature for 1874 is relatively simple.
- The time series of 47,574 observations (all 130 years) is not stationary. Interpreting the correlogram is not easy, and you must analyze the results for several periods..

We analyze the 365 day correlogram of daily temperature for 1/16/1874 to 1/15/1875.

The sample autocorrelations have two components:

- The true autocorrelations from the time series.
- The white noise error term (sampling error).

Determine the significance of the figures, or the measure of materiality. The white noise error term has a standard deviation of $1/\sqrt{T}$. The one year time series has 365 observations; the full Atlantic City file has 47,574 observations.

- For T = 365, the standard deviation is $365^{-0.5}$ = 5.23%.
- For T = 47,574, the standard deviation is $47{,}574^{-0.5}$ = 0.46%.

A 95% confidence interval uses ± 1.96 standard deviations.

- For a one year correlogram, random fluctuations distort the sample autocorrelations up to 1.96 × 5.23% = 10.25% (up or down) 95% of the time. A 10% sample autocorrelation may just be random fluctuation.
- For a 130 year correlogram, even a 1% sample autocorrelation is significant, since it occurs less than 5% of the time.

A common rule of thumb is to replace each figure by a range that is ±1 standard deviation about the observed value. The sample autocorrelations for the first lag or two have higher error terms, and we might use a range of ± 2 × standard deviation.

Illustration: Replace the first sample autocorrelation (lag 1) in the one year time series of 40% by the range 35% to 45% (or 30% to 50% using 2 standard deviations). The longer time series suggest a ϕ_1 parameter of 50%, not 40%. We assume the ϕ_1 parameter is 50% for 1874 as well, but the high stochasticity of daily temperature gave 40%.

The first sample autocorrelation of 50% in a 130 year correlogram is the range 49% to 51% (or 48% to 52% using 2 standard deviations). The 50% parameter is supported by the correlograms for sub-periods, so it is robust. The non-stationary time series reflects long-term trends or cycles, not a random walk.

The 50% sample autocorrelation for lag 1 seems reasonable. If it was hotter than usual one day, it will probably be hotter than usual the next day.

Spells of hot or cold weather dissipate after a few days. If Monday is hotter or colder than usual, Tuesday and Wednesday will also be somewhat hotter or colder, but the effect dies out by the weekend.

Take heed: An effect that dissipates in N days does not mean an autoregressive process with N terms. An AR(1) process has an infinite memory: the effect continues indefinitely, though it may be overwhelmed by the random fluctuation after a few lags.

Autoregressive vs Moving Average

The sample autocorrelation of lag 1 may reflect an autoregressive or a moving average process. In your project write-up, explain what the ARIMA process implies.

Illustration: We contrast AR(1) and MA(1) processes for daily temperature.

Scenario: Suppose the average daily temperature for May 5, 6, and 7 is 70°, the expected temperature for May 5, 20X8, is 70°, and the actual temperature is 80°. The sample autocorrelation of lag 1 is 40%.

For both AR(1) and MA(1) processes, the expected temperature for May 6, 20X8, is 74°.

- For an AR(1) process, the expected temperature for May 7, 20X8, is 71.6°.
- For an MA(1) process, the expected temperature for May 7, 20X8, is 70°.

The MA(1) process dies out in a day; the AR(1) process dies out gradually.

- The AR(1) process depends on the previous value.
- The MA(1) process depends on the residual.

Illustration: Suppose the average daily temperature for May 5, 6, and 7 is 70°, the expected temperature for May 5, 20X8, is 70°, and the actual temperature is 90°. The sample autocorrelation of lag 1 is 40%.

For both AR(1) and MA(1) processes, the expected temperature for May 6, 20X8, is 78°. Suppose the actual temperature is 75% on May 6, 20X8.

- For an AR(1) process, the expected temperature for May 7, 20X8, is 72°.
- For an MA(1) process, the expected temperature for May 7, 20X8, is 68.8°.

The MA(1) process says that if it was colder than expected one day, a cold front is moving in, and we should expect colder weather the next day.

An MA(1) process is not reasonable for daily temperature, but an ARMA(1,1) process is. The 40% sample autocorrelation of lag 1 may be $\phi_1 = 20\%$ and $\theta_1 = -20\%$.

