Name: Noah Kohn
Course: Time Series
E-mail: noah.kohn@trustmarkinsurance.com

TS Project: Murders in Chicago

Introduction

Chicago has recently earned the ignominious designation of the murder capital of the United States. I decided to look at how the rate of murders in Chicago have developed over time, and whether or not it fits well with an autoregressive time series model. To evaluate predictive power, I fit the models using 2001-2009 data, and then test the predictions of the model against (known) 2010 data. This provides an empirical backdrop to the theoretical underpinnings. This test is performed via Monte Carlo simulations, using the mean of one-thousand runs of the resultant distribution as the prediction for each month.

Data

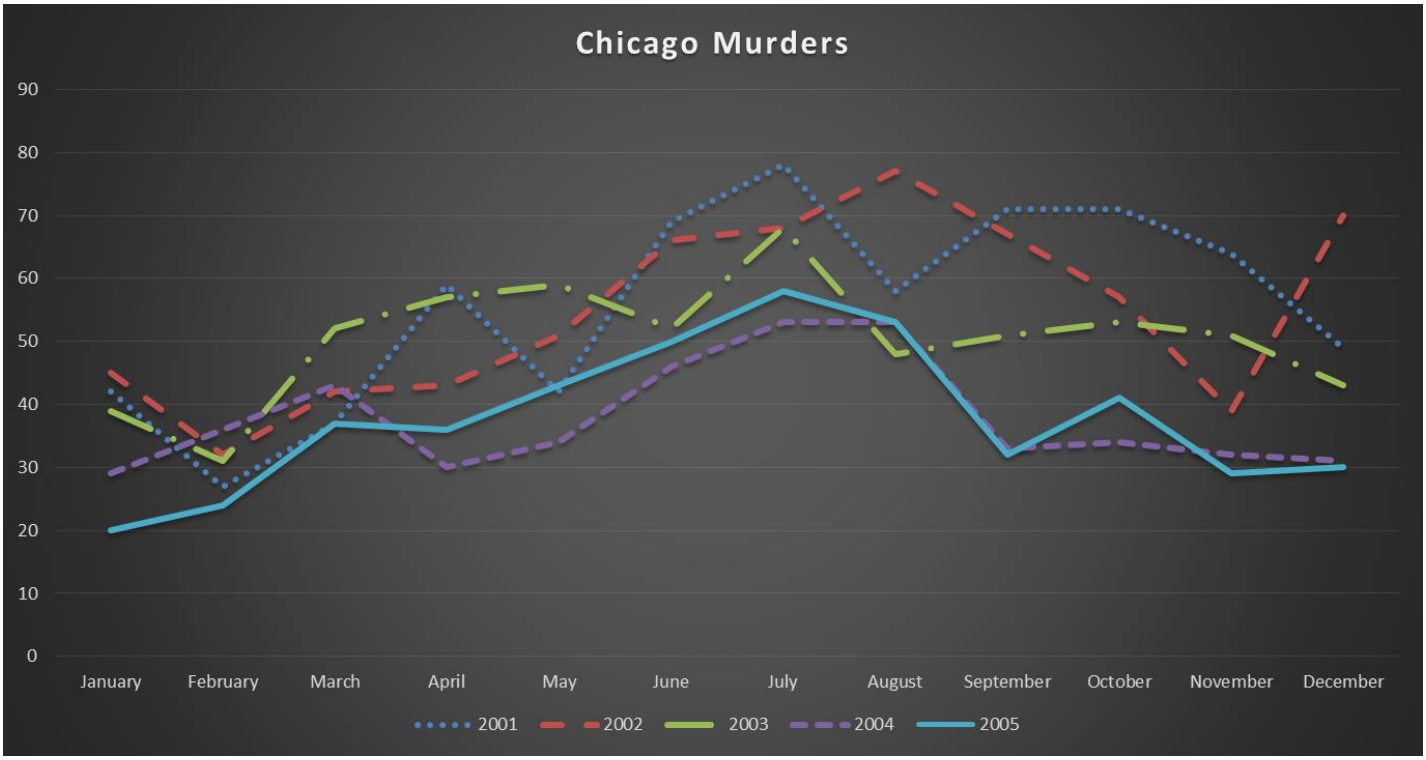
The data used is crime date from 2001-2010 in Chicago, IL. It is from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

All crimes are present, but I filtered for only homicide. The models I fit are an AR(1) and an AR(2). They are of the form $Y_t = \Phi_1 Y_{t-1} + e_t + C$ and $Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + e_t + C$ respectively.

