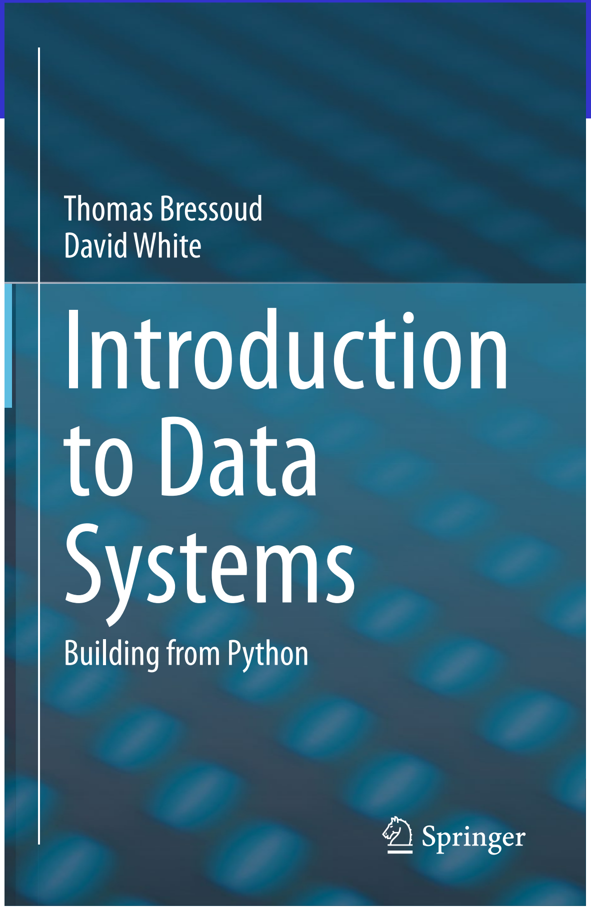
Mathematical and statistical models for applied time series analysis

David White
Denison University

A bit about me

- Bowdoin --> Wesleyan --> Denison
- Trained in pure math: algebraic topology
- Worked briefly outside academia building systems to analyze big data
- Teach math, computer science, and statistics
- Co-authored a book. Basically “all the computer science you need to know to do data science”
- Wrote a few papers on data science pedagogy and a few applied statistical analyses (often with students)
- Now frequently asked to consult on time series analyses
- Strong personal interest in social justice.



Plan for today

- What is a time series data set and how can it be analyzed?
  - Model it as a function of time (detrend)
  - Fourier Analysis (find seasonal patterns)
  - ARIMA models (impact of the past on the present)
  - Regression of one time series on another.
- Examples of applied time series analyses from my research
  - Police seizures of drugs predict for overdose deaths
  - Police behavior at protests predicts for number and violence of protests

Key Take-Aways

- Time series models are not that hard (for math majors), but social scientists often need help with them. Liberal arts training is critical.
- There are tons of freely available datasets that have never been analyzed. Lots of low-hanging fruit, and nowhere near enough math/stats people working in this area.
- Even simplistic analyses are valuable to social scientists and harm reduction professionals, can save lives, and can get published. Great for students.
- Great line of research to justify “broader impacts” in grant proposals.
- Much easier to talk to your friends about than abstract homotopy theory!
- I created a GitHub repository with R Markdown files to carry out dozens of applied time series models on real-world data sets. Happy to share!

Time Series Definition and Examples

A time series is a sequence of numbers Y_t where t is time. Examples:

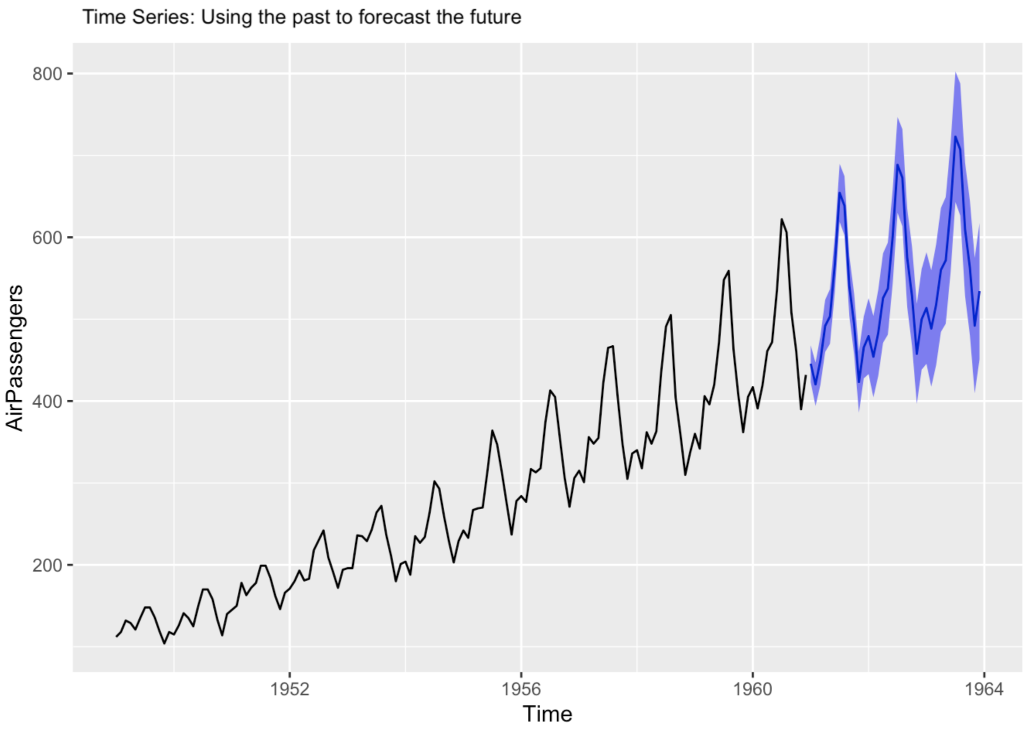
- Price of a gallon of fuel, t measured in days
- Financial data like price of a stock, or inflation index
- Sound waves, e.g., Y_t measured in decibels
- Climate change: amount of $CO2$ in atmosphere
- Traffic: number of cars every minute
- Number of COVID cases/hospitalizations/deaths
- Number of protesters each day
- Number of drug overdose deaths each month

Time series data is everywhere!

Very high-level view of applied statistics

- Given a data set, choose and fit an appropriate model that captures the essential features, is useful, and is not overfit to the data.
- The residuals (what the model misses) should be random and independent
- Use the model for prediction/forecasting
- Do inference, e.g., determine whether explanatory variables really matter, quantify how much they matter for predicting the response variable, etc.
- Try to maximize how much of the variability in the response variable is explained (or minimize residual sum of squares), without overfitting.
- Principle of Parsimony: simple models are better! Think about your final audience and the take-away message.

One application: Forecasting

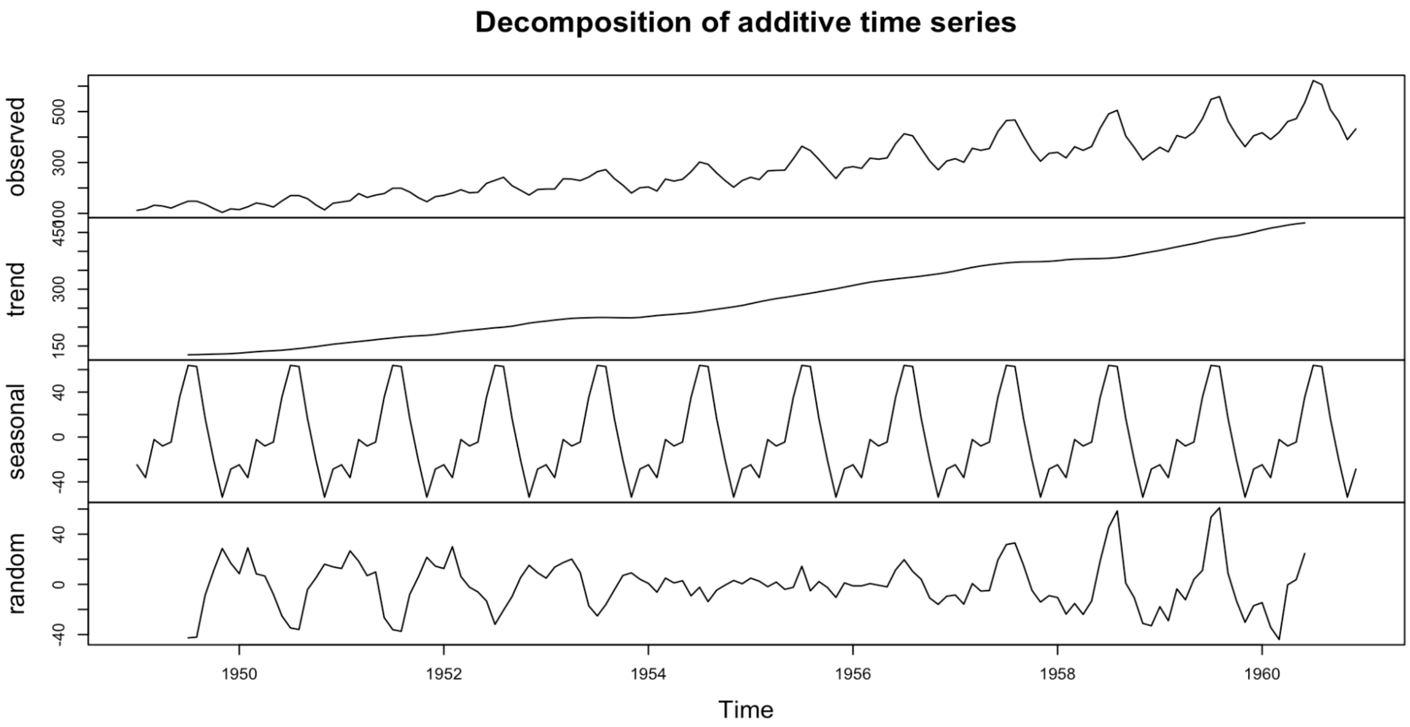


Example: Data = number of airline passengers each month from 1949-1960.

Fit a model that explains the growth and seasonal patterns, with random and independent residuals.

Forecast the next three years, plus 95% confidence interval. Use that to make money!

Fit a model then use it to forecast



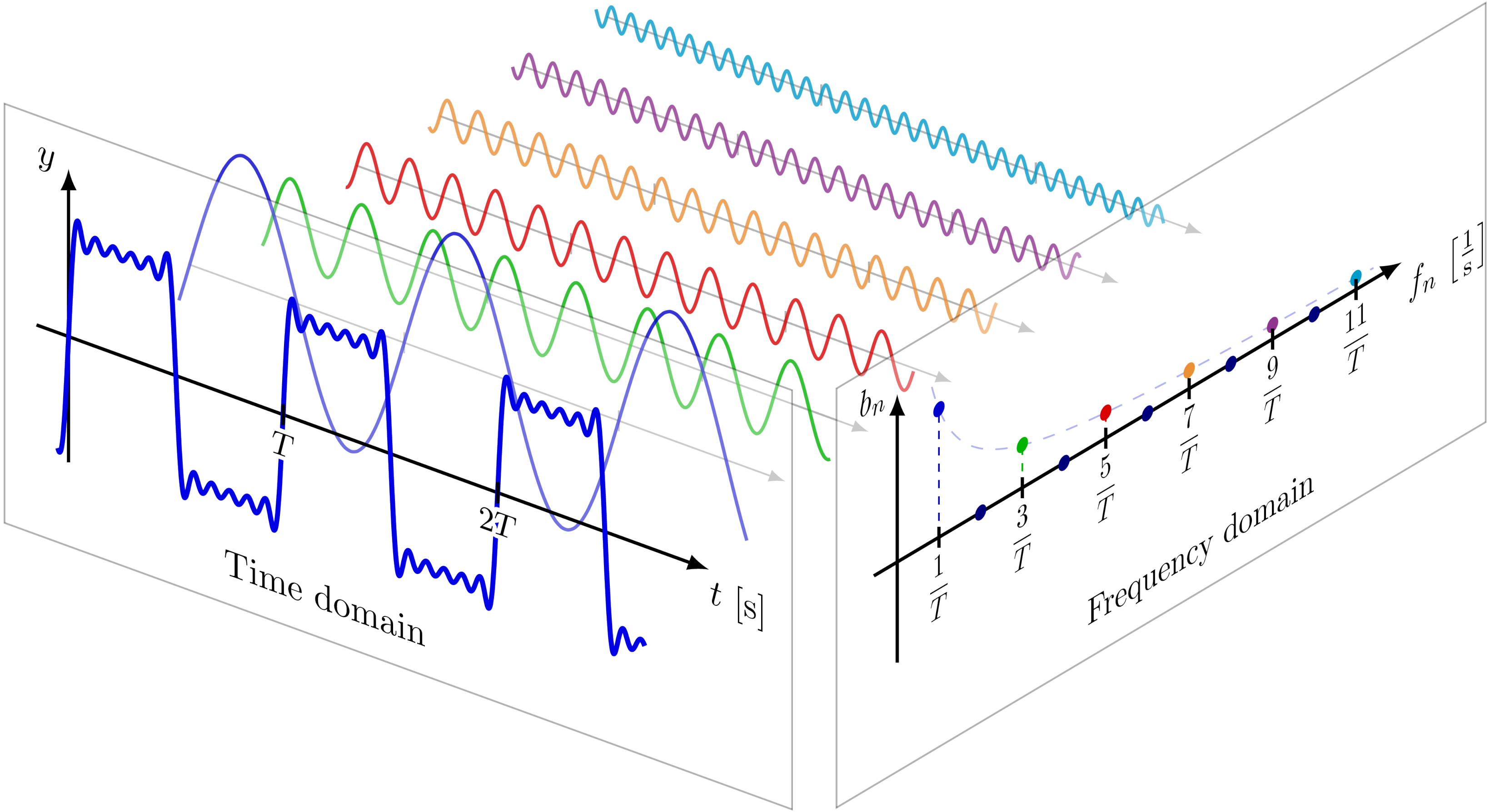
Monthly data.

What math functions can model the seasonal part?

