Time series analysis

Crime rates and neighborhood demographics

Jason Chang, Jae Hyun, Zhouyang Lian, Sohil Shah, Melanie Tosik

Business understanding

Task

- Predict crime types and locations from spatiotemporal data
- Identify correlations between demographics indicator variables and crime frequencies
- Time series analysis

Use cases

- Predict changes in neighborhood “health” based on demographic trends
- Prevent leading causes of mortality, morbidity and social problems among youth in the city
- Improve police dispatch efficiency
- Improve turn-by-turn directions based on type and timing of criminal activity

Hypotheses

- Demographic factors are general indicators of a neighborhood’s potential susceptibility to crime

Data understanding

NYPD Complaint Data Historic

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2016).

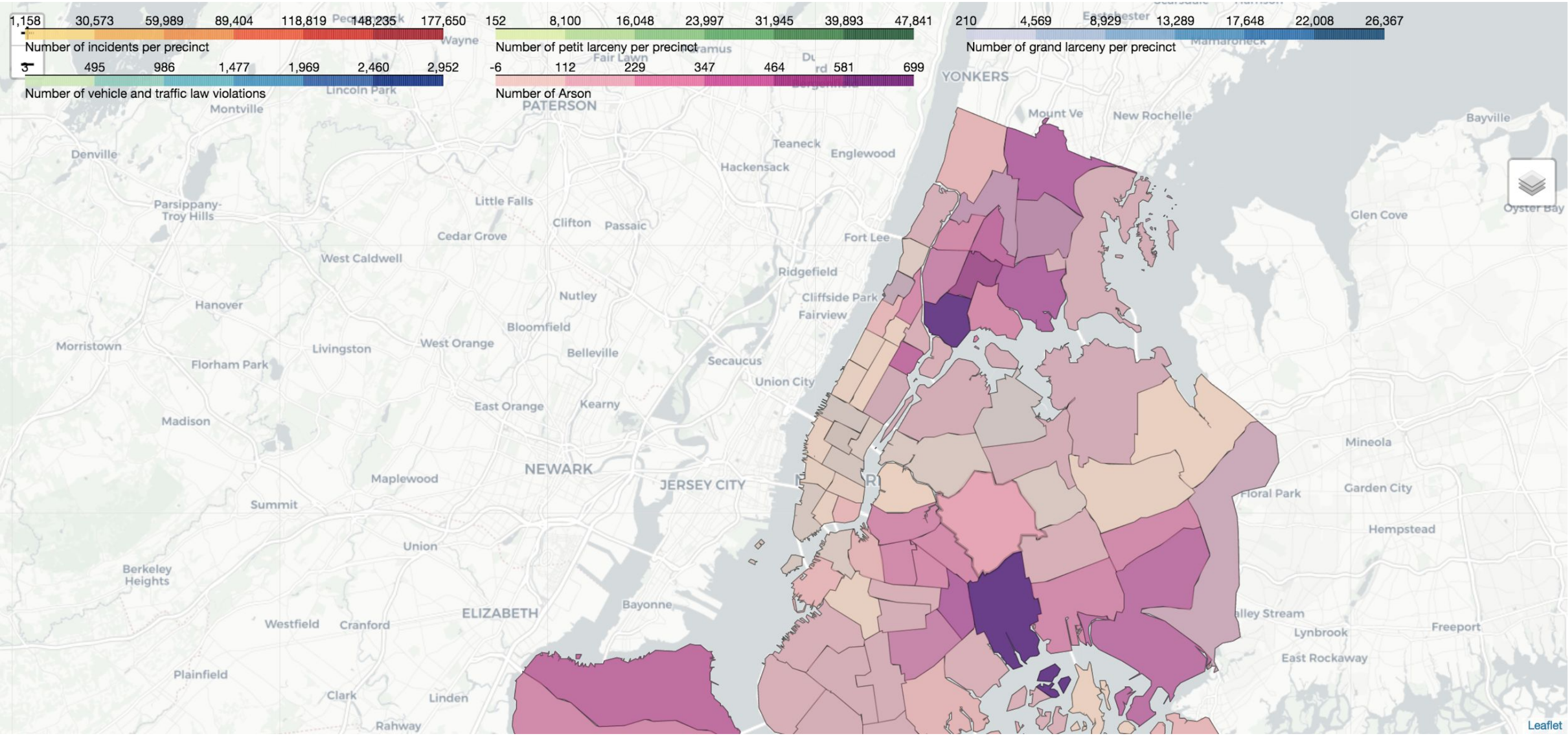
- 2006 to 2016
- 5.58M rows across 24 columns
- Includes date, coordinates, level of offense, description, and responsible jurisdiction

CoreData.nyc

New York City’s housing and neighborhoods data hub, presented by the NYU Furman Center.

