Baby names time series student project.

Some candidates have asked about this web site: “What sort of ARIMA models might we expect?” Here are some ideas.

If a person does something remarkable and is widely praised by the media, some couples may name their babies after this person. This is a stochastic event.

Suppose the name Rebecca is given to 1% of babies, on average. If a person named Rebecca does something remarkable in January 20X5, and she is praised by news anchors during 20X5, perhaps 2% of babies born in 20X6 will be named Rebecca. This is a stochastic event; nothing in the past history of baby names suggests a higher percentage of Rebeccas in 20X6.

We consider the effects on *future years*. One possibility is

- In 20X7, some news about the remarkable Rebecca lingers, and 1.1% of babies are named Rebecca.
- In 20X8, 1.0% of babies are named Rebecca, as before 20X6.

This process has a *memory of one period*. It is a moving average process, with $\theta_1 = -0.100$.

Another possibility is

- In 20X7, couples still like the name Rebecca, and 1.5% of babies are named Rebecca.
- In 20X8, 1.25% of babies are named Rebecca.
- In 20X9, 1.125% of babies are named Rebecca.

This is an autoregressive process with a $\phi_1 = 50\%$.

A student project may ask: “Do increases in a particular name dies out quickly, as in a moving average process, or do they die out slowly, as in an autoregressive process?”

Jacob: I looked at this web site, and the percentages of different names shows short term and long-term trends. The percentage of babies named Rebecca may be 1.0% in 20X5, 1.2% in 20X6, 1.4% in 20X7, and so forth. Which process is this?

Rachel: This process has a trend, so it is not stationary. Take first differences to remove the trend, and analyze the time series of the first differences.

Jacob: The trends are not constant. The incidence of a name may increase for 10 years, then remain level for 15 years, then decrease for 10 years.

Rachel: You may analyze this two ways.

- The first differences may be an autoregressive process with a slow decay. The expected first difference may be zero. If a stochastic event causes the first difference to be positive, it may decay back to zero over the next ten years. If another stochastic event causes the first difference to be negative in the seventh year, it may remain negative for a while, decaying slowly back to zero.
- If the processes differ in the two or three periods, you may separate the time series into two or three periods.

Jacob: Do we use all the names on the web site?

Rachel: You are not doing a thesis on baby names.

Select one, two, or three names. You will find it easier to select names that show different patterns. Model the time series for each name with an ARIMA process. Explain the statistical tests you perform.

You can select a name that shows different patterns in two or more periods. Model each period with an ARIMA process.

If one name particularly intrigues you, examine that name in more detail. Graph the pattern for the name, its first differences and second differences. Use correlograms and the statistical tests to select an optimal ARIMA process.

Jacob: Do we have to explain why a particular process makes sense?

Rachel: The student project is a statistical project. You demonstrate that you can model a time series with an ARIMA process. You can add comments about why a process makes sense, but that is not necessary.

Jacob: What non-statistical items should we be careful about?

Rachel: Baby names have large random effects. We do not expect perfect fits with the ARIMA process. High order processes, such as ARIMA(4,1,6) instead of ARIMA(1,1,0), is generally a poor solution.

Jacob: How can we tell which ARIMA process is the better solution?

Rachel: Suppose the more complex process fits better over an interval of 40 years. Fit the processes over the first 35 years and forecast the next 5 years. Very often, the more complex model has the better in-sample fit but the worse out-of-sample forecast. If so, we choose the more parsimonious sample with the better forecast.

Jacob: Do we expect the same ARIMA process to apply to all names?

Rachel: Most statisticians would say no. In any case, your goal is not to determine the proper ARIMA process for baby names. Your student project shows that you understand

how to use the statistical techniques. You can show this by a study of one or two baby names; you don't have to use them all.