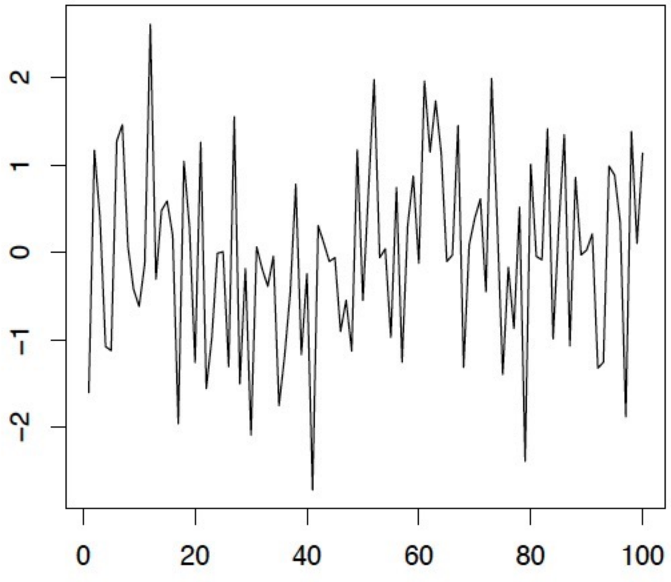
Hint: functions that oscillate.

What if you don't know the period?

Breaking a signal up into sum of its sinusoidal pieces, like Taylor series



Fourier models



Fit a model like:

$$\sum_{j=1}^{m} \left[A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t)\right]$$

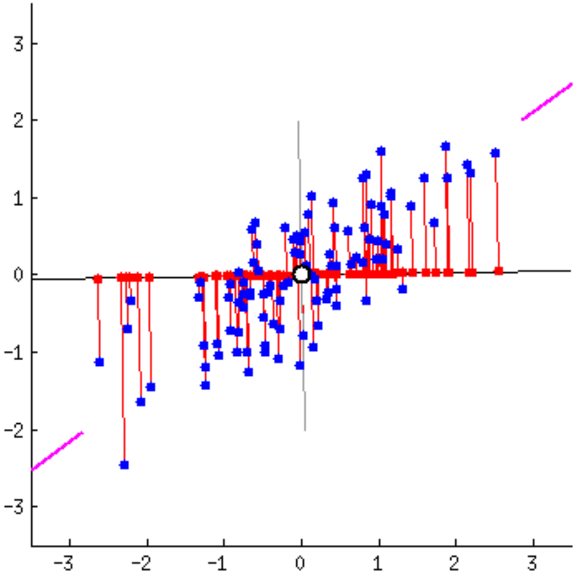
Problem: what's m? What are the periods?

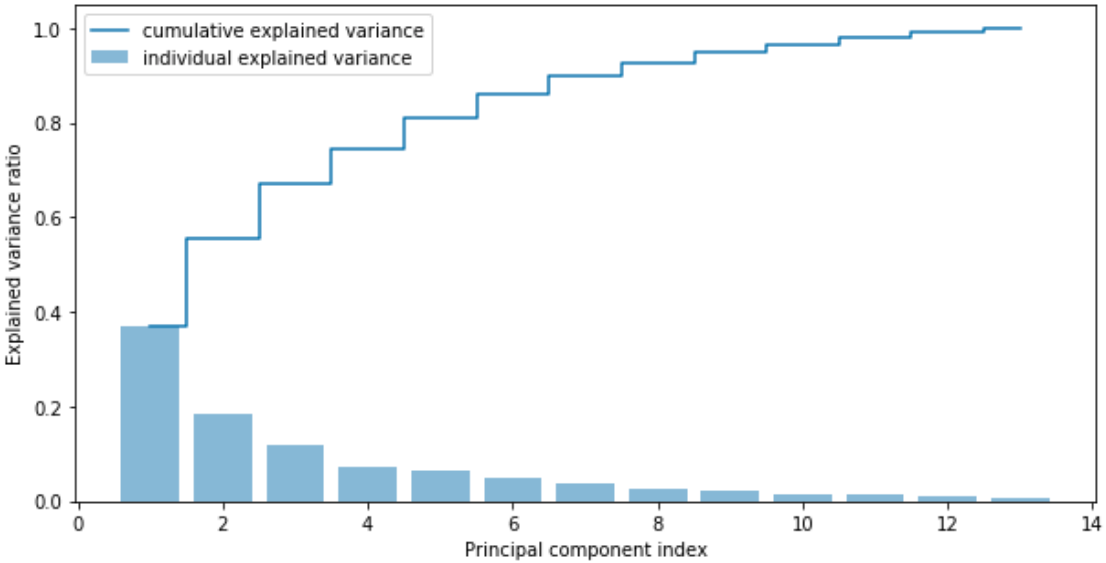
Solution: take the Fourier transform!

Idea: change to a basis where basis vectors correspond to periods, ordered by how much variability they explain. Then just keep the first few.

Same math as Shazam or Sound Hound!

Principal Component Analysis (PCA)





Exact same concept as PCA. Accomplish via Singular Value Decomp.

Another application: Bivariate time series analysis

How do changes in explanatory variable X_t affect response variable Y_t? Examples:

- Using news data (e.g., word count of certain words) to predict the stock market
- Using atmospheric CO2 to predict temperature
- Using police behavior to predict number of protesters
- Using drug market to predict overdose deaths

Answer requires time series regression models.

Kenyon Mission: build strong foundations for lives of purpose and consequence.
Let's use our math/stats skills to make the world a better place!

My own research: data science for social good

Research Question 1 (from 2019):

- Every year, thousands of people die from drug overdoses in Ohio
- Death data is often delayed by six months or more, but police crime lab tests of drug composition of seized drugs is immediate.
- Regress Deaths_t on Seizures_t and build an "early warning system" to warn people when dangerous drugs appear, *before* those drugs kill.

Research Question 2 (from 2021 and 2023):

- Does use of rubber bullets by police suppress or inflame protests?
- Let Y_t = number of protesters and X_t = number of rubber bullets shot
- We find a statistically significant correlation and quantify the impact of X_t
- Use this to lobby to change police practices to negotiated management model

Question 1: opioid epidemic in USA

The yearly number of drug overdose deaths surged in the US from 16,849 cases in 1999 to 107,941 cases in 2022.

In 2022, more than 295 people died every day in the US after overdosing on opioids.

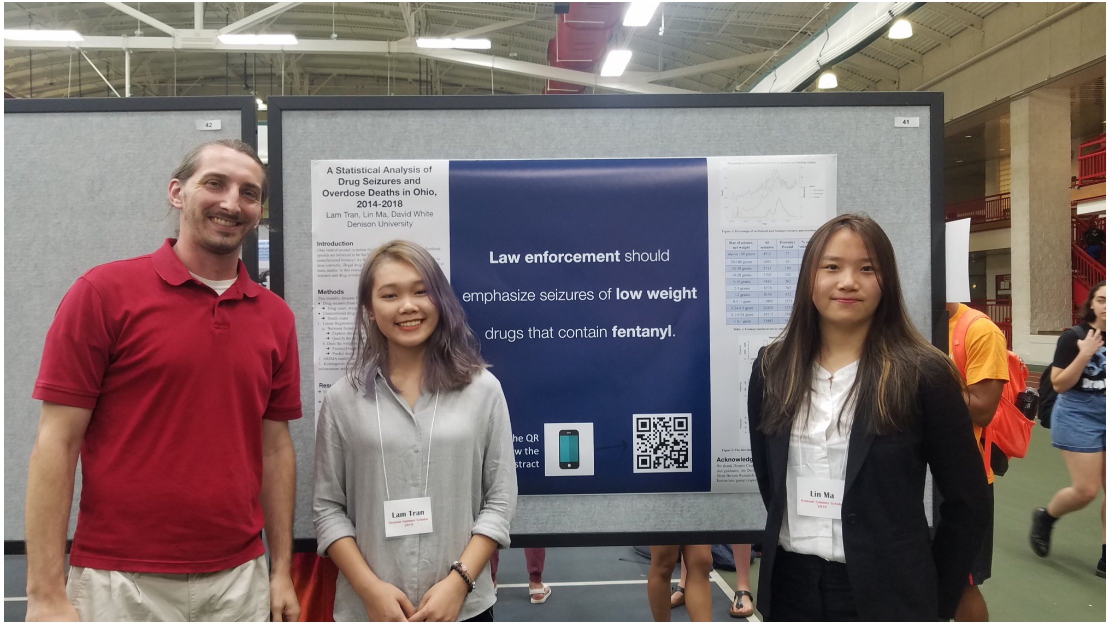
The number of fentanyl encounters has been increasing exponentially, from 5,343 in 2014 to 117,045 in 2020. It has continued to increase.

As of now, an American is more likely to die from an unintentional drug overdose than in a car accident.

Ohio has average drug use patterns, but in 2017 was second in the country for unintentional drug overdose deaths, and now seventh highest. Why?

Iron Law of Prohibition: if you crack down on one type of drugs, dealers will select more potent drugs to traffic in.

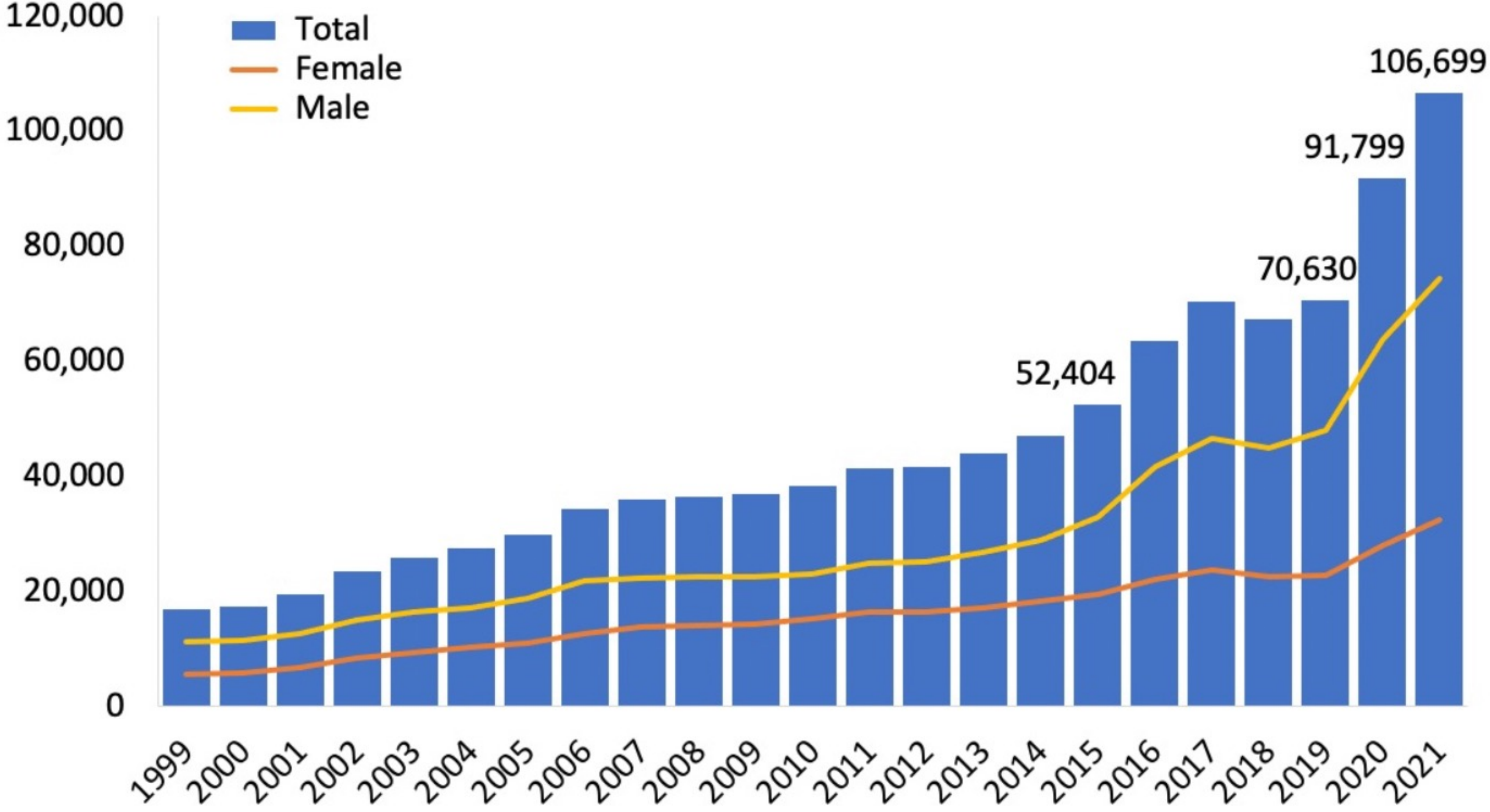
Joint with Denison students Lin Ma (‘20), Lam Tran (‘21)
Inspired by work of Dennis Cauchon (Harm Reduction Ohio)



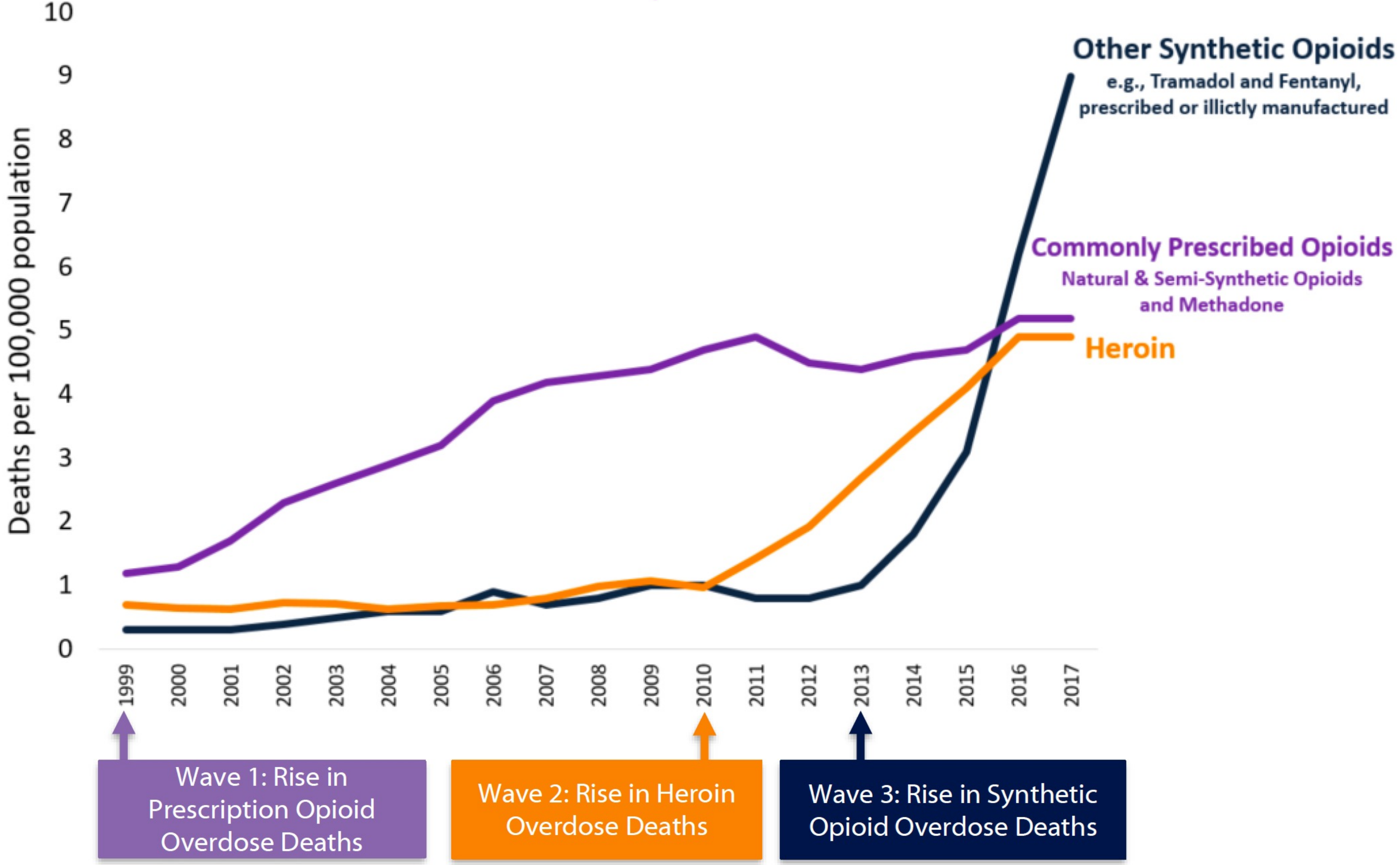
Funded by Summer Scholars Program, Narrative Journalism concentration, and the Andrew W. Mellon Foundation.