- 20 standardized datasets from city, state, and federal sources
- Demographic data, including household composition, income, education, poverty, and race/ethnicity

Data understanding



Visualization based on GeoJSON data of police precincts and Leaflet combined with OpenStreetMap

Data understanding

SIGN UP

CoreData.nyc

New York City's housing and neighborhoods data hub, presented by the NYU Furman Center.

Share Feedback

Search by address, place name, or BBL SEARCH

SUBSIDIZED HOUSING DATABASE 1 | DEMOGRAPHICS | HOUSING MARKET AND CONDITIONS | LAND USE AND DEVELOPMENT | NEIGHBORHOOD SERVICES AND CONDITIONS | RENTERS

Map View

Subsidized Properties

Neighborhood Indicators

INDICATOR

Population aged 25+ with a bachelor's degree or higher

Year: 2016

Region: Sub-Borough Area

4.8% 83.5%

No data

Geographic Boundaries

CoreData Visualizations
Curated by the NYU Furman Center

Download

Sub-Borough Area	2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	201
Astoria	24.6%	29.2%	30.5%	33%	36.8%	35.8%	35.2%	38.6%	41.5%	42.1%	43.5%	46.2%	45.
Bay Ridge	30.6%	35.5%	35.5%	31.2%	34.3%	38.3%	32.7%	36.6%	36.3%	39.6%	38.4%	42.2%	40.
Bayside/Little Neck	37.4%	43.2%	40%	36.4%	41.3%	41.7%	44.8%	45.8%	43.5%	44.3%	43.5%	42.5%	41
Bedford Stuyvesant	10.6%	16%	16.9%	20.2%	21%	20.3%	23%	20.6%	25.2%	26.7%	28.6%	29.1%	36.
Bensonhurst	20.5%	22.9%	23%	22%	17.7%	20.1%	18.6%	23.3%	27.6%	26.6%	30.1%	29.1%	30.
Borough Park	20.8%	22.2%	25.2%	24.3%	26%	25.6%	22.7%	23.5%	23.6%	28.5%	29%	26.9%	28.
Brooklyn Heights/Fort Greene	42.6%	49.7%	54.5%	52.8%	55.3%	58.2%	54.6%	58.2%	58.6%	59.2%	58.9%	62.1%	62.
Brownsville/Ocean Hill	7.6%	9.8%	8.6%	10.8%	11.6%	15.5%	10.2%	9.1%	11.6%	12.5%	12.2%	14.6%	16.
Bushwick	6.9%	11.4%	14%	13.5%	13.3%	14.5%	16.6%	17.5%	19.1%	20.7%	23.6%	26.9%	29.
Central Harlem	14.8%	23%	25.6%	30.8%	30.6%	36.7%	34.1%	30.8%	33.5%	33.4%	37.8%	38%	37
Chelsea/Clinton/Midtown	60.6%	68.6%	68.1%	69%	66.1%	68.7%	69.1%	71.6%	73%	75.4%	71.8%	76.1%	78.
Coney Island	23.5%	31.7%	33.4%	32.6%	27.2%	27%	27.6%	34.6%	32.2%	34.3%	32.1%	37.1%	40.
East Flatbush	15.1%	20.8%	18.6%	17.4%	18.4%	19.6%	17.6%	20.3%	19%	22.4%	21.8%	21.8%	27.
East Harlem	14.2%	25.7%	22%	21.8%	23.7%	25.7%	27.5%	26.4%	32.1%	34.9%	27.3%	32%	34.