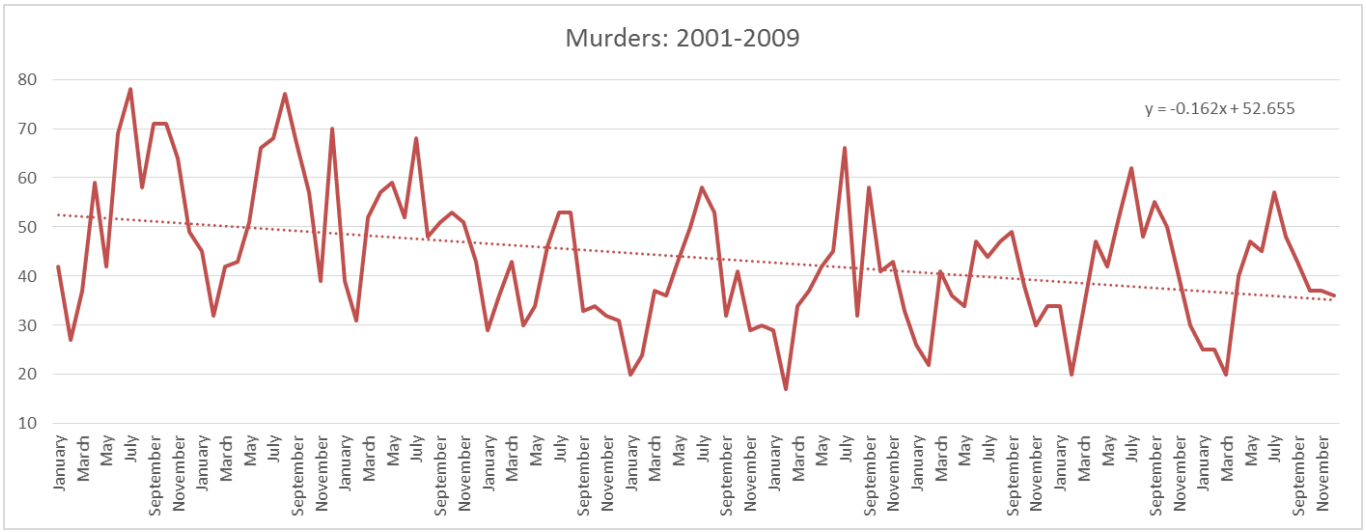
Seasonality

I suspected that there would be seasonality present in the data, on the theoretical basis that more crimes would be committed in warmer months than in colder months, as more people would be outdoors. Summing the murders by month and year, we see the following tabular data. Beneath, I overlay the murder rates across a representative sample of years. Though there is naturally some variability, we see a clear trend for more murders in the summer months, and fewer in the winter months—in every year but one the maximum occurs in July or August, and that one exception has only 2 more murders in September than August. Further, we can see that in aggregate the summer months have noticeably more murders than the winter months. For example, compare July to December or January.

	January	February	March	April	May	June	July	August	September	October	November	December	**Total**
2001	42	27	37	59	42	69	78	58	71	71	64	49	**667**
2002	45	32	42	43	51	66	68	77	67	57	39	70	**657**
2003	39	31	52	57	59	52	68	48	51	53	51	43	**604**
2004	29	36	43	30	34	46	53	53	33	34	32	31	**454**
2005	20	24	37	36	43	50	58	53	32	41	29	30	**453**
2006	29	17	34	37	42	45	66	32	58	41	43	33	**477**
2007	26	22	41	36	34	47	44	47	49	38	30	34	**448**
2008	34	20	33	47	42	52	62	48	55	50	40	30	**513**
2009	25	25	20	40	47	45	57	48	43	37	37	36	**460**
2010	22	22	31	46	46	49	42	57	31	36	33	23	**438**
Total	**311**	**256**	**370**	**431**	**440**	**521**	**596**	**521**	**490**	**458**	**398**	**379**	***5171***