- The autocorrelation of lag 1 is $\phi_1 - \theta_1$.
- The autocorrelation of lag 2 is $\phi_1 \times (\phi_1 - \theta_1)$.

Take heed: A student project on daily temperature may compare AR(1), AR(2), & ARMA(1,1) processes. For the AR(1) and AR(2) processes, the linear regression gives the autoregressive parameters.

- Use the Yule-Walker equations for an estimate of the ϕ_1 and θ_1 parameters.
- Model the time series with this ARMA(1,1) process.
- Compute the expected value and residual for each day,
- From the residuals, compute the Box-Pierce Q statistic for the ARMA(1,1) process.
- Compare this Box-Pierce Q statistic with those for AR(1) and AR(2) models.

For other student projects, fluctuations complicate the choice of ARIMA process. The 130+ years of daily temperature observations smooth the random fluctuations in the correlogram. Although the Yule-Walker equations do not give optimal parameters for short time series, they give reasonably exact parameters for a time series of 50,000 observations.

Take heed: Some of the sample autocorrelation reflects the long-term trend or cycle, or the joining of distinct time series. Your fitted model is more accurate if you eliminate the patterns in the time series.

AR(1) MODEL: VERIFICATION

Verify the β coefficient of the linear regression with the sample autocorrelation of lag 1 to ensure that you correctly form the sample autocorrelations and the regression analysis.

- A material difference indicates an error.
- The two figures do not match exactly. Reconcile the figures to confirm your work.

The illustrative work-sheet shows AR(1) models for several time series:

- 365 days (1/16/1874 to 1/15/1875): a practice model so you can replicate the fitting procedure with simple cell formulas and quick response.
- The full period (130+ years, or 47,574 values), which is not stationary.
- Various sub-periods: the illustrative worksheet shows three arbitrary sub-periods of 10,000 values each. You judge how best to form stationary time series,

The 365 day AR(1) model is fit with a regression analysis of 364 pairs of figures.

Take heed: The one period lag converts 365 values to 364 pairs of values. The last day (day #365) is not used as an independent variable, since the one year time series has no day #366 as a dependent variable.

- The β coefficient for the regression analysis is 0.402636.
- The sample autocorrelation of lag 1 is 0.39841.

The two figures are close but not identical. The β coefficient is (0.402636 – 0.39841) / 0.39841 = 1.06% higher.

Principle: The β coefficient should be higher by a factor of about 1/T, where T is the number of observations in the time series. In this example, 1/T = 1/365 = 0.27%.

Confirm that the difference is correct. Both figures are the sumproduct divided by the sum of squared deviations.

- The sumproduct is 7,290.336 for the sample autocorrelation and slightly lower for the linear regression (in this example).
- The sum of squared deviations differs, reflecting the degrees of freedom.

The sample autocorrelation has one more observation.

- The sample autocorrelation of lag 1 uses the sum of squared deviations for 365 days, which is 18,298.767.
- The linear regression uses the sum of squared deviations for 364 days. The last day (day #365) is not used as an independent variable, since the one year time series has no day #366 as a dependent variable. The last day (#365) has a deviation of -13.897

and a squared deviation of 193.130. The sum of squared deviations for 364 days is 18,105.636.

Dividing the sumproduct of 7,290.336 by 18,105.636 gives 0.402656. This is almost the same as the ordinary least squares estimator for β (0.402636). We have reconciled the two figures, and we proceed with the ARIMA analysis.

Some candidates wonder why the two figures are not identical. The difference between them is (0.402656 – 0..402636) / 0.40266 = 0.005%.

The numerator and the denominator of the sample autocorrelation and the ordinary least squares estimator for β depend on the values used for the mean.

- The adjusted sum of squared deviations above (18,298.767 – 193.130 = 18,105.636) used all 365 days for the mean deviation. The last deviation of –13.897 is high. Eliminating the last deviation of –13.897 raises the mean deviation by 13.897 / 364 = 0.03807. The revised sum of squared deviations used for the regression analysis is slightly lower: 18,078.701 instead of 18,205.636.