« 1 2 3 » Total Rows: 55

Data preparation

NYPD Complaint Data Historic

- Feature reduction
  - Missing Values Ratio (MVR)
  - Feature Correlation Threshold (FCT)
- Data aggregation
- Data integration/fusion
- Data transformation
- Visualization

CoreData.nyc

- Data normalization
  - Population
  - Crime frequency
- Visualization

Modeling correlations

Pearson correlation coefficient

- Measures linear relationship between two datasets
- Varies between -1 and +1
- -1 or +1 imply exact linear relationship
- 0 implies no correlation
- P-value: probability of uncorrelated system producing data sets with same correlation

Tools

- Python
  - pandas
  - sklearn
  - OpenStreetMap
  - Folium (Leaflet.js)
- R
- Tableau

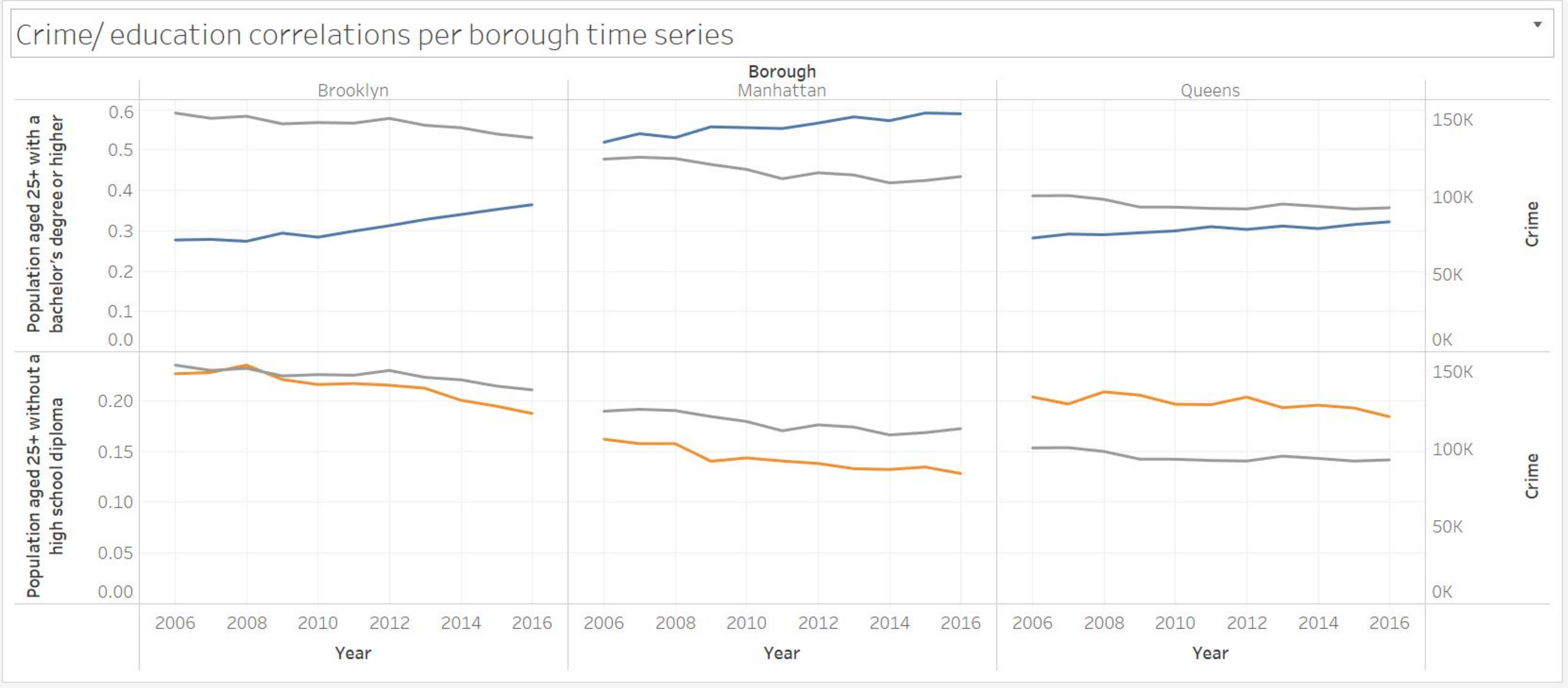
Evaluation

Correlations between crime frequencies and demographic indicators

- Income diversity ratio	[-]*	
- Median household income		
- Poverty rate	[-]	
- Unemployment rate		
- **Population aged 25+ with a bachelor's degree or higher**	[-]	Average: -0.795
- **Population aged 25+ without a high school diploma**	[+]	Average: 0.722
- Population	[-]*	
- Population aged 65+	[-]	
- Foreign born population		
- Racial diversity index		

* 4 of 5 boroughs showed similar trends

Evaluation



Crime/ education correlations per borough time series
Borough
Brooklyn
Manhattan
Queens
Population aged 25+ with a bachelor's degree or higher
Population aged 25+ without a high school diploma
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.20
0.15
0.10
0.05
0.00
150K
100K
50K
0K
Crime
2006
2008
2010
2012
2014
2016
Year

Introduction to time series analysis

Definition

A times series is an ordered sequence of values of a variable at equally spaced time intervals.