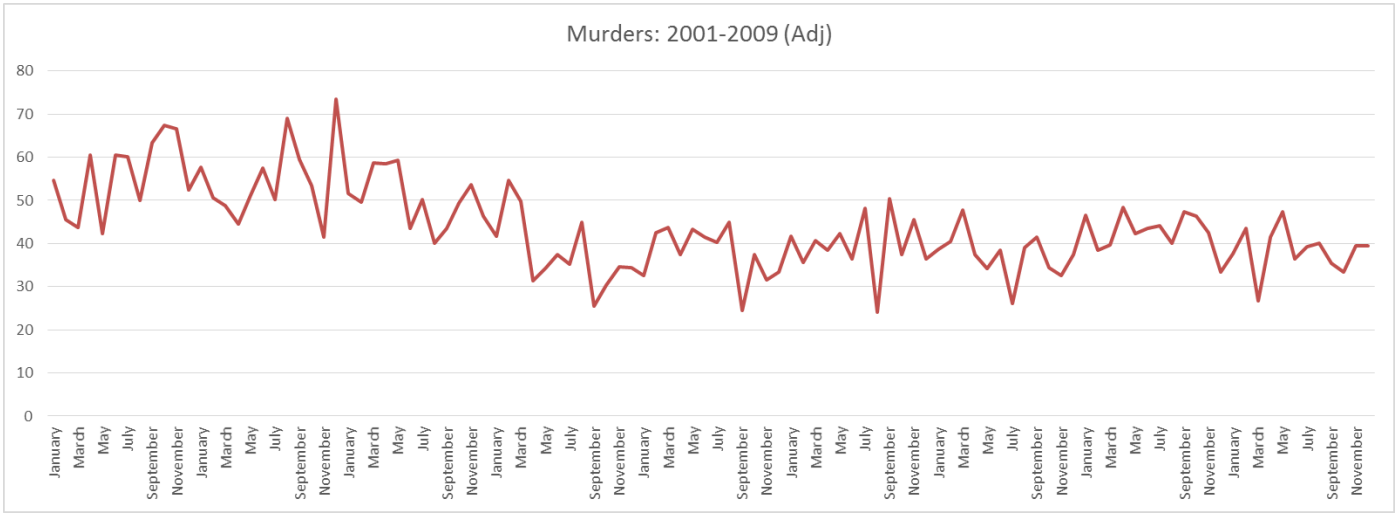


To adjust for seasonality, it is first necessary to examine the entire time series and notice any trend. In this data, there is a slight downward slope over time. Performing a linear regression on the data, we find that the trend is $Y = -0.162X + 52.655$. After determining the trend at each point, we consider the residual of the data point and the trend. For example, if there were 42 murders but the trend would expect 52.49, then the residual is -10.49.[1] The average of these monthly residuals, by month, becomes the seasonal adjustment factor for each month, which is then subtracted from the original series to arrive at the seasonally adjusted series. It is this resultant series which is then considered for the remainder of the analysis.



