Chicago Crime Data

Joshua Gomberg
Prepared for Cotiviti
Data Source: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

In the interests of public safety, the city of Chicago posts the details of reported crimes as a publically available data set. From this data, predictive time series models may be constructed. The ability to forecast the prevalence of crime is of great importance to police departments. If an increase in crime can be predicted beforehand, then it can be mitigated through an increase in police presence. The types of crimes being committed and the locations where they occur are of particular interest to modelers. Predictive models incorporating location data can help identify specific regions within the city where additional police presence is needed. Crime classification data (sex crimes, drug offenses, theft, etc.) can help identify specific behaviors that police officers should be trying to identify as they patrol the streets.

For the purposes of this study, a time series model predicting the average daily number of reported crimes per month in Chicago as a function of time is constructed, as a generalized, illustrative example. Submodels based on location and crime classification can be built using analogous methods to the ones outlined here. The data was sampled monthly to incorporate local smoothing, and crime counts were represented as the daily average for each month to account for the fact that months vary in length. The resulting data is depicted in Figure1 below:

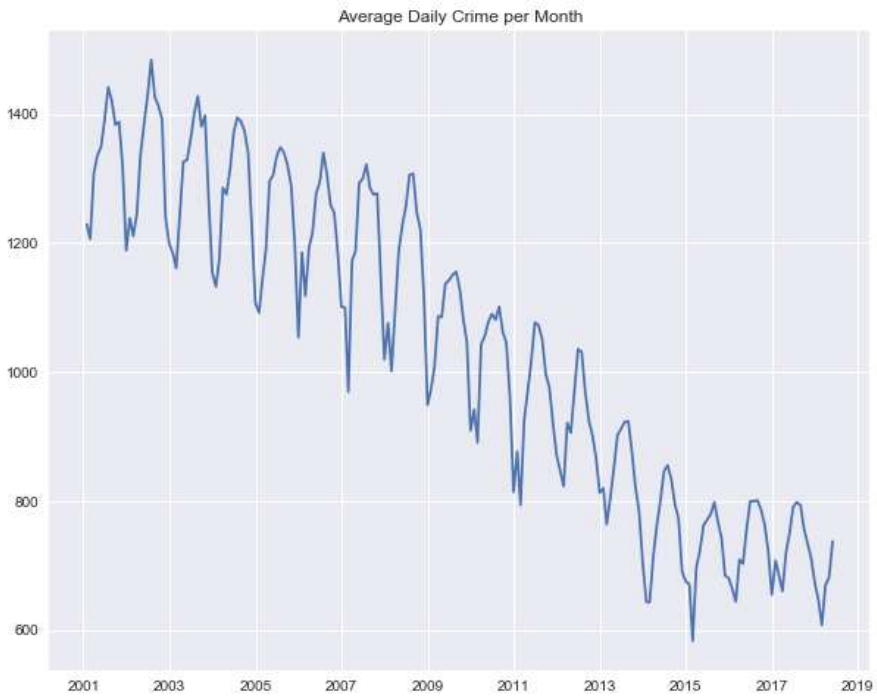

Figure 1: Average daily crime reports per month

From this plot it is clear that crime follows a seasonal pattern, with peaks during the warmer months in the middle of the year and valleys during the colder winter months. Crime has also been trending downward. For the purposes of modeling, only data taken from 2012 onward is investigated, so the model better reflects the more recent data with lower crime rates. The decomposition of the seasonal and trend components of crime are visualized in the plots in Figure 2:

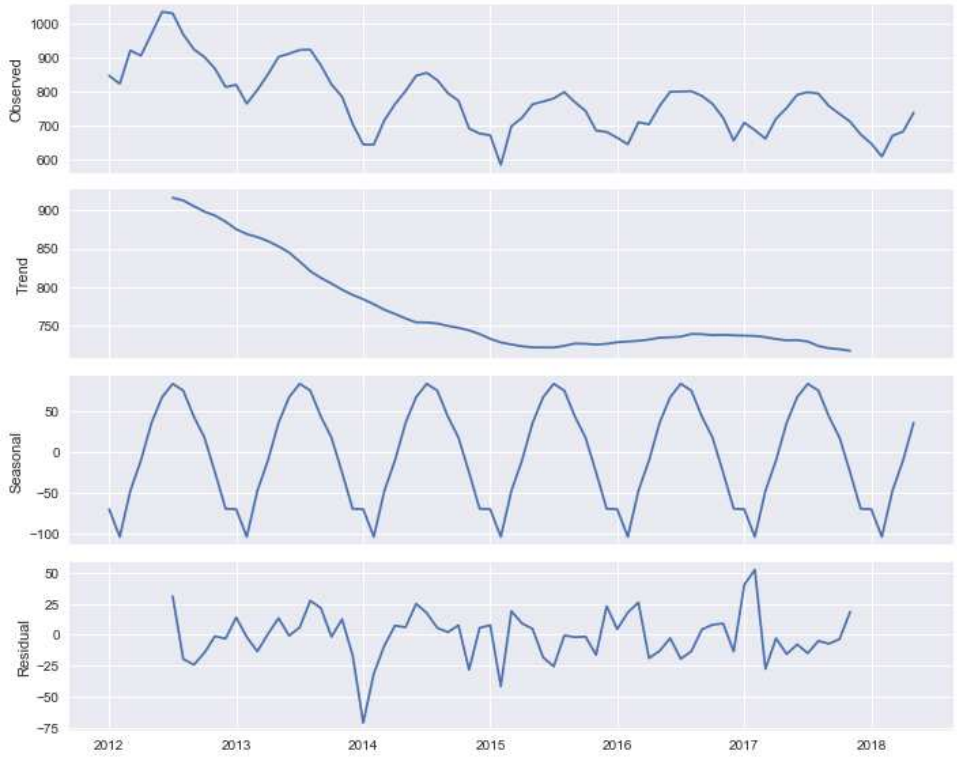

Figure 2: Decomposition of the seasonal and trend components of average daily crime reports per month, from 2012 to present.

Based on these observations, candidate models for predicting crime rates were drawn from the seasonal ARIMA (autoregressive integrated moving average) class of models. A seasonal ARIMA $(p, d, q) \times (P, D, Q)_S$ model for a time series X_t is defined as:

$$\left(1 - \sum_{i=1}^{p} \phi_i L_1^i\right)\left(1 - \sum_{i=1}^{p} \Phi_i L_S^i\right)(1 - L_1)^d (1 - L_S)^D X_t = \left(1 + \sum_{i=1}^{q} \theta_i L_1^i\right)\left(1 + \sum_{i=1}^{q} \Theta_i L_S^i\right)\varepsilon_t \quad (1)$$

where ε_t are the residuals at time t and the lag operator L_j is defined such that:

$$L_j X_t = X_{t-j} \quad (2)$$

As the data is sampled monthly and the seasonal pattern repeats yearly, the value of S was selected to be 12. Investigating all models fitted to data from 2012 with parameter such that $\max(p, d, q, P, D, Q) < 2$ and selecting the one with the lowest value of the Akaike information criterion (AIC) revealed that an ARIMA $(0,1,1) \times (0,1,1)_{12}$ model was best. The AIC is defined such that

$$\text{AIC} = 2k - 2\ln\hat{\ell} \qquad (3)$$

where k is the number of parameters and $\hat{\ell}$ is the maximum of the likelihood function.

An ARIMA $(0,1,1) \times (0,1,1)_{12}$ model was fitted to the average daily crime report data from January 2012 through December 2017, leaving the 2018 data for model validation. The model was found to have coefficients θ_1 = -0.5007 and Θ_1 = -0.5253. Diagnostic plots for this model are illustrated in Figure 3.