Objectives

1. Compact description of data, e.g. classical decomposition: $y_t = T_t + C_t + S_t + I_t$
2. Interpretation, e.g. long term trends or seasonal variation
3. Forecasting, e.g. "Twitter predicts the stock market" (Zeng et al., 2010)
4. Control, e.g. impact of monetary policy on unemployment
5. Hypothesis testing, e.g. global warming
6. Simulation, e.g. estimate probability of catastrophic events

Industry applications

Economic forecasting, sales forecasting, budgetary analysis, stock market analysis, yield projections, process and quality control, inventory studies, workload projections, or census analysis

Time series techniques

Decomposition
Decompose time series into long term trends, seasonal variation, repeated but non-periodic (cyclic) fluctuations, and residuals (irregular components)

Forecasting
Forecast future events based on historic data, based on autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models

Clustering
Partition time series data into groups based on similarity or distance (Euclidean, Manhattan, Hamming, or dynamic time warping (DTW) distance)

Classification
Build a classifier based on labeled time series, use it to predict label of unlabeled time series (e.g. k-NN)

Forecasting and statistical stationarity

Definition

A stationary time series is one whose statistical properties are constant over time.

- Stationarity is often required and can be achieved through mathematical transformations
- Stationarized series are easy to predict
- Predictions can be "untransformed" to obtain predictions for the original series
- Stationarizing through differencing is an important part of fitting an ARIMA model

First difference

The first difference of a time series is the series of changes from one period to the next.

Example:

If Y_t denotes the value of time series Y at period t, first difference of Y at period t is equal to $Y_t - Y_{t-1}$.

ARIMA models

Definition

General models for forecasting a time series that can be stationarized through differencing.

- Can be viewed as a “filter” that tries to separate the signal from the noise
- Signal is then extrapolated into the future to obtain forecasts

Equation

Linear equation in which the predictors consist of lags of the dependent variable and/or forecast errors.

- Predicted value of *Y* = constant and/or weighted sum of one or more recent values of *Y* and/or weighted sum of one or more recent values of the errors
- If predictors consist only of lagged values of *Y*, it is a pure autoregressive model (i.e. a special case of a regression model)

ARIMA models

Acronym

ARIMA stands for **A**uto**R**egressive **I**ntegrated **M**oving **A**verage.