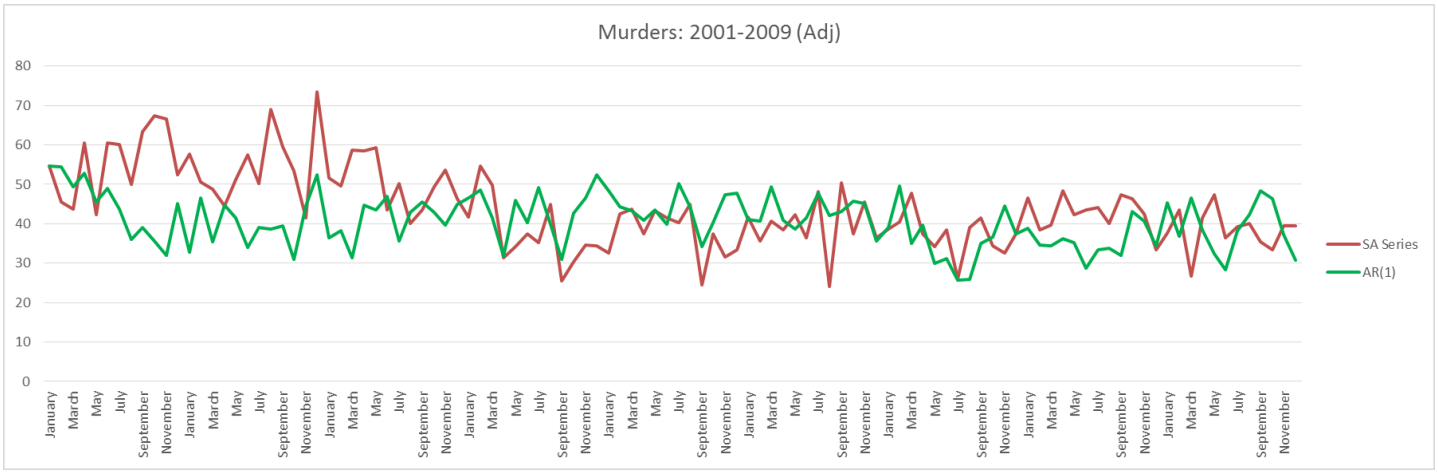
[1] Obviously murders can only be whole numbers, but for the sake of mathematical precision, decimal values are used throughout. When doing a prediction or speaking to real-world values, these would be rounded.

After smoothing for the seasonality, we end up with:



AR(1)

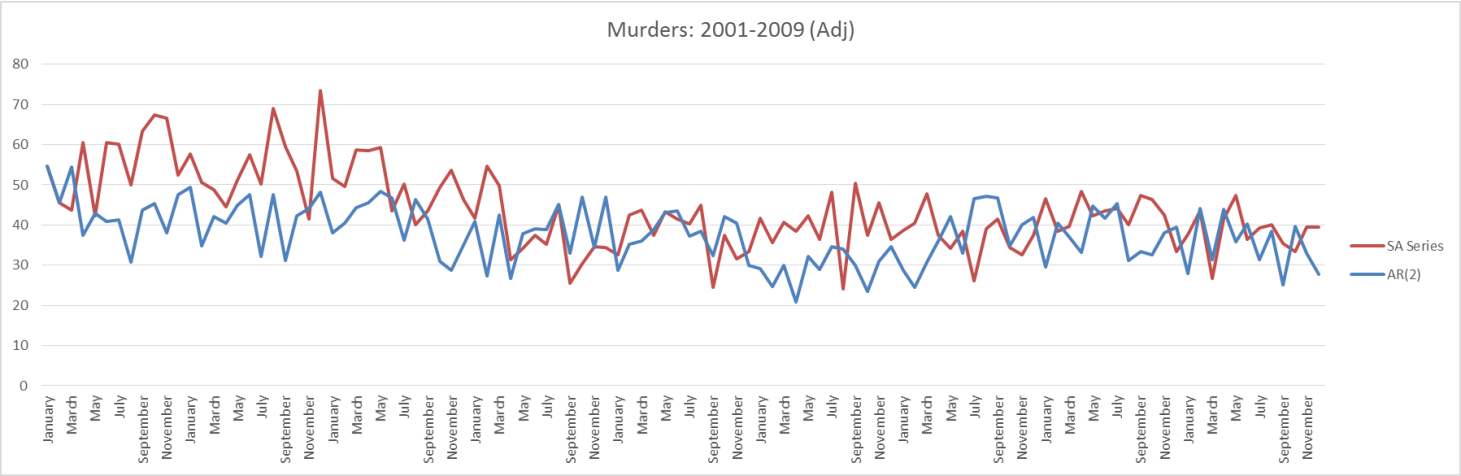
To fit the data to an AR(1) model, we need to estimate the Φ parameter from the data. This is accomplished by using the method of moments. For an AR(1) case, $\hat{\Phi} = r_1$ where r_1 is the sample autocorrelation at lag 1. To compute this, we compute the deviation of each observation from the average of all observations. The autocorrelation is then the sum of all offset products (the deviation at time *t* times the deviation at time *t-1*) over the sum of the squares of the deviation. In this case, $r_1 = 0.50$, leading to $\Phi = 0.50$. Because true AR processes have an error term, they are stochastic. This will be of importance when forecasting, because we are therefore best served by using Monte Carlo simulation to look at a probability distribution of possible future murder rates, as opposed to a single deterministic projection. One sequence of error terms gave rise to the following illustrative AR(1) process, fit to our murder rate data.