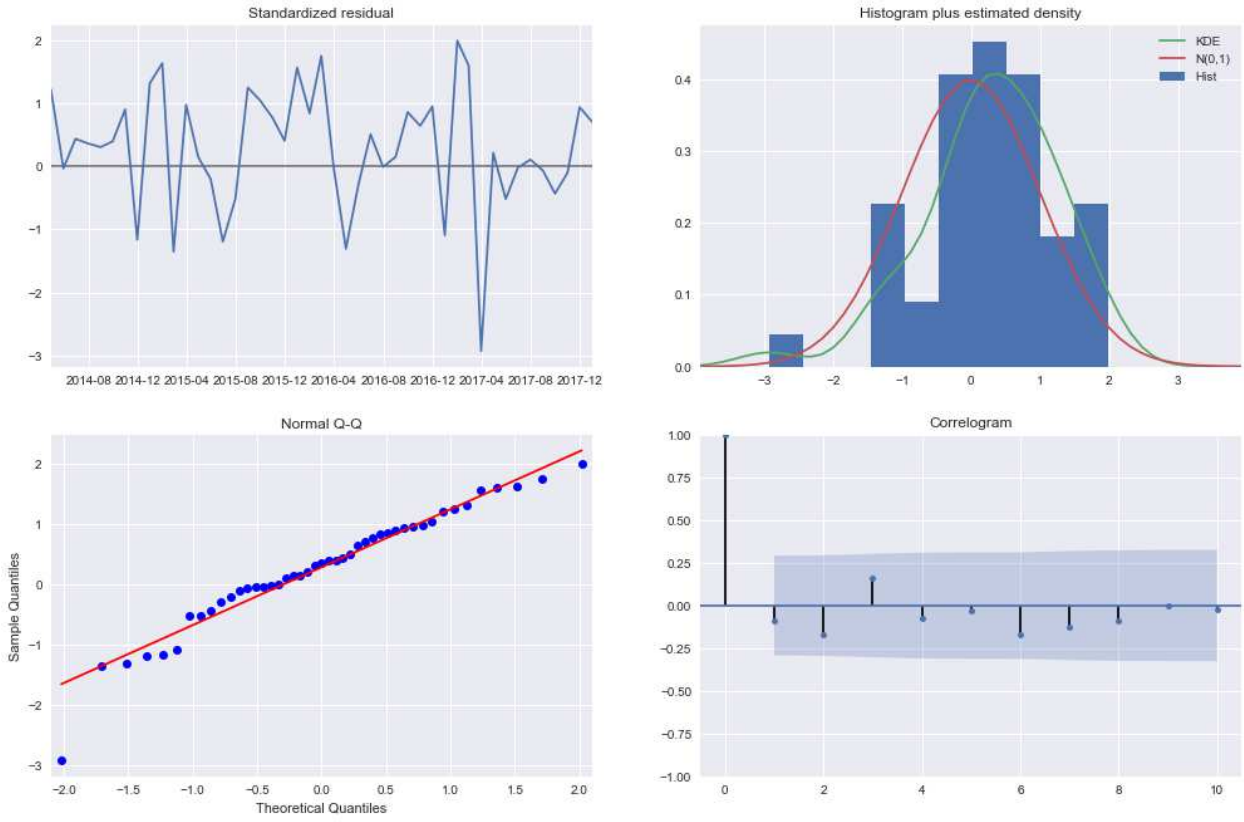

Figure 3: Diagnostic plots for an ARIMA $(0,1,1) \times (0,1,1)_{12}$ fitted to the data from January 2012 to December 2017.

Figure 4 shows the fitted results of the model and the forecasted 2018 results overlaying the actual average daily crime reports per month from 2012 onwards. The sum of squared prediction errors for the 2018 data is 3909.27.