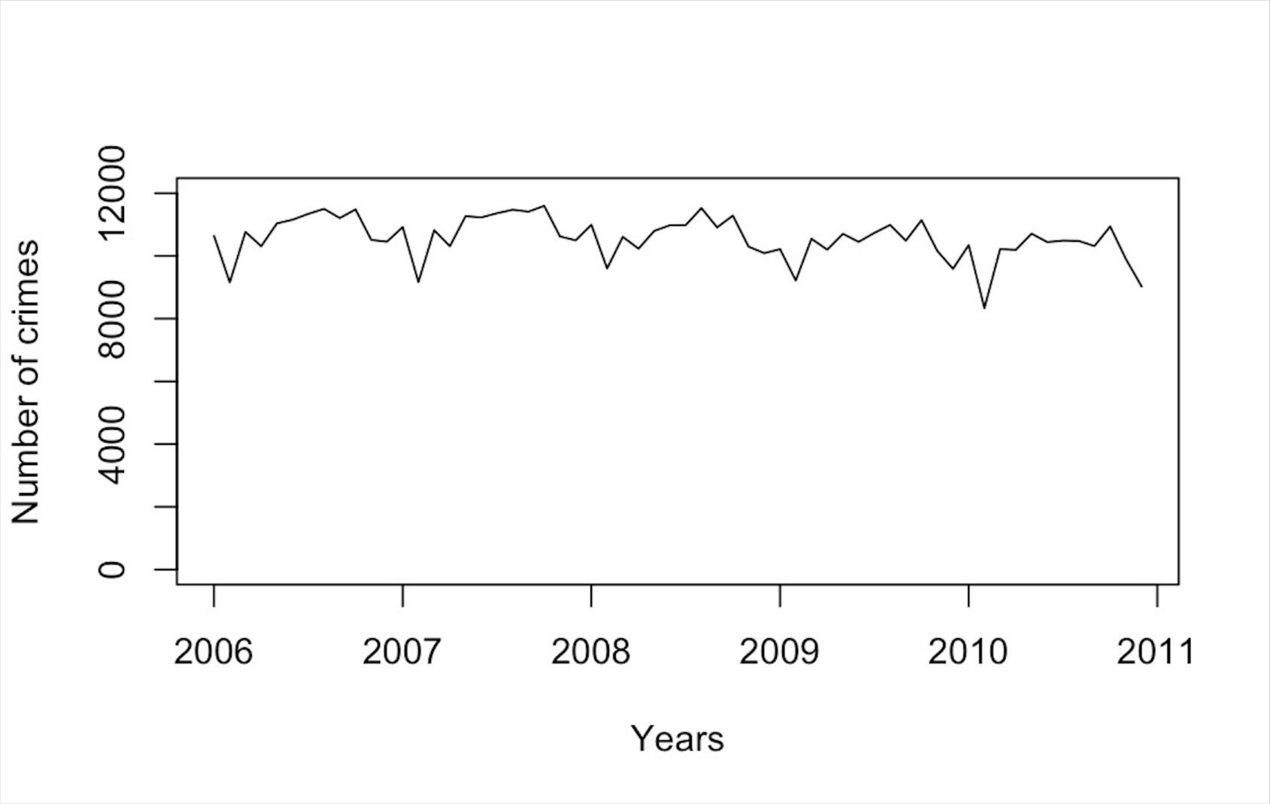
- Lags of the stationarized series in the forecasting equation are called "autoregressive" terms
- Lags of the forecast errors are called "moving average" terms
- Time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series

Equation

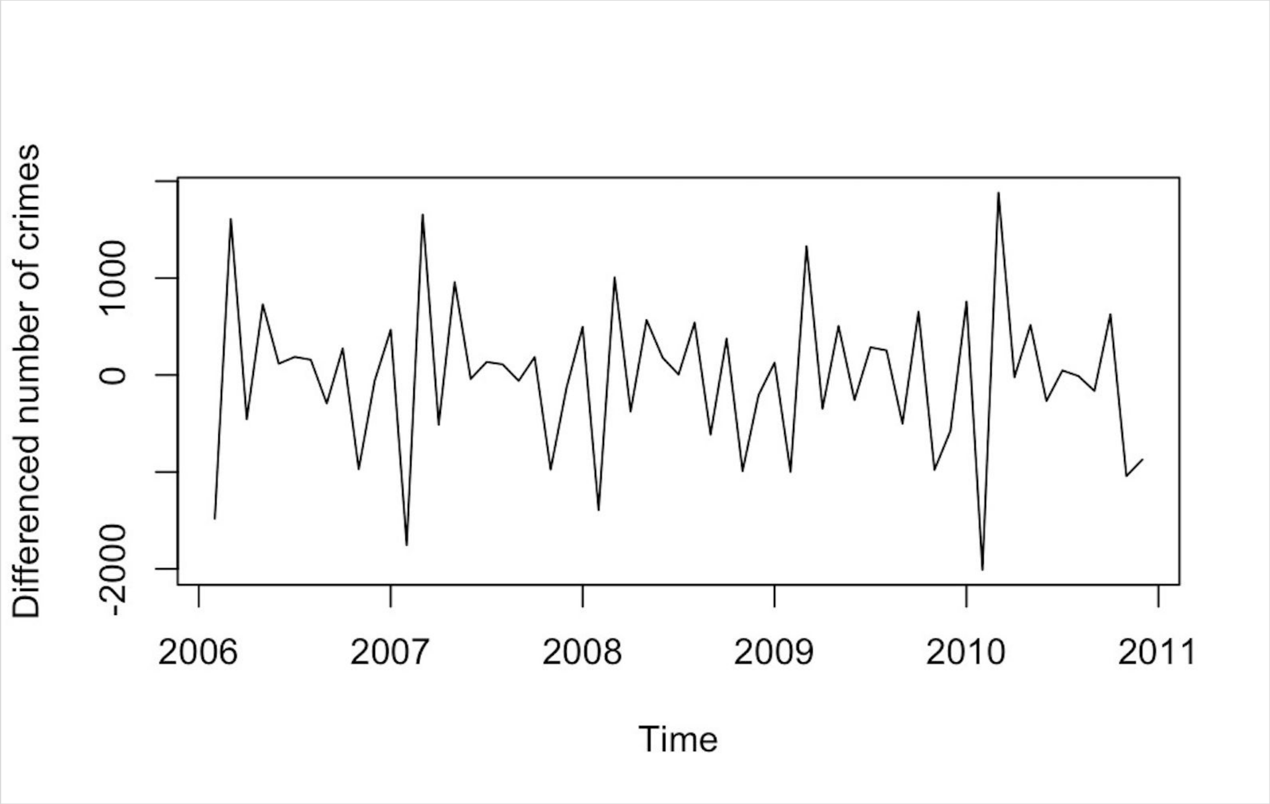
A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- p is the number of autoregressive terms,
- d is the number of nonseasonal differences needed for stationarity, and
- q is the number of lagged forecast errors in the prediction equation.