- The sumproduct of 7,290.336 uses the 365 day mean for both series: the first 364 days and the last 364 days. The regression analysis uses the mean of the first 364 days for the first series and the mean of the last 364 days for the second series. The sumproduct in the regression analysis is 7,278.895. The quotient of 7,278.895 / 18,078.701 = 0.402623.

Take heed: The reconciliation above is not needed for your student project. Just check that the ordinary least squares estimator of β in your regression analysis is close to the sample autocorrelation of lag 1.

AR(1) PROCESS

Copy the seasonally adjusted temperatures from Cells P12:P47585 to Cells W13:W47586.

- Use Paste Special and select *VALUES*. Cells P12:P47585 contain relative formulas. Ordinary copy and paste puts meaningless formulas in Cells W13:W47586.
- Be sure to paste the values one row down. The value in Column P, Row 12, goes to Column W, Row 13.

Use Excel's *REGRESSION* add-in or the *SLOPE* and *INTERCEPT* built-in functions to estimate the AR(1) model.

- Excel's *REGRESSION* add-in uses more memory. Depending on your machine's CPU, you may not be able to run the regression.
- The *SLOPE* and *INTERCEPT* built-in functions are more efficient. They can be used in VBA macros, allowing you to automate the work. You can run the regression analysis once for each year, to test if the time series analysis is robust.

Begin with a regression on two or three years of observations to be sure you use the Excel functions correctly.

- If you can use the *REGRESSION* add-in for all 130 years, Excel shows the residuals.
- If you use the *SLOPE* and *INTERCEPT* built-in functions to estimate the AR(1) model, create the residuals manually. Yo do this procedure for the ARMA(1,1) model, and you can test the validity of your code on the AR(1) model.
    - From the observed value in Period T-1 and the parameters of the ARIMA process, determine the forecast (expected value) for Period T.
    - From the actual value in Period T, determine the residual in Period T.

TRENDS AND CYCLES

We don't expect a consistent trend in daily temperature. The average daily temperature hasn't changed much over the past hundred million years. Estimates of global warming or cooling over the past hundred years suggest a change of 1 or 2 degrees Fahrenheit.

But daily temperature has strong medium term cycles. Meteorologist believe the earth has been slowly warming since the last Ice Age. Every century, the earth may warm or cool a few degrees. Cities (with smog) and coastal regions may have other cycles or trends.

Short term cycles from ocean currents and short term trends from urbanization and smog may have a strong effect on daily temperatures.

Illustration: The build-up in auto exhaust during the week in metropolitan areas may raise rainfall on weekends, as water vapor condenses on exhaust particles. The change in daily temperature is harder to model, since smog has both warming and cooling effects.

Ideally, we separate the changes in daily temperature into seasonality, long-term trends, long-term and medium term cycles, and random fluctuations.

For medium-term trends and cycles, split the time series into eras with no trend or cycle.

COMMENT ON SEASONAL ADJUSTMENTS

Fajlure to deseasonalize gives high autoregressive (ϕ) parameters and sub-optimal ARIMA models. The illustration below shows the rationale.

Illustration: Suppose the proper model is an AR(1) process, where μ is the average daily temperature for that day, ϕ_1 is 20%, the average daily temperature varies through the year from 25° F to 90° F, and σ is high.

The δ parameter varies from 25° F × 80% = 20° F to 90° F × 80% = 72° F, depending on the time of year. Instead of a δ parameter, the ARIMA process might use a centered moving average of twenty values.

An ARIMA model may have $\phi_1 = 20\%$ and $\phi_2 = \phi_3 = \ldots = \phi_{10} = \phi_{356} = \phi_{357} = \ldots = \phi_{366} = 4\%$. A better ARIMA model may have $\phi_1 = 20\%$ and values of 1% for 80 other lags, using 20 day moving averages from four years. An even better model may have $\phi_1 = 20\%$ and values of 0.1% for 800 other lags, using 20 day moving averages from 40 years.

This is wrong. Don't use autoregressive parameters to de-seasonalize the data.

- Separate the seasonality and the autoregressive process by deseasonalizing the data.
- Then fit an AR(1) or AR(2) process.