AR(2)

Fitting the AR(2) distribution is slightly more complicated. Now, we need to calculate the autocorrelation at lag 1 and at lag 2. Then, using method of moments leads us to estimates of $\hat{\Phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2}$ and $\hat{\Phi}_2 = \frac{r_2-r_1^2}{1-r_1^2}$. The calculation of the autocorrelation at lag 2 is identical to that performed at lag 1 (explained above) except that the product is between times *t* and *t-2*, instead of *t* and *t-1*. We find in this data that $r_2 = 0.55$, leading to $\Phi_1 = 0.305531$ and $\Phi_2 = 0.39319$.

Notwithstanding the same caveats regarding stochastic simulations, below is an illustrative example of the AR(2) process fit to the murder data.



Simulation

Using the data for the proceeding years (2001-2009) to develop the series, I then simulated 1000 runs of random error variables for each of the next 12 months. Using the average result for each month as the simulated variable, I think backed out the seasonal adjustment to return to a prediction on the same basis as the original data. The results are below. We see that both models can be calibrated to be fairly accurate. In both cases, the many years in which July had more murders than August pulls the results toward this sort of prediction via the seasonal adjustment. When, in 2010, the heaviest month was August instead of July, this resulted in large and opposite errors for both July and August.

	AR(1)						
Obs	Year	Month	Avg Simulated Murders	SF	Predicted Murders	Actual Murders	Error
1	2010	January	35.78	-12.61	23	22	1
2	2010	February	35.80	-18.56	17	22	-5
3	2010	March	35.64	-6.73	29	31	-2
4	2010	April	36.03	-1.45	35	46	-11
5	2010	May	35.81	-0.29	36	46	-10
6	2010	June	35.65	8.54	44	49	-5
7	2010	July	35.52	17.81	53	42	11
8	2010	August	36.18	7.97	44	57	-13
9	2010	September	36.42	7.58	44	31	13
10	2010	October	36.09	3.63	40	36	4
11	2010	November	36.11	-2.54	34	33	1
12	2010	December	35.94	-3.38	33	23	10
Total					431	438	-7

	AR(2)						
Obs	Year	Month	Avg Simulated Murders	SF	Predicted Murders	Actual Murders	Error
1	2010	January	36	-12.61	23	22	1
2	2010	February	36	-18.56	18	22	-4
3	2010	March	35	-6.73	28	31	-3
4	2010	April	36	-1.45	35	46	-11
5	2010	May	36	-0.29	36	46	-10
6	2010	June	39	8.54	48	49	-1
7	2010	July	40	17.81	58	42	16
8	2010	August	38	7.97	46	57	-11
9	2010	September	31	7.58	39	31	8
10	2010	October	39	3.63	42	36	6
11	2010	November	31	-2.54	29	33	-4
12	2010	December	32	-3.38	29	23	6
Total					431	438	-7

Both models are plausible, and arguments could be made for either. On the one hand, the sample variance for the AR(2) model is slightly lower, and there is perhaps theoretical justification for believing that both previous months have something to do with the current murder rate (in terms of a local trend). Personally, however, I would be more swayed by the principle of parsimony—given that there is not a strong difference between the models, I would choose the simpler model, the AR(1).