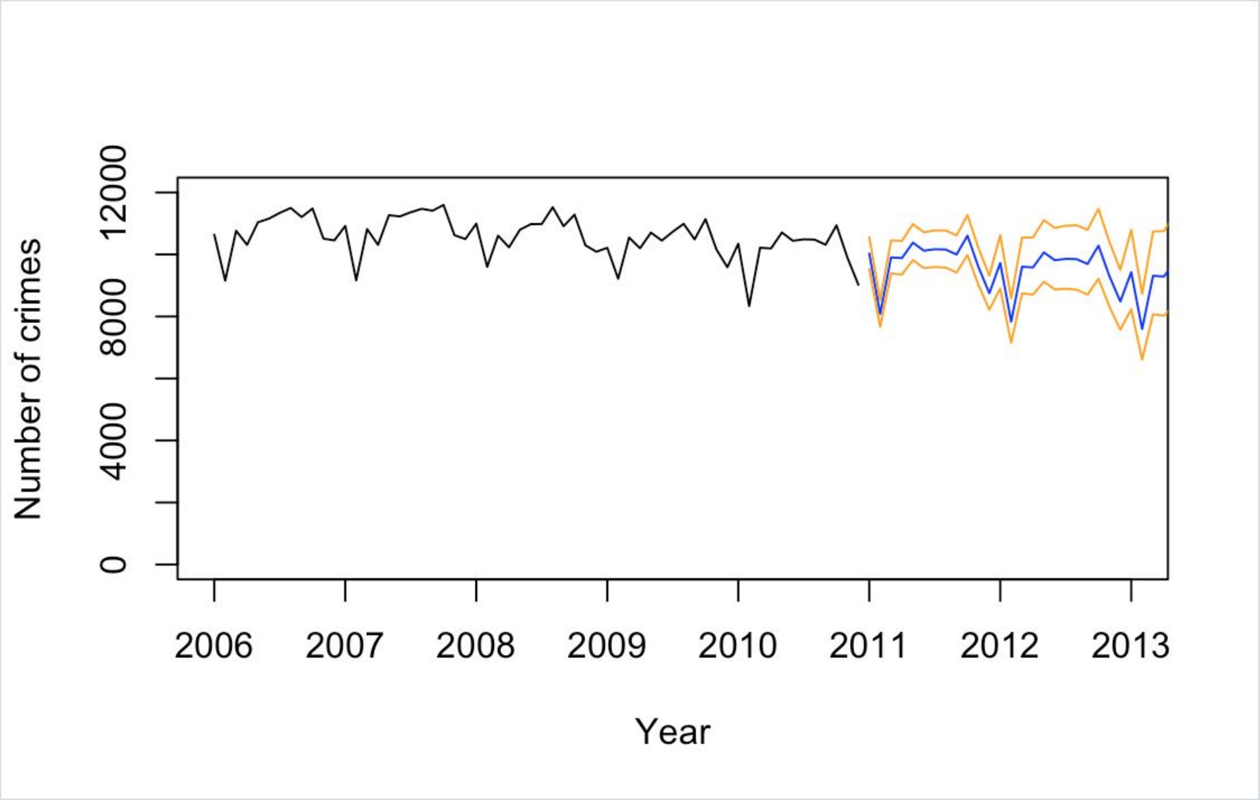
Crime time series analysis [original]