Our summer research team:
Data driven journalism

Figure 1. National Drug-Involved Overdose Deaths*, Number Among All Ages, by Gender, 1999-2021



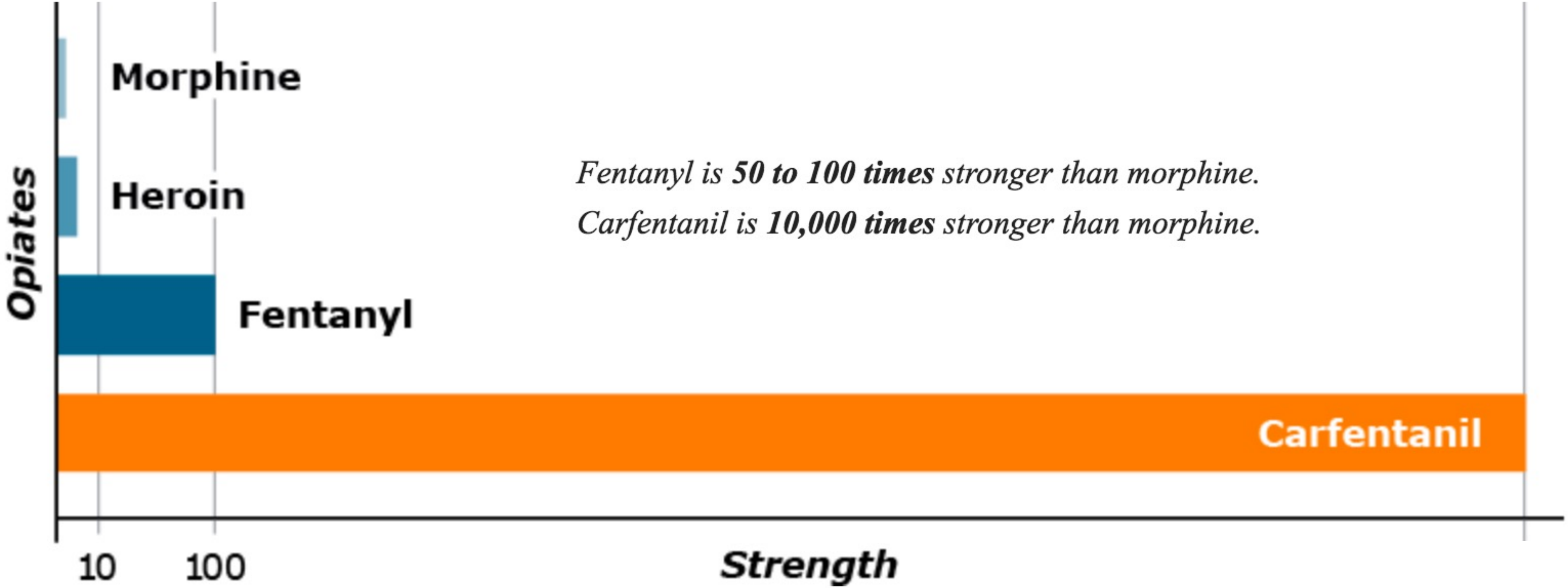
3 Waves of the Rise in Opioid Overdose Deaths



Fentanyl is a synthetic opioid

- Cheap to manufacture.
- Very powerful.
- Easy to mix with other drugs.
- Many variants of unknown strength.
- Strongest known variant is Carfentanil.

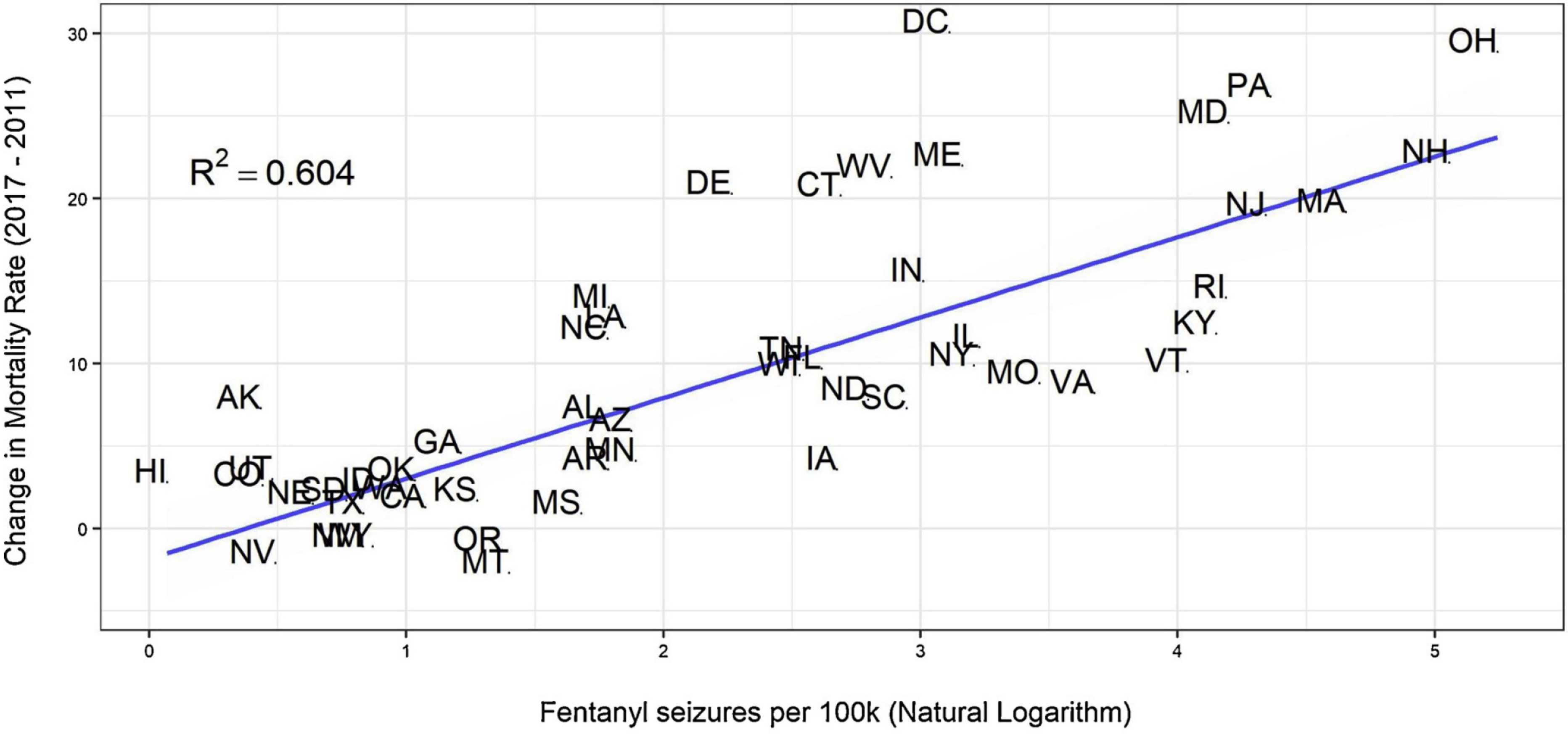
Relative Strength Compared to Morphine



*Fentanyl is **50 to 100 times** stronger than morphine.*

*Carfentanil is **10,000 times** stronger than morphine.*

Fentanyl & Increased Overdose Mortality (2011 vs 2017)



Source: NFLIS/CDC

Analyzing BCI dataset alongside Ohio Mortality data

Reference: Ma, Tran, and White. “A statistical analysis of drug seizures and opioid overdose deaths in Ohio from 2014 to 2018,” *JSR,* vol. 10(1), 2021.

1. Exploratory: Fentanyl seizures and deaths track together over time.
2. Quantify the amount of variability in deaths that it explained by drug seizures and by fentanyl seizures. Drug seizures is a powerful predictor for deaths.
3. Low weight drug seizures are more likely to contain fentanyl than higher weight seizures. The weight variable adds predictive power.
4. Use time-series analysis to quantify lag between seizures and deaths.
5. Fit an ARIMA model for deaths and seizures, then a general linear mixed model.
6. Compare the efficacy of different types of law enforcement, including national law enforcement (FBI/DEA), drug task forces, and local police and sheriff’s departments

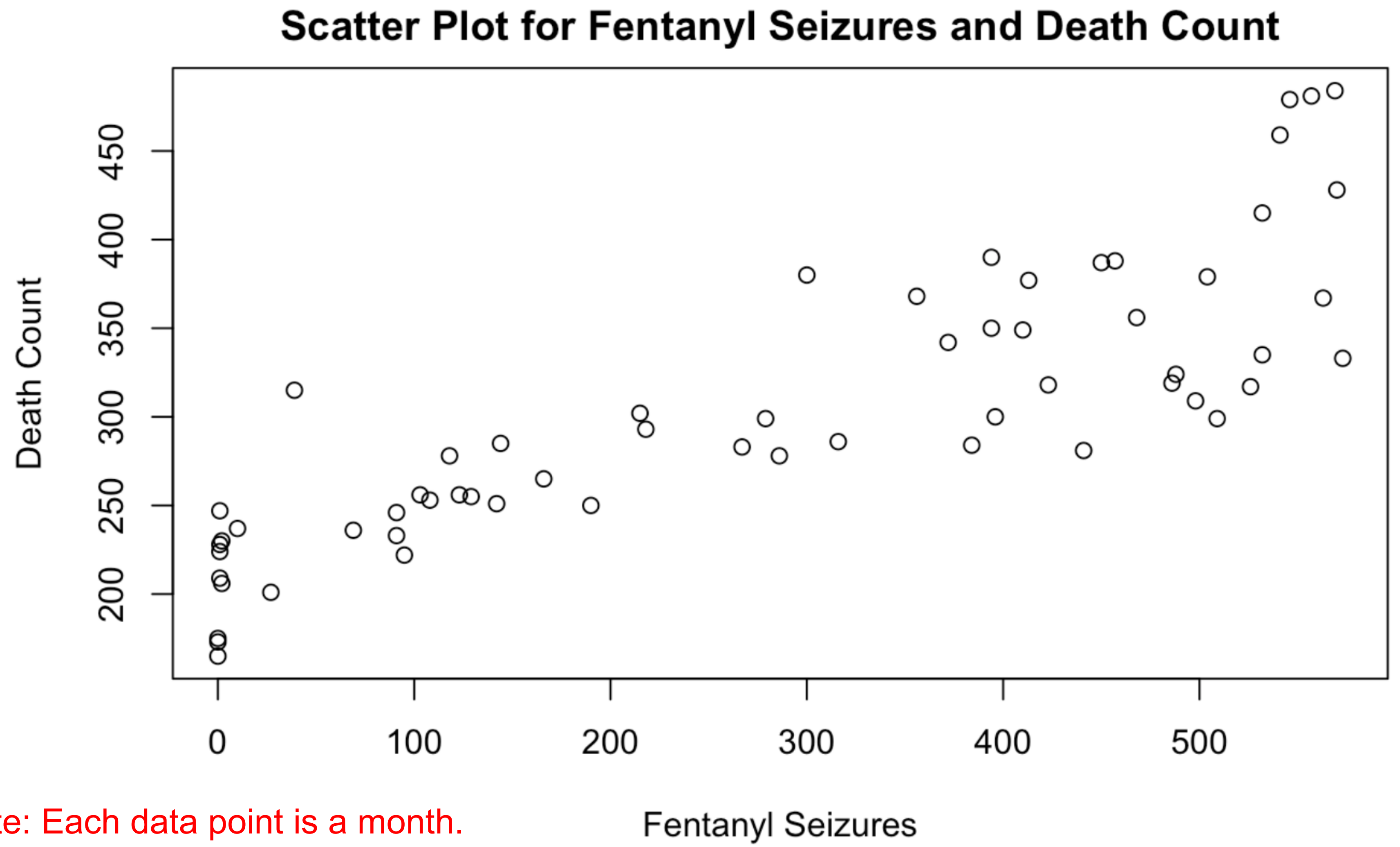
Discussion of the two time series

Ohio Dept of Health: Number of overdose deaths per month.
Bureau of Criminal Investigations (BCI): one row per drug seizure by police, with date, county, list of drugs taken, and weight.