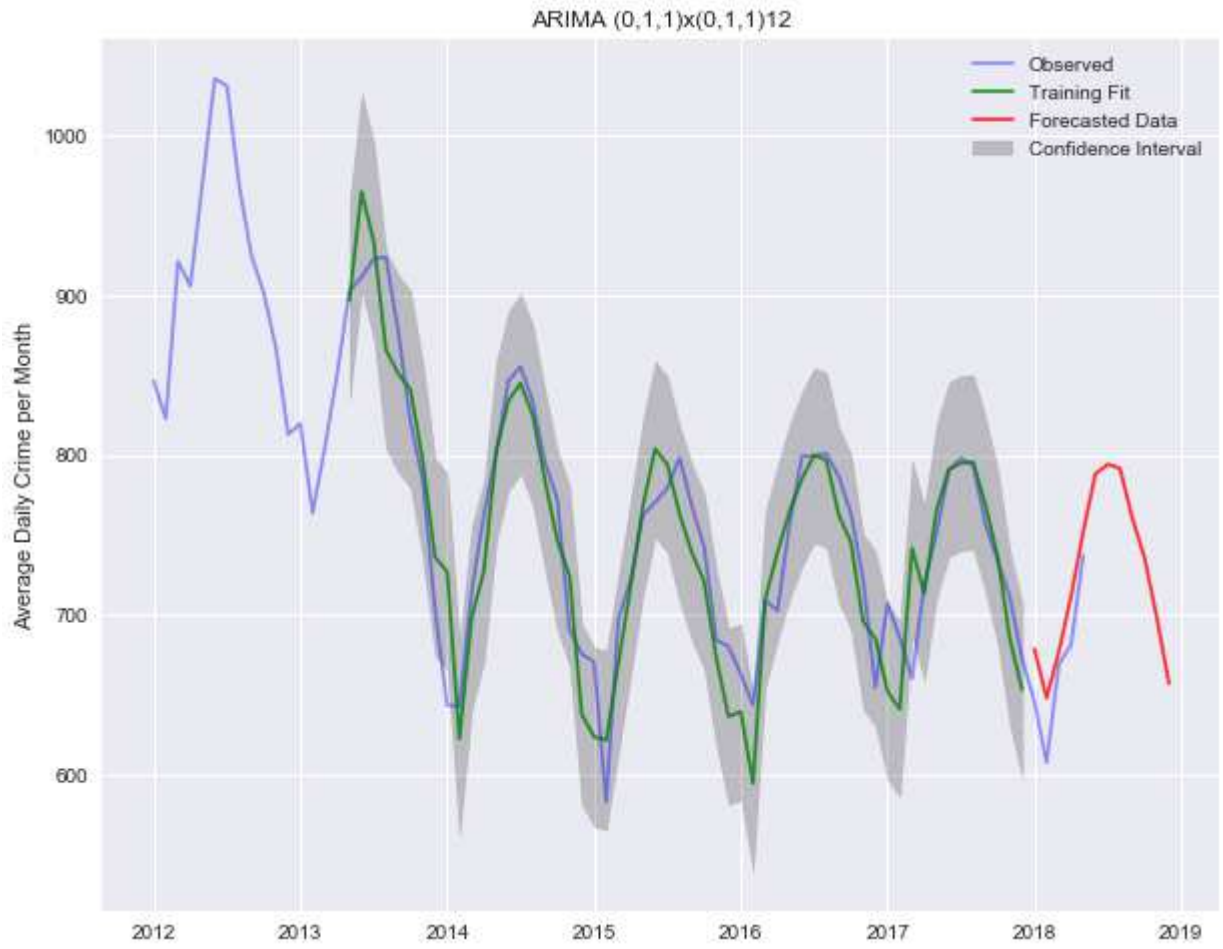

Figure 4: Fitted, forecasted, and observed data from 2012 onward.

In conclusion, seasonal ARIMA models perform well for cases of crime report data. These kinds of models can help predict times when additional police involvement is needed. If location and crime category data is included in the models, the particular regions within a city needing said involvement can be identified, and the nature of the sort of police engagement needed to reduce crime can be identified.