Crime time series analysis [smoothing and difference]



Crime time series analysis [forecasting]



Practical considerations

Missing values

- ARIMA models and smoothing cannot be applied to time series with missing values
- Linear/logistic regression and neural network models do not require imputation

Unequally spaced series

- Many time series are naturally discrete (e.g. bus time arrivals, concerts, bid timings)
- Some forecasting methods might require interpolation

Extreme values

- Unusually large or small values in the series can affect forecasting
- Determine source of outliers (e.g. data entry errors, unusual events) and remove accordingly

Choice of time span

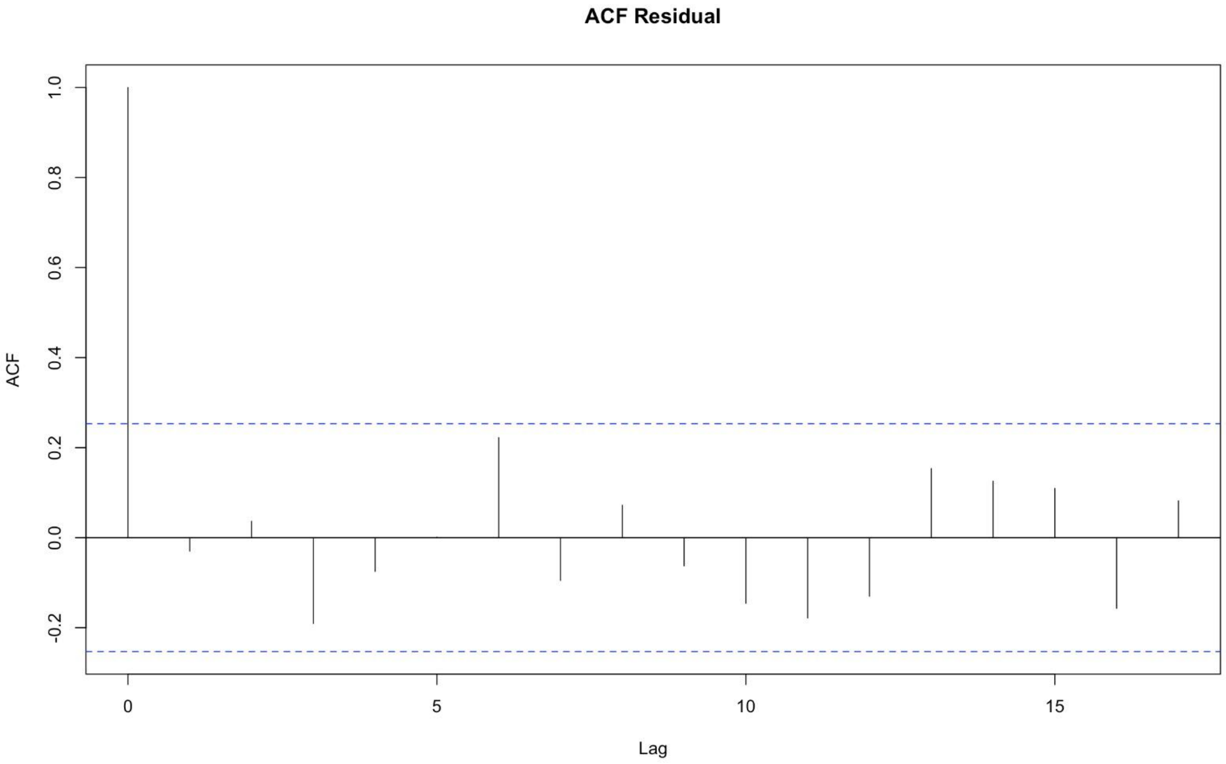
- Long past of the series might deteriorate forecasting accuracy (changing context/environment)