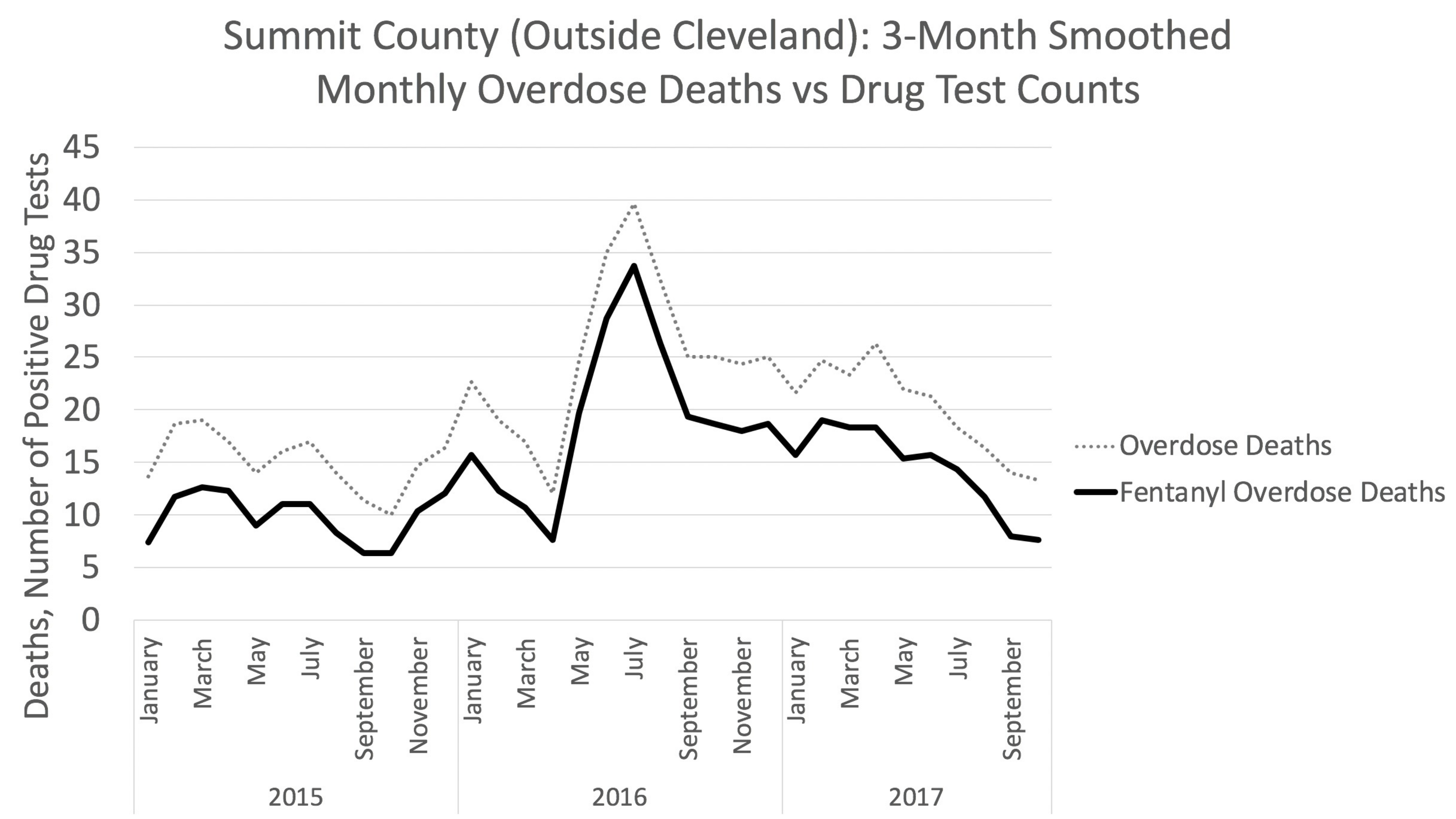
What we did:

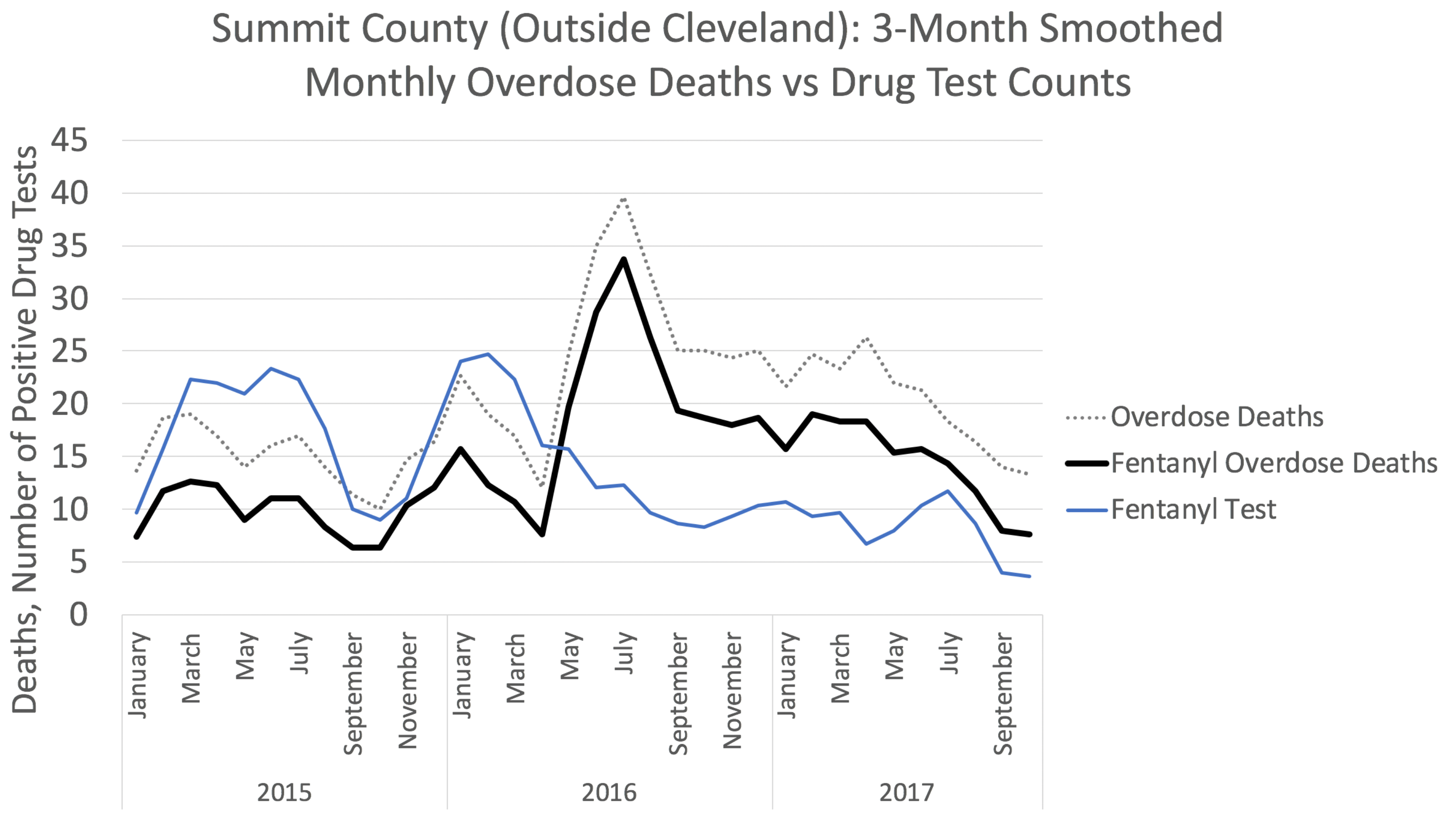
1. Aggregate/wrangle the BCI data to the monthly level.
2. In the BCI data, use text-matching algorithms to identify the seizures that contained fentanyl and other fentanyl variants.
3. Merge the data sets together then regress deaths on seizures.
4. We got an R^2 of 80%, and learned:

One additional positive BCI test of carfentanil predicts 0.45 more deaths.
One additional positive BCI test of fentanyl/fentanyl analog predicts 0.2 more deaths.

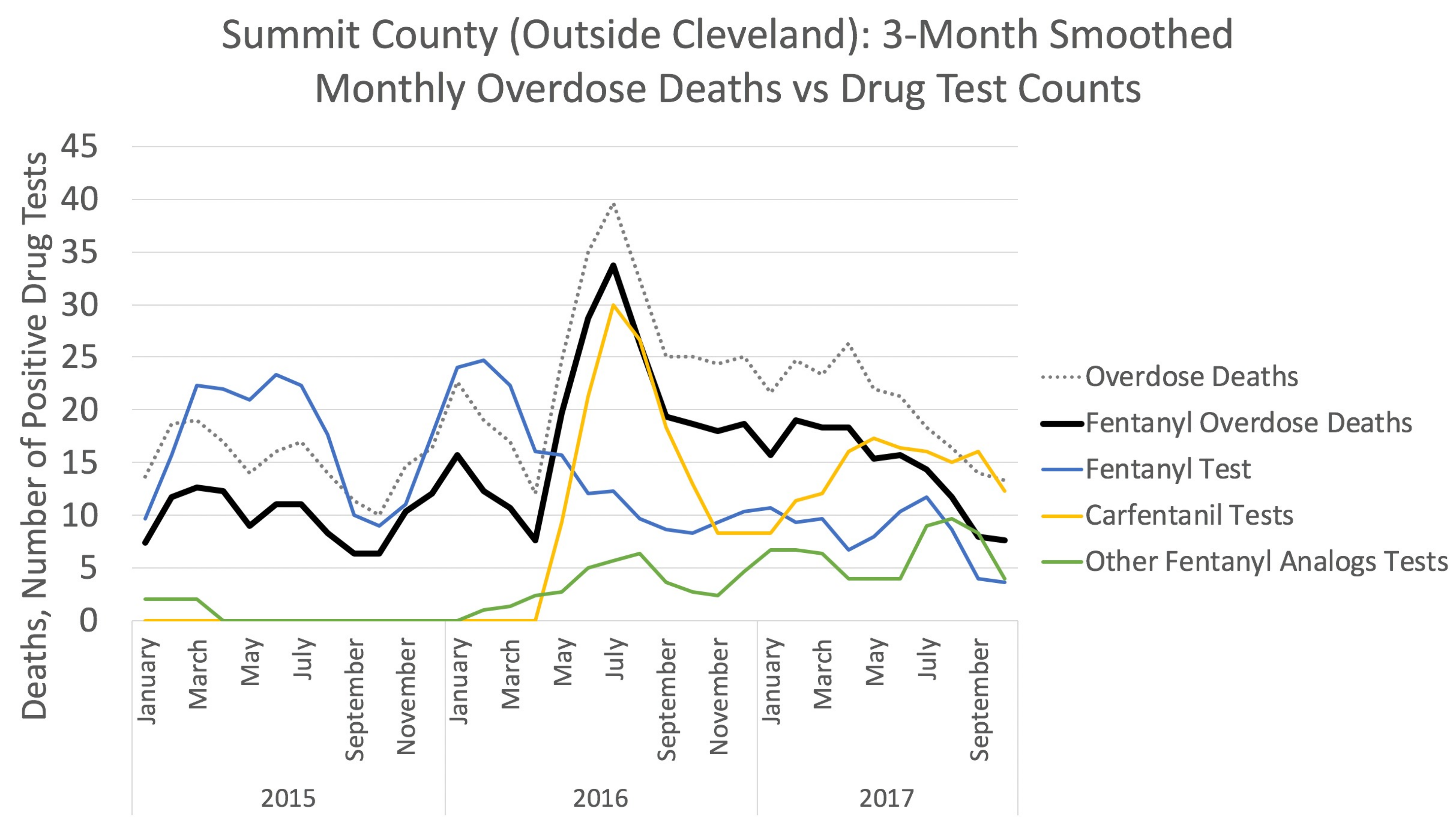


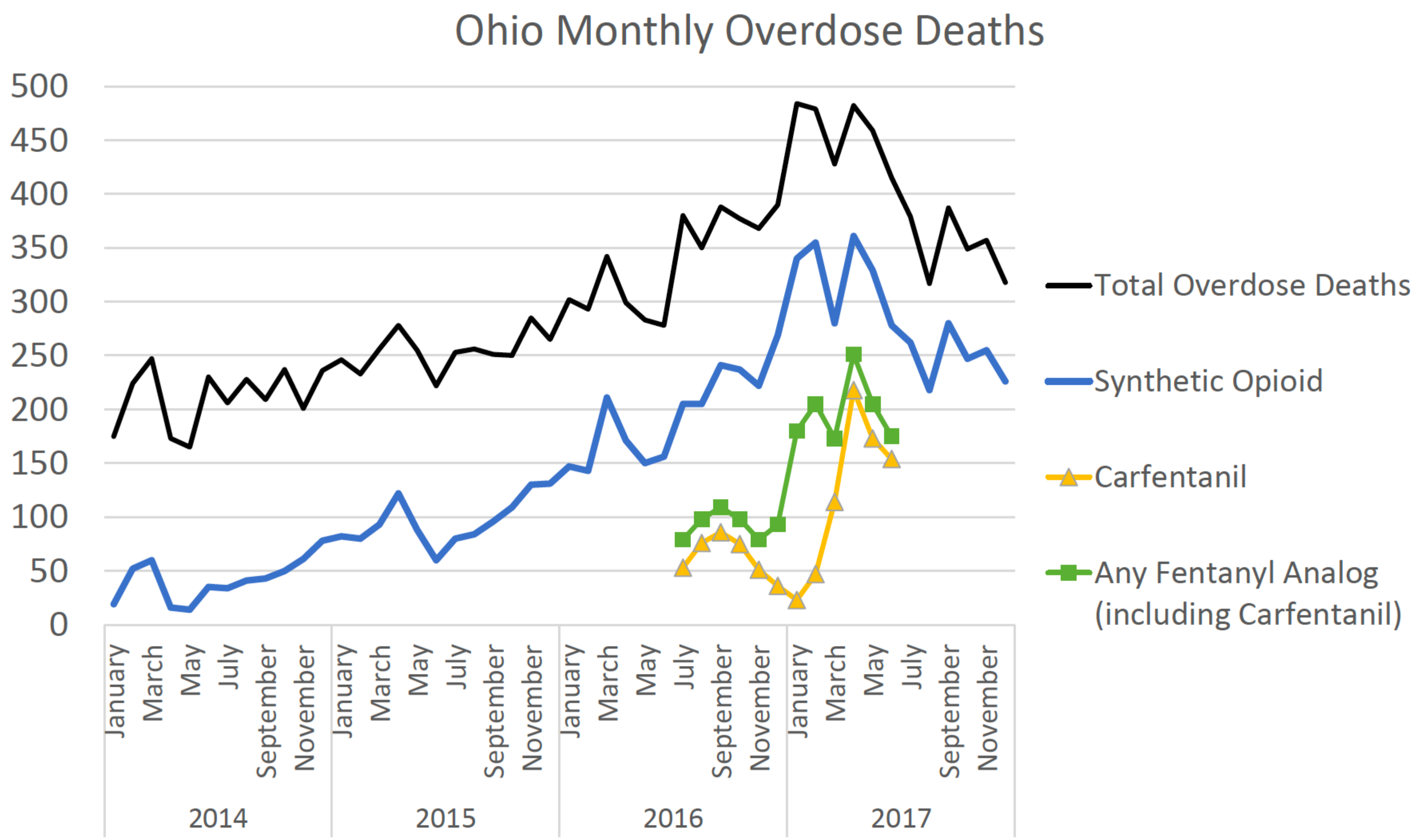
Note: Each data point is a month.


Summit County (Outside Cleveland): 3-Month Smoothed Monthly Overdose Deaths vs Drug Test Counts
Deaths, Number of Positive Drug Tests
45
40
35
30
25
20
15
10
5
0
January
March
May
July
September
November
2015
2016
2017
Overdose Deaths
Fentanyl Overdose Deaths



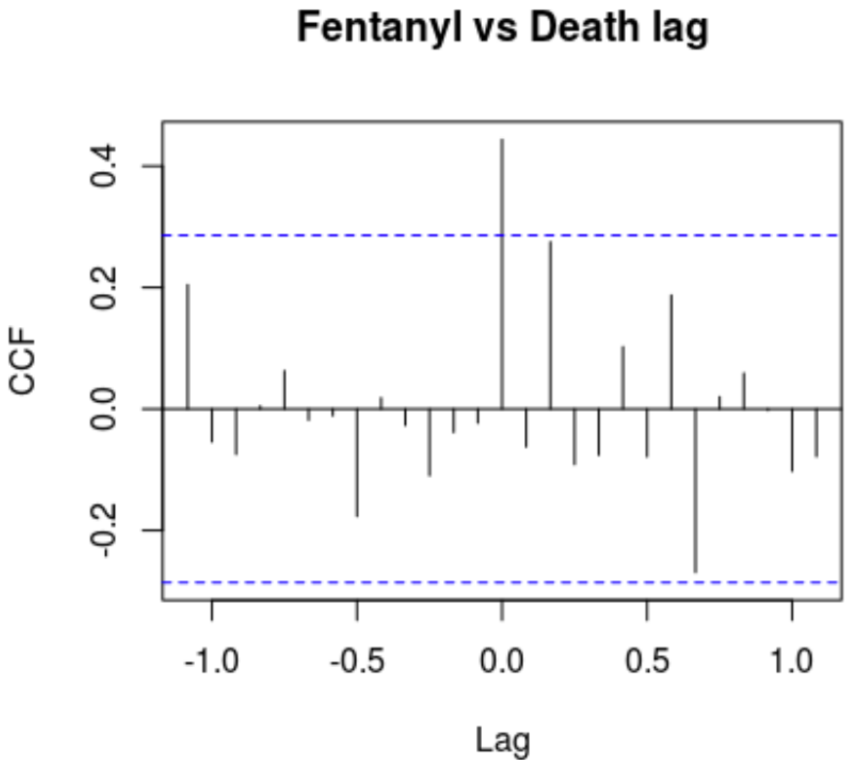
Summit County (Outside Cleveland): 3-Month Smoothed Monthly Overdose Deaths vs Drug Test Counts
Deaths, Number of Positive Drug Tests
45
40
35
30
25
20
15
10
5
0
January
March
May
July
September
November
2015
2016
2017
Overdose Deaths
Fentanyl Overdose Deaths
Fentanyl Test


Summit County (Outside Cleveland): 3-Month Smoothed Monthly Overdose Deaths vs Drug Test Counts
Deaths, Number of Positive Drug Tests
45
40
35
30
25
20
15
10
5
0
January
March
May
July
September
November
2015
2016
2017
Overdose Deaths
Fentanyl Overdose Deaths
Fentanyl Test
Carfentanil Tests
Other Fentanyl Analogs Tests



Ohio Monthly Overdose Deaths
500
450
400
350
300
250
200
150
100
50
0
January
March
May
July
September
November
January
March
May
July
September
November
January
March
May
July
September
November
January
March
May
July
September
November
2014
2015
2016
2017
Total Overdose Deaths
Synthetic Opioid
Carfentanil
Any Fentanyl Analog (including Carfentanil)

A natural question: are police seizures lagging behind deaths?
Cross-correlation function (CCF) shows highest correlation is at lag 0.



For every integer, h, the CCF at h is the correlation between:
x_t = fentanyl seizures at time t, and
the shifted time series
y_{t-h} = deaths h months ago,
after both are pre-whitened.

Here, the only statistically significant cross-correlation is h = 0. There is no lag between seizures and deaths.

- Findings: Drug seizure composition and weight have strong predictive value for drug overdose deaths.
- Before interpreting R^2 or quantifying impact of each additional BCI test, we need to be sure our residuals are random and independent.
- Our data points represent months. Number of deaths in January is probably related to number of deaths in February! So, not independent!
- The solution: time series regression.
- First, we will need to determine how Y_t depends on its own history, and build that into the model.

Time Series Regression

How much do deaths at time t depend on deaths in previous months?
Goal: fit an ARIMA(p,d,q) model ("autoregressive integrated moving average")
AR(p) is Deaths_t depends on Deaths_{t-1}, Deaths_{t-2}, …, Deaths_{t-p}
I(d) if you have to do "differencing" d-times to make the time series stationary.
MA(q) if Deaths_t depends on ε_{t-1}, ε_{t-2}, …, ε_{t-q}

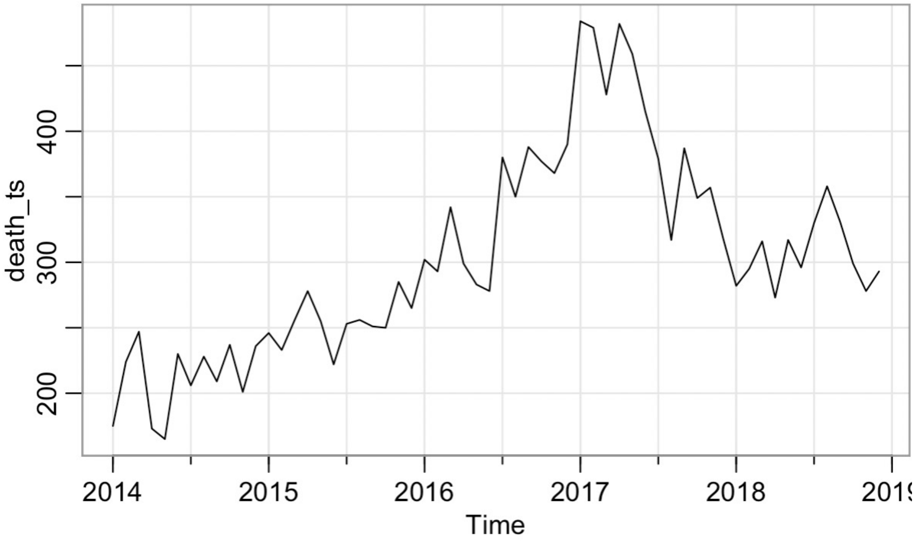
Find optimal (p,d,q) using

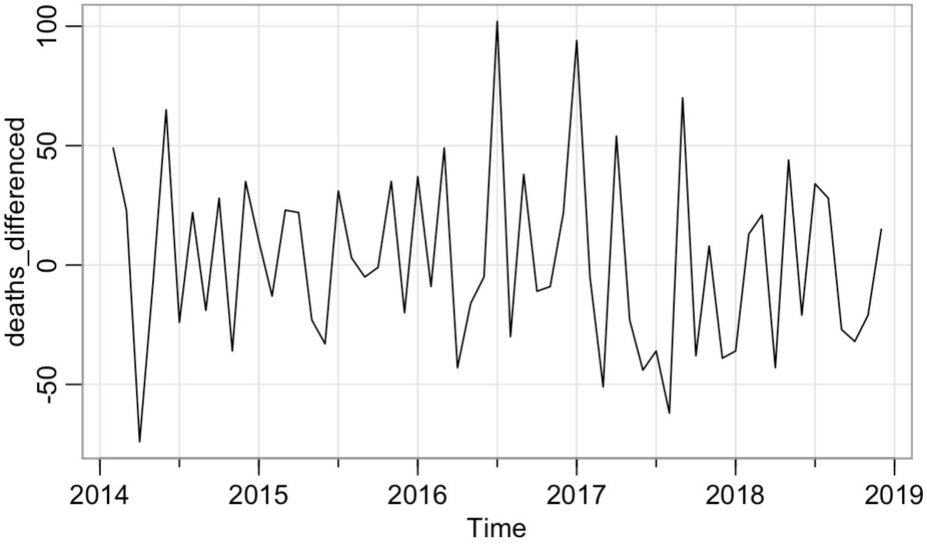
1. Autocorrelation function (ACF) = $\text{corr}(\text{Deaths}_t, \text{Deaths}_{t-h})$ for all lags h.
2. Partial autocorrelation function (PACF) = autocorrelation that remains after removing "carried over" autocorrelation; useful for error terms ε_{t-h}

Next, use Seizures_t to predict Deaths_t
Recall: no lag between police seizures and overdose deaths.

Death time series is not stationarity, so transform it

First difference operator: $\Delta\ \text{Deaths}_t = \text{Deaths}_t - \text{Deaths}_{t-1}$

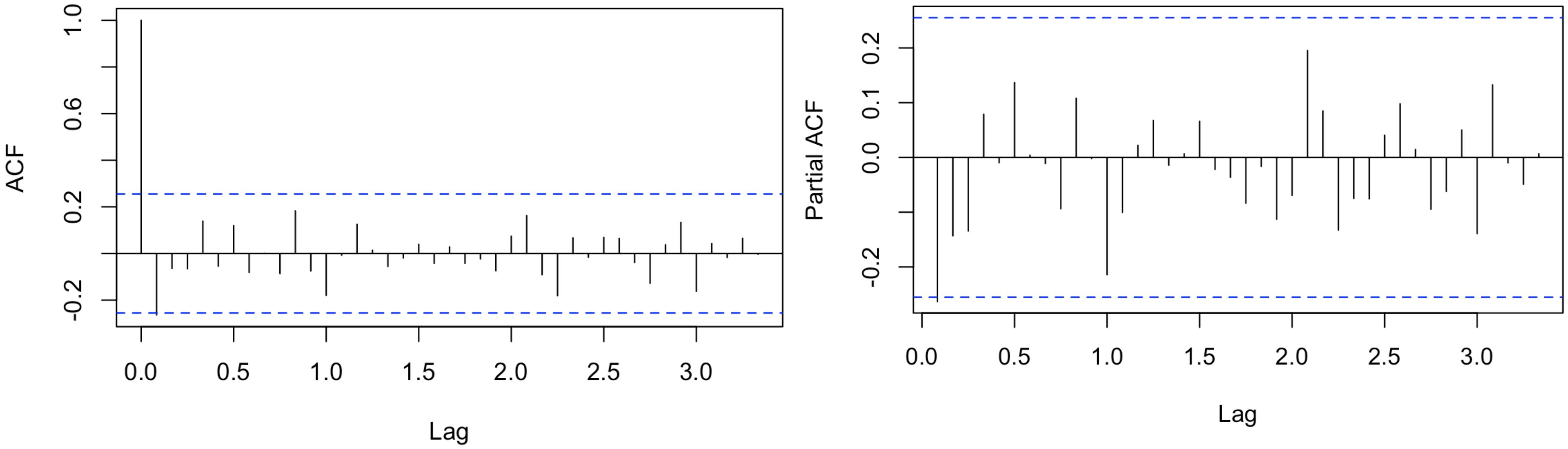




Δ Deaths$_t$ passes test for stationarity.

Next question: does Δ Deaths$_t$ depend on its own history? If so, how much?
We want random and independent residuals at the end of the day.

Autocorrelation function for $y_t = \Delta$ Deaths$_t$ and partial ACF



These graphs suggest we model Deaths$_t$ via ARIMA(0,1,1)
That is, Δ Deaths$_t$ depends on ε_{t-1} i.e., whether last month was unusual
This model also had the best AIC, and random and independent residuals.

Best models (ARIMA and GLMM)

- The best ARIMA model for drug overdose deaths is an ARIMA(0,1,1) model.
- Best time series model for using seizures to predict deaths:

$$\text{Deaths}_t \sim \text{ARIMA}(0,1,1) + \beta_1 * \text{Seizures}_t + \beta_2 * \text{Weight}_t^{0to0.1} + \beta_3 * \text{Weight}_t^{0.1to0.24} + \varepsilon_t$$

- Here $\text{Weight}_t^{0to0.1}$ is the number of seizures in month t of weight 0 – 0.1 grams (smallest weight), and $\text{Weight}_t^{0.1to0.24}$ similar.
- There is also a county-level model (GLMM) with R^2 of 0.88
- 20 more seizures of fentanyl predict for 3 more deaths.
- 20 more seizures of carfentanil predicts for 5 more deaths.
- We did all this work in eight weeks, and now the early warning system is in place
- Separate paper: what traits are correlated with overdoses

Ohio SUDORS Data (2016-2018)

- 9,300 individuals who died of drug overdose, 750 attributes for each: demographics, mental health/substance abuse history, personal problems, relationship status, job status, bystanders, Naloxone, polysubstance abuse, etc.
- Data curated by the Ohio Department of Health, but they were uninvolved with our paper.
- Includes data from law enforcement, coroners, hospitals, prisons, mental health treatment centers, etc.
- Never previously analyzed in Ohio. Our analysis is based on a 2018 paper on the Rhode Island SUDORS dataset.
- We made a bunch of summary tables like “what % of people had X”
- There is much left to be done! No one else has looked at this data.

Harm Reduction talking points

- Naloxone – convincing people to carry and use it
- Fentanyl test strips – determine if drugs have fentanyl
- Good Samaritan laws
- Needle exchanges + educating those who come
- Medication-Assisted Treatment for addiction
- Drug Courts, and treating drug users like human beings
- Counseling for those with mental health disorders
- Alternative treatments for pain
- Marijuana legalization – does it help?
- Study fentanyl analogues: there are many; unknown strength.
- Punchline: Harm Reduction saves lives and is much more cost effective (and ethical) than letting people suffer/die.

Much more research that can be done:

- Any question featuring time series analysis, e.g. interrupted time series, changepoint detection, spectral/Fourier.
- Any question using the spatial (geolocation) component of the data. Topological data analysis approaches.
- Polysubstance abuse
- Interaction terms, e.g., polysubstance and race.
- Danger for recent releases from prison to overdose
- Overdoses and local laws
- How many drug users are there? We only see deaths.
- How many lives would it save to open one more needle exchange? Where to open it? Is it cost effective?