Future work

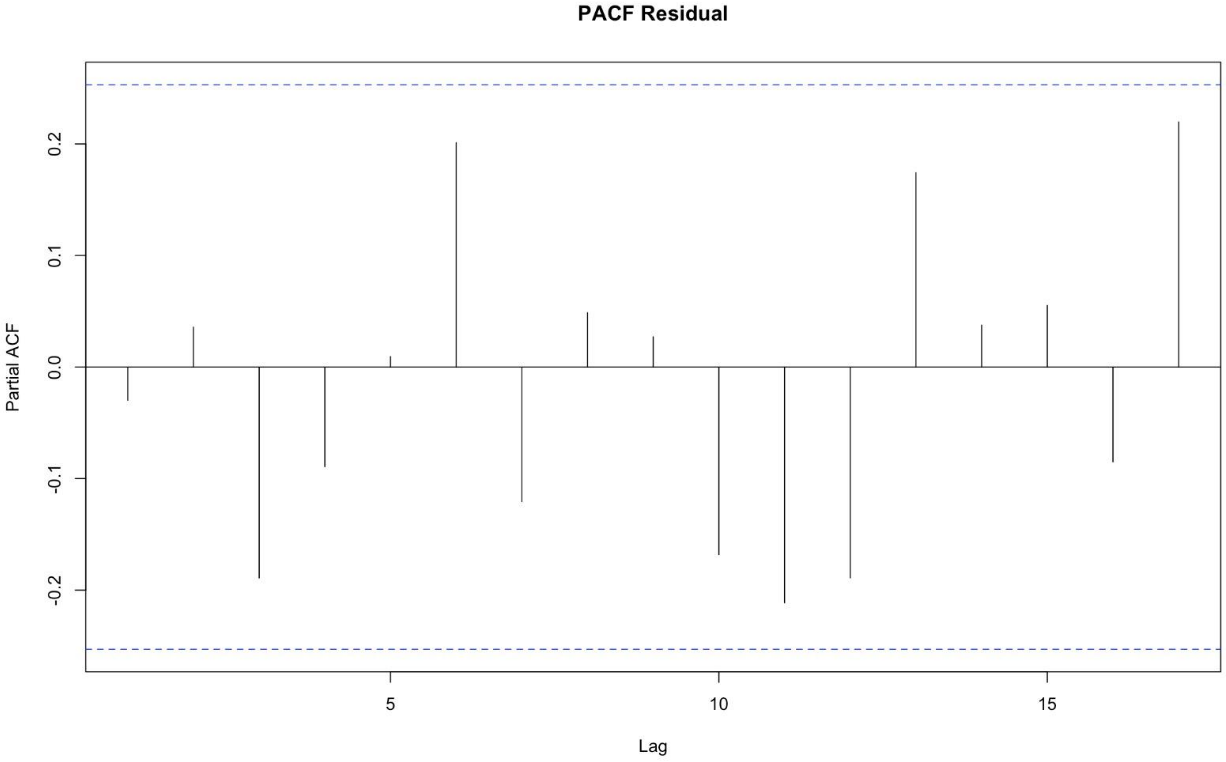
- Verify fit of ARIMA model by examining residuals to ensure no more information is left for extraction (residuals should be random with no visible trends)
- Forecast heat maps for crime frequency at fine-grained spatial and temporal resolution
- Include other relevant data sources as exogenous variables to have correlated time series
- Classify crime types on time series, e.g. using long short-term memory (LSTM) with fuzzy logic
- Cluster time series of various different crime types based on individual crime type time series

Questions?

ARIMA residuals [ACF diff log]



ARIMA residuals [PACF diff log]



Related work

J. R. Hipp, and C. E. Kubrin, “**Crime Report for Southern California**,” UCI Irvine Laboratory for the Study of Space and Crime, 2015.

- Analyze the level of crime for all cities in the Southern California region with the population size of at least 4000
- Presents top/bottom cities for a specific crime type, and the increases/decreases of crime type
- Adjust the rates for the different socio-demographic characteristics of the city

H. Wang, D. Kifer, C. Graif, and Z. Li, “**Crime Rate Inference with Big Data**,” Pennsylvania State University.

- Uses large-scale (“Big Data”) Point-Of-Interest data as well as taxi flow data for Chicago
- Observed consistently improved performance in predicting crime rates for multiple years compared to using just demographic features

H. Kang, and H. Kang, “**Prediction of crime occurrence from multi-modal data using deep learning**,” 2017.

- Uses feature-level data fusion method based on deep neural network (DNN)
    - Train DNN with spatial, temporal, environmental context, and joint feature representation layers
- Dataset consists of: crime statistics, demographic and meteorological data, and images of Chicago
- Improved accuracy in predicting crime