Key Take-Aways

- There are tons of freely available datasets about the opioid epidemic: CDC, SUDORS, BCI, NSDUH, hospitals, prescriptions (OARRS), marijuana laws, NFLIS (DEA), FBI, OMAS Medicaid survey, mental health, etc.
- Many have never been analyzed – lots of low-hanging fruit, and nowhere near enough math/stats people working on this.
- Even simplistic analyses are valuable to harm reduction professionals, can save lives, and can get published. Great for students.
- Huge need to educate people about what actually works to combat the overdose epidemic.
- I’m happy to help you get involved, sharing datasets, analyses, papers, connecting you to experts, GitHub repository with time series models.
- Email: david.white@denison.edu

Second research project using ARIMA models

Summer of 2020, protests driven by a desire for more racial justice in policing (following the murder of George Floyd by police).

Police used tear gas and rubber bullets (KIPs) to try to control protesters. A group of eye doctors wrote a paper calling for police to stop using KIPs.

Referee asked (paraphrased) “how do you know things wouldn’t have been worse without the use of KIPs? Police claim they only shoot the bad protesters, and protests become less violent after KIP use.”

Joint with Nancy Rodriguez of Colorado.

ACLED protest data set

ACLED data set gathered from news reports. One row per protest (going back many years). Columns telling:

- Date of protest
- How many protesters were there
- Were police there?
- Did police use KIPs? Tear gas? Etc.
- Was protest violent?
- How many protesters injured? How many police injured/died?

Wrangle the data to have one row per day with columns for number of protests, protesters, injuries to protesters/police, KIP use.

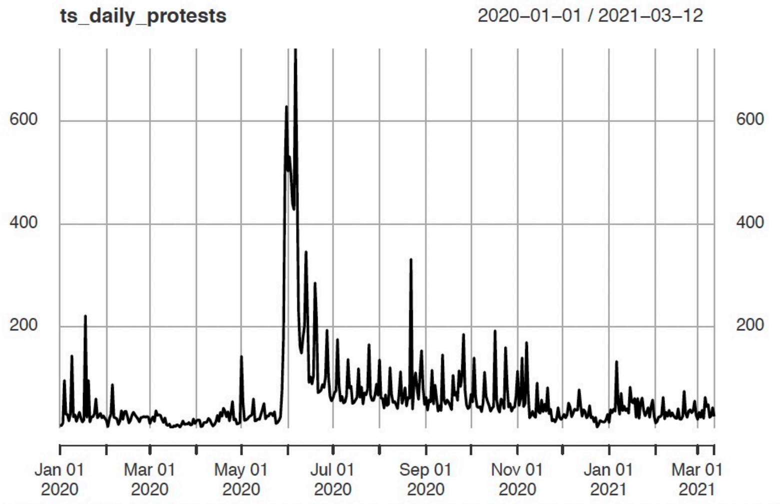
Protests and KIP use

Time series for number of protests per day and for KIP use per day.

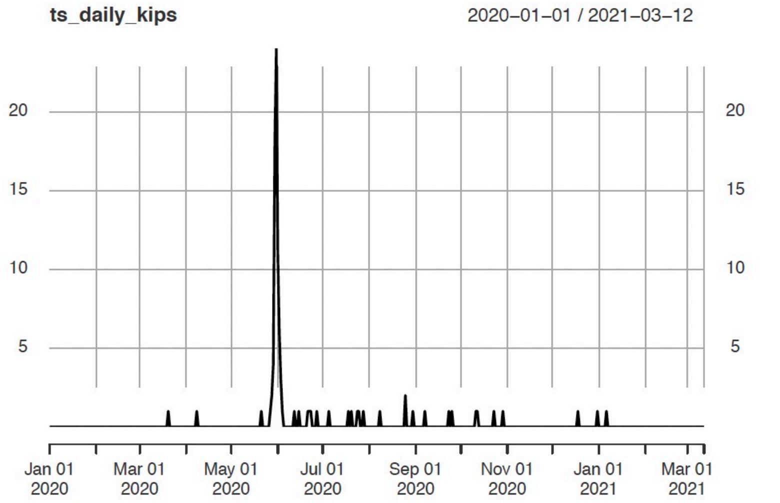
We see the big spike after George Floyd’s death, in May 2020.

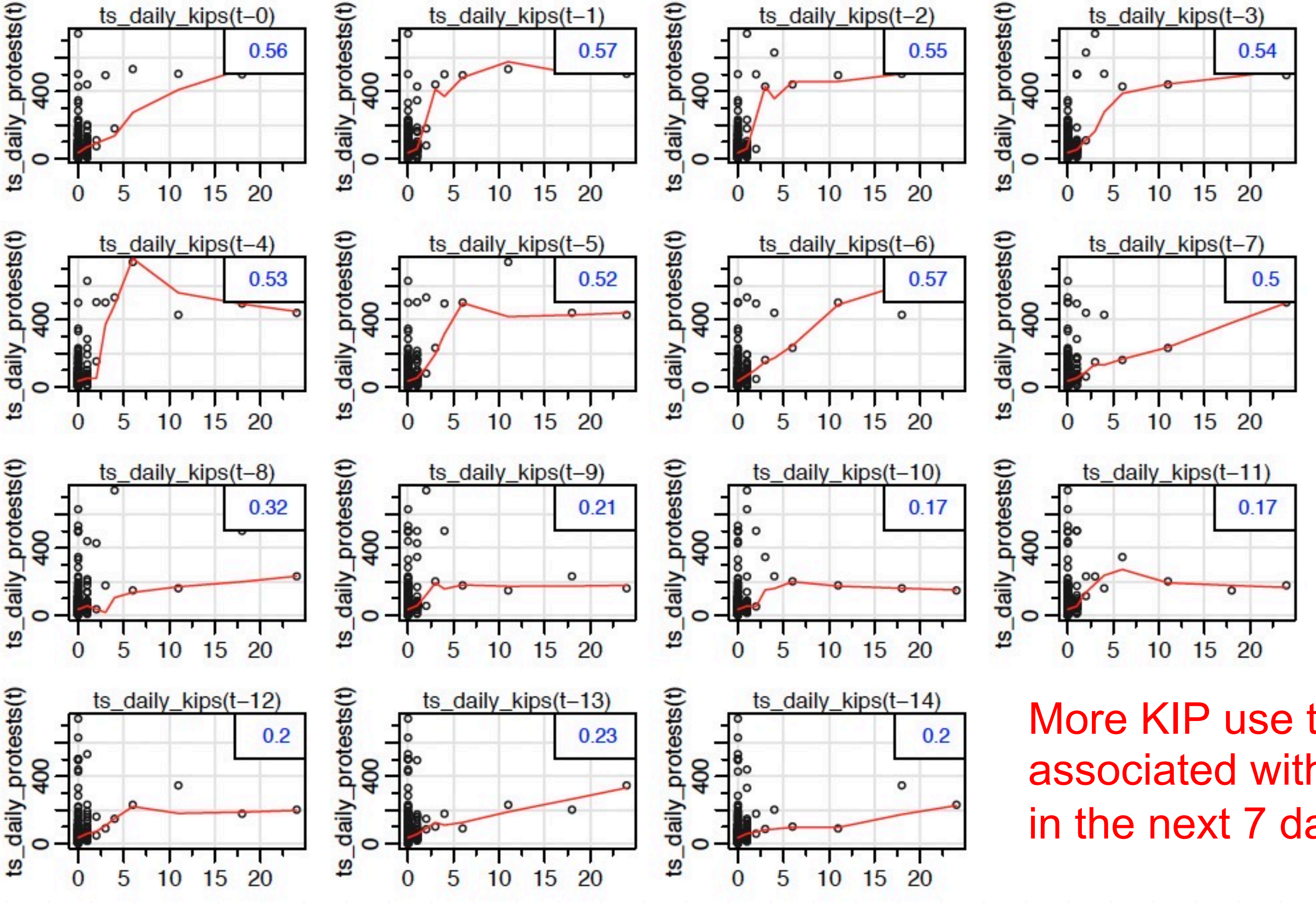
Does one time series lag behind the other?
Many protests lead to more KIP use?
KIP use leads to more protests the next day?

Can KIP use predict for number of protests?
For violence of protests?



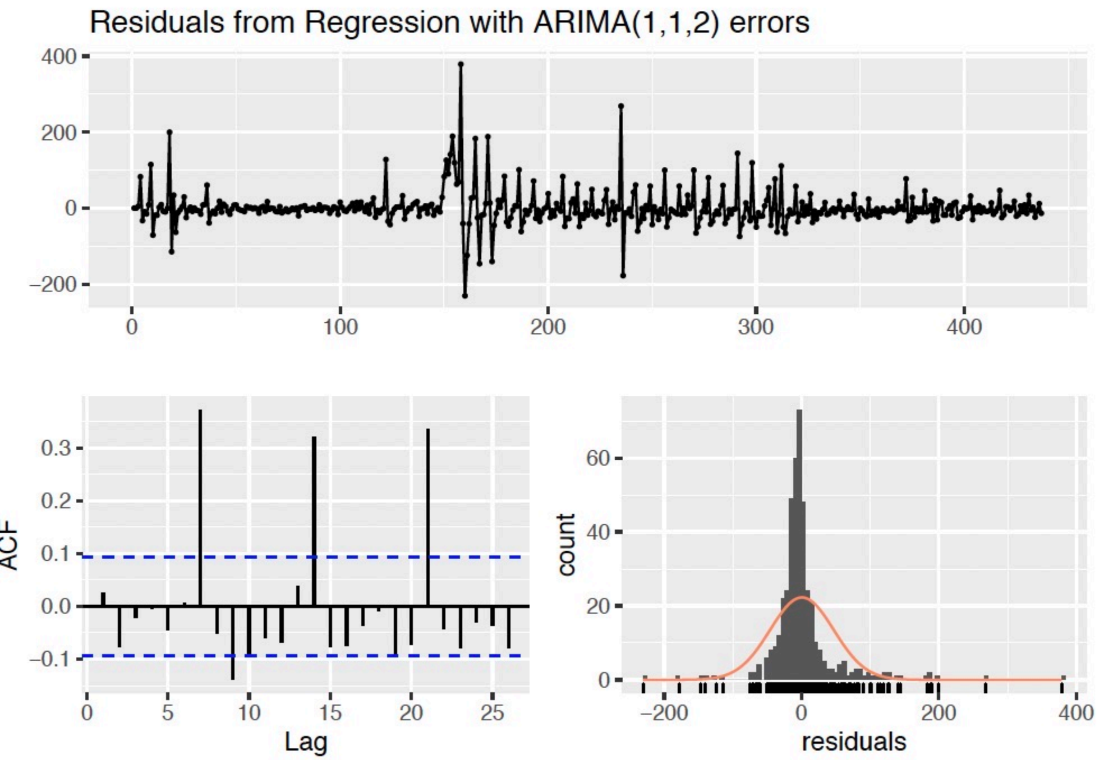

```
plot(ts_daily_kips)
```





More KIP use today is associated with more protests in the next 7 days.

First ARIMA model for protests alone

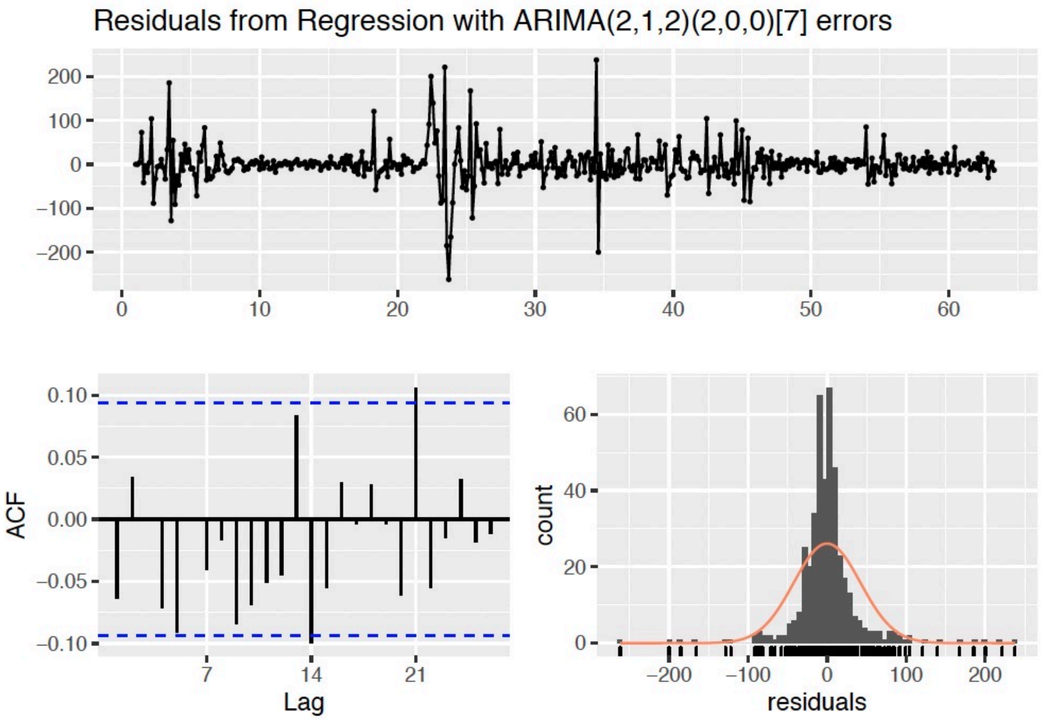


Can't see any obvious pattern in residuals except big spike in May 2020

ACF has a problem: significant lags at days 7, 14, 21, etc.

Also, residuals are not normally distributed.

Seasonal ARIMA to get at weekly effects



Autocorrelation: P_t depends on P_{t-1} and P_{t-2}

Seasonal autocorrelation: P_t depends on P_{t-7} and P_{t-14}

ACF is fixed now

Next: threshold model to get at self-exciting nature, then Hawkes Process

Punchline: KIPs don’t suppress protests

```
auto_mod = auto.arima(ts_daily_protests, xreg = ts_daily_kips)
summary(auto_mod)
```

```
## Series: ts_daily_protests
## Regression with ARIMA(1,1,2) errors
##
## Coefficients:
##          ar1      ma1     ma2     xreg
##       0.8728  -1.3048  0.3123  16.5311
## s.e.  0.0874   0.1273  0.1043   2.0386
##
## sigma^2 estimated as 2165:  log likelihood=-2291.62
## AIC=4593.25   AICc=4593.39   BIC=4613.64
```

Each use of KIPs is associated with 16.53 more protests.

Punchline: KIPs do cost life

```
auto_mod_d = auto.arima(ts_daily_deaths, xreg = ts_daily_kips)
summary(auto_mod_d)
```

```
## Series: ts_daily_deaths
## Regression with ARIMA(0,0,0) errors
##
## Coefficients:
##        intercept    xreg
##           0.3333  0.1349
## s.e.      0.0299  0.0187
##
## sigma^2 estimated as 0.3843:  log likelihood=-410.13
## AIC=826.25   AICc=826.31   BIC=838.49
```

Each use of KIPs is associated with 0.1349 more deaths

Third research project

2013 Euromaidan protests in Ukraine.

Police responded with arrests, beatings, rubber bullets, etc.
Protests grew; president fled.

A Ukrainian sociologist asked us to do a similar analysis about the impact of police procedures on the protests.

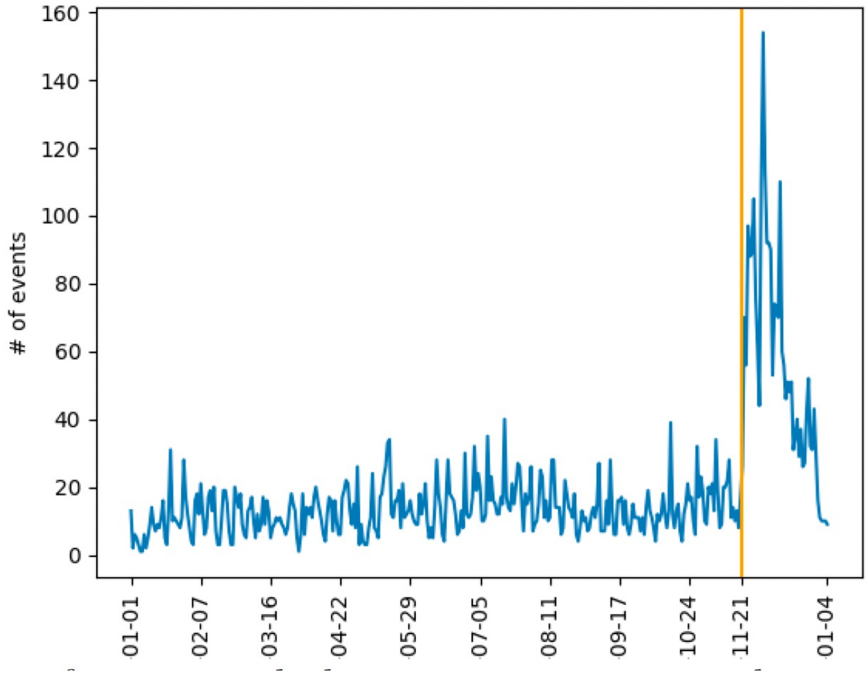
Research Team

Nancy Rodriguez
University of Colorado
Applied Mathematics

Yassin Bahid
University of Colorado
Applied Mathematics

Olga Kutsenko
Taras Shevchenko National
University of Kyiv
Dept of Sociology

The time series of protests (CSLR)



Euromaidan started Nov 21, 2013.

This “hockey stick” pattern is common in real-world time series. Like George Floyd protests.

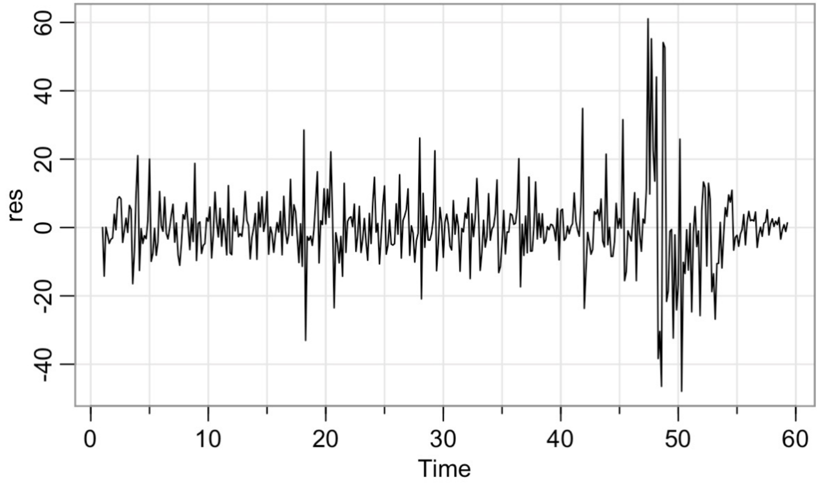
The time series is not “stationary” so it’s harder to model.

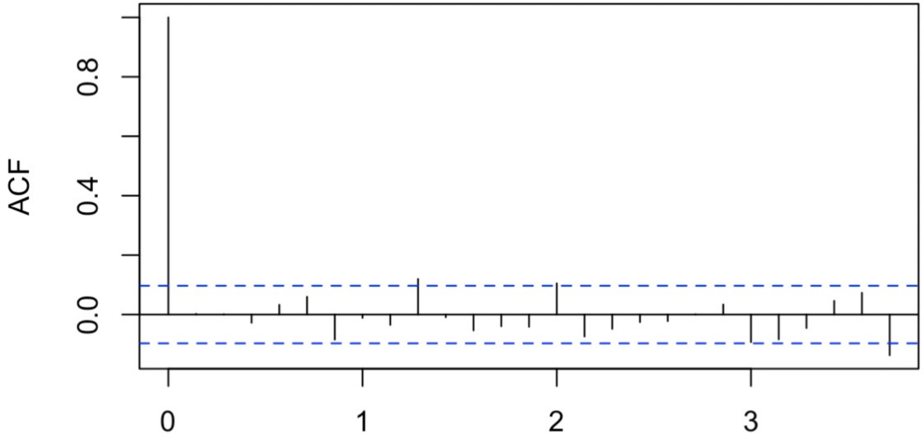
It’s “self-exciting” like the spread of an epidemic.

SARIMA model is good; shows self-excitation

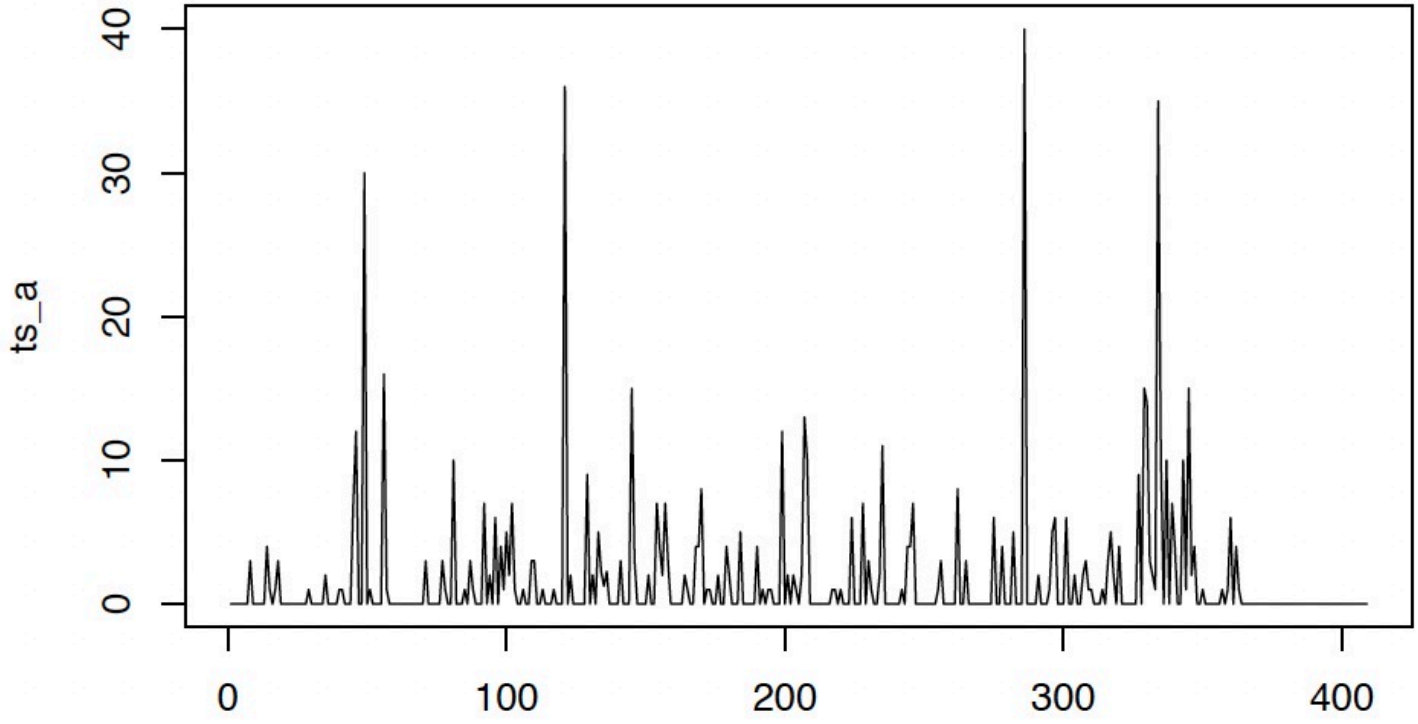
First-order differencing removed trend: $\Delta p_t = p_t - p_{t-1}$

	ar1	ar2	ma1	ma2	ma3	sar1	sma1	sma2
	1.4594	-0.7899	-1.8331	1.1981	-0.1720	0.9191	-0.9114	0.0902
s.e.	0.0702	0.0688	0.0913	0.1512	0.0742	0.0867	0.1048	0.0576



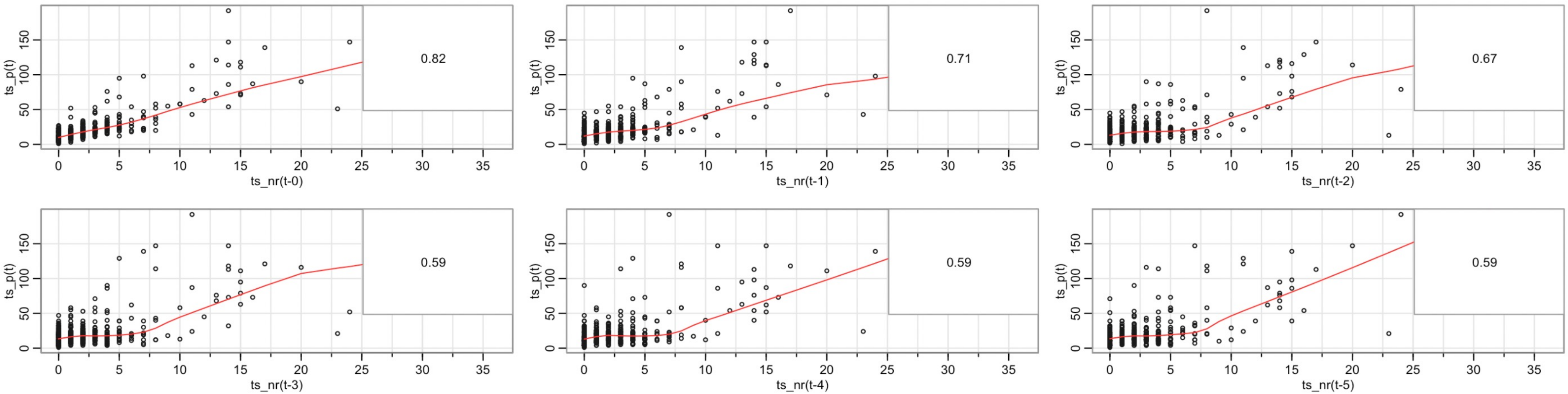


No data on KIPs but we do have on “negative response” events



Lagplots

- The strongest relationship between p_t and nr_{t-h} is at $h = 0$ and $h = 1$
- Negative responses today are correlated with protests



- Same for i_{t-h} (injuries) and e_{t-h} (Euromaidan events)

Multivariate model; useful for prediction

	lag 0	lag 1	lag 2	lag 3	lag 4
e_t	0.91	0.79	0.75	0.72	0.7
nr_t	0.82	0.71	0.67	0.59	0.59
i_t	0.24	0.16	0.09	0.08	0.09

Each of i_t, nr_t, and e_t has a statistically significant effect on p_t:

```
         ar1      ma1    sar1    sar2    ts_i   ts_nr    ts_e
      0.2341  -0.9466  0.1829  0.2399  1.0326  1.1902  0.7970
s.e.  0.0536   0.0188  0.0497  0.0485  0.2669  0.1144  0.0279
```

The model can be spelled out as:

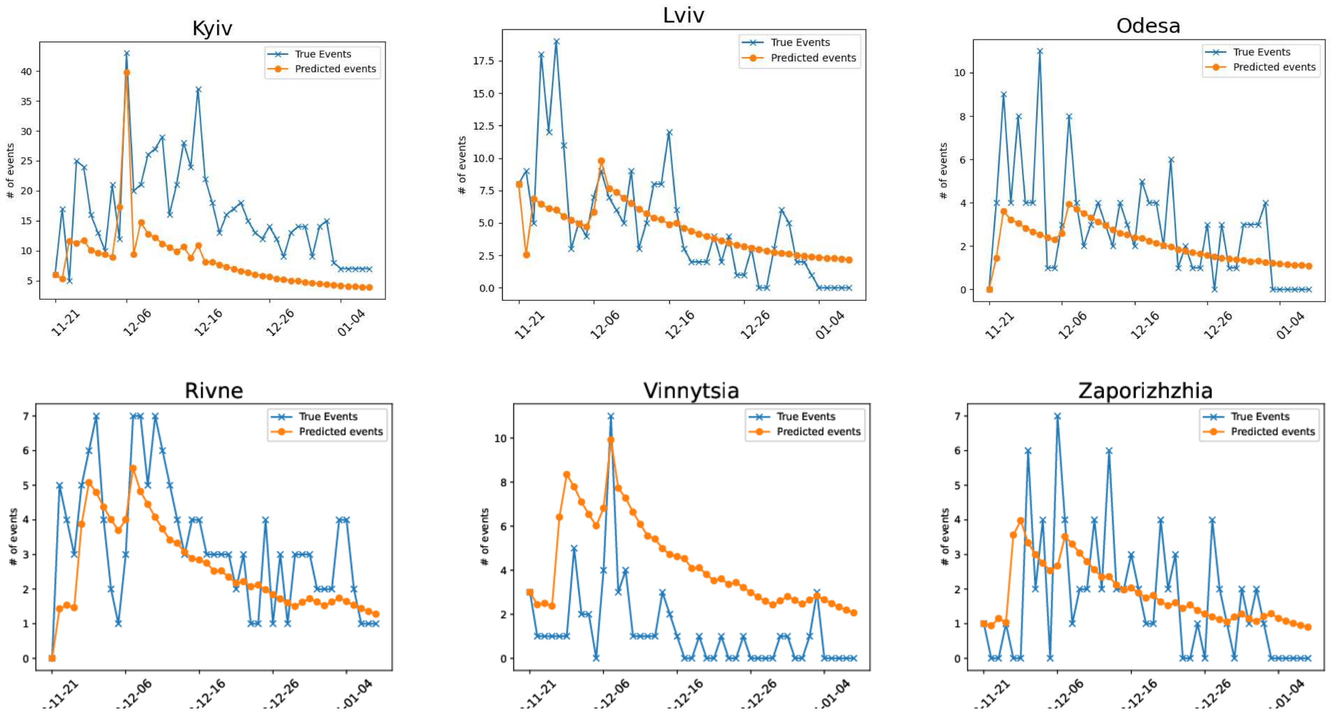
$$p_t = 1.0326 * i_t + 1.1902 * nr_t + 0.7970 * e_t + x_t$$

where the differenced series $X_t = \Delta x_t$ satisfies:

$$X_t = 0.2341 * X_{t-1} - 0.9466 * \epsilon_{t-1} + 0.1829 * X_{t-7} + 0.2399 * X_{t-14} + \epsilon_t$$

We also fit a threshold model, which “found” Nov 21, 2013.

Mathematical modeling via Hawkes process



Discussion

- The model excelled in predicting the spatial spread of events.
- The best spike times took into account the specific reactions of each oblast.
- Model accurately captures the spike in Kyiv on December 1st, 2013, and subsequent spread throughout other oblasts the following day.
- The political affinity between oblasts was a far more significant factor than the geographical distance between oblasts in determining the spread of protests.
- The fast spread of information through news and the internet makes physical distance less relevant.

Future work

- Get access to the 2014 data and use it to verify our model even more.
- Get access to missing data regarding number of protesters, and also magnitude of media coverage. Extend model to include these terms.
- Apply our framework to other countries and other protest time series.
- Find a data set with more granular information on “negative response” to quantify specific effects of rubber bullets, body armor, beatings, arrests, etc.

Key Take-Aways

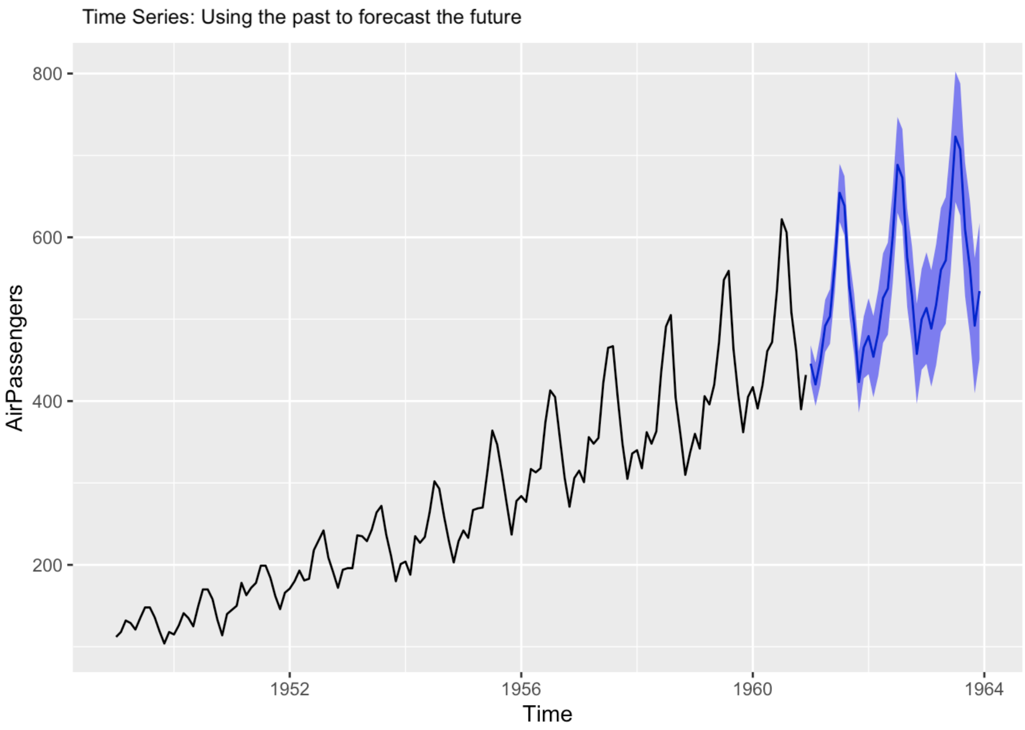
- Time series models are not that hard (for math majors), but social scientists often need help with them. Liberal arts training is critical.
- There are tons of freely available datasets that have never been analyzed. Lots of low-hanging fruit, and nowhere near enough math/stats people working in this area.
- Even simplistic analyses are valuable to social scientists and harm reduction professionals, can save lives, and can get published. Great for students.
- Social justice: in line with Kenyon’s mission!
- Much easier to talk to your friends about than abstract math!
- I created a GitHub repository with R Markdown files to carry out dozens of applied time series models on real-world data sets. Happy to share!

THANK YOU

References

1. Lin Ma, Lam Tran, David White, “A Statistical Analysis of Drug Seizures and Opioid Overdose Deaths in Ohio from 2014 to 2018,” *Journal of Student Research*, 10(1), 2021.
2. Lin Ma, Lam Tran, David White, State Unintentional Drug Overdose Reporting Surveillance: Opioid Overdose Deaths and Characteristics in Ohio, 2020.
3. Rodriguez and White: An analysis of protesting activity and trauma through mathematical and statistical models, *Crime Science* 12(17), 2023.
4. Bahid, Kutsenko, Rodriguez, White, The statistical and dynamic modeling of protests in Ukraine: the Revolution of Dignity and preceding times, accepted.

One application: Forecasting

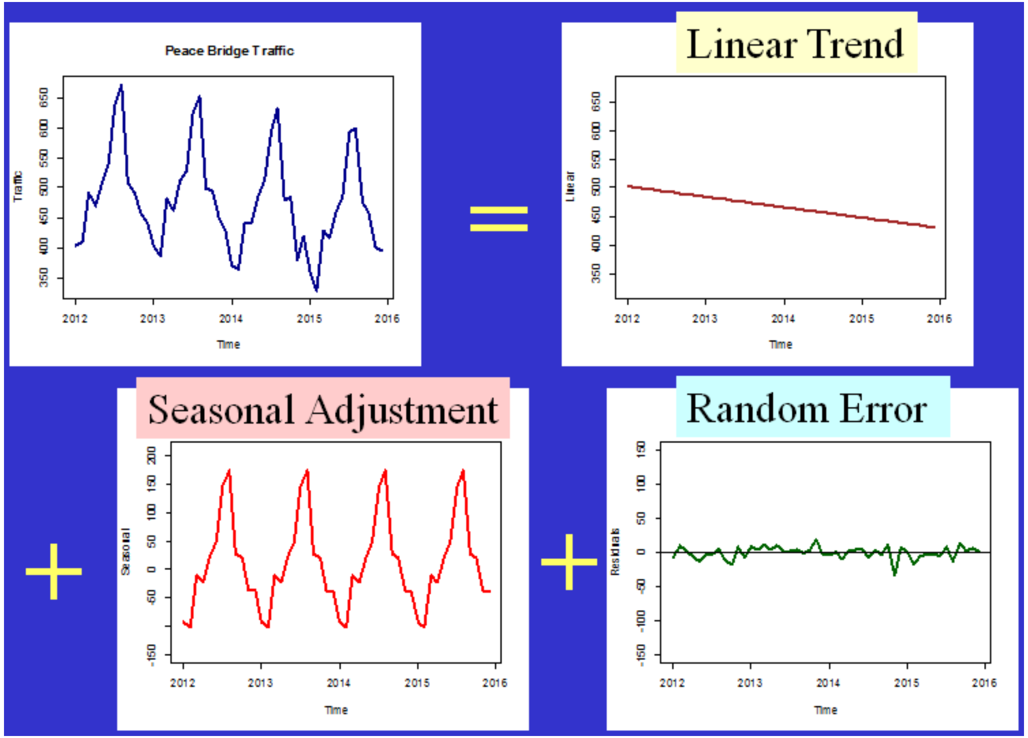


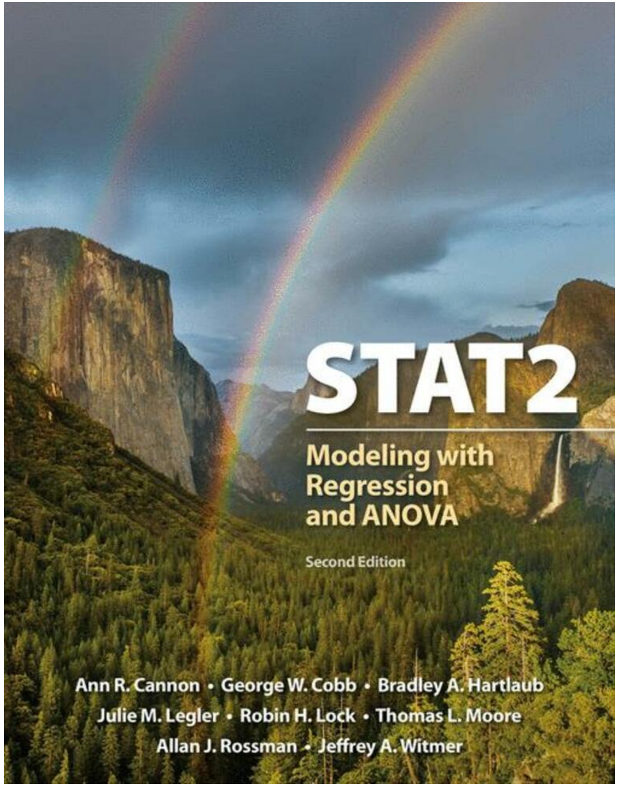
Example: Data = number of airline passengers each month from 1949-1960.

Fit a model that explains the growth and seasonal patterns, with random and independent residuals.

Forecast the next three years, plus 95% confidence interval. Use that to make money!

Decomposing a Time Series to get random residuals





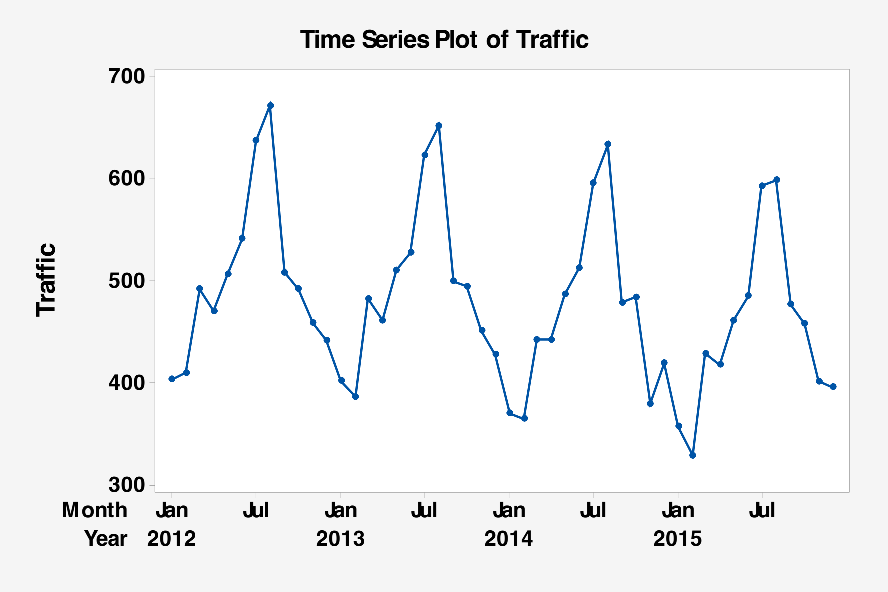
Example: Peace Bridge Traffic

Dataset: **PeaceBridge2012**

Monthly traffic (both directions in thousands of vehicles) between U.S. and Canada, 2012 to 2015

http://www.peacebridge.com/index.php/historical-traffic-statistics/yearly-volumes

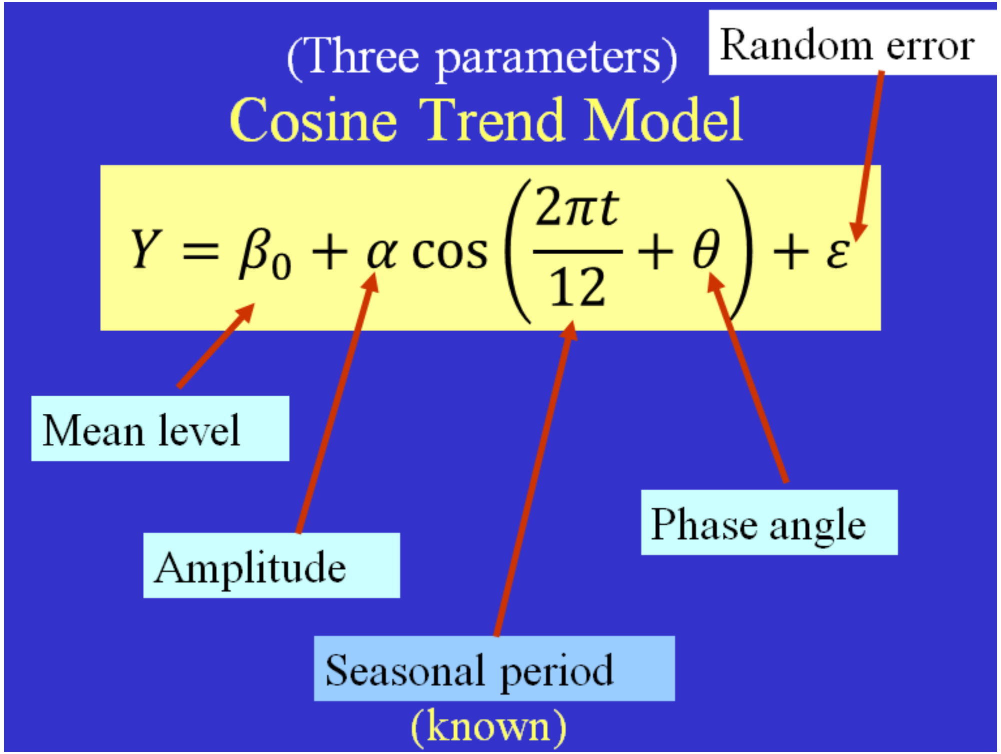
Time Series Plot: Bridge Traffic



What does this plot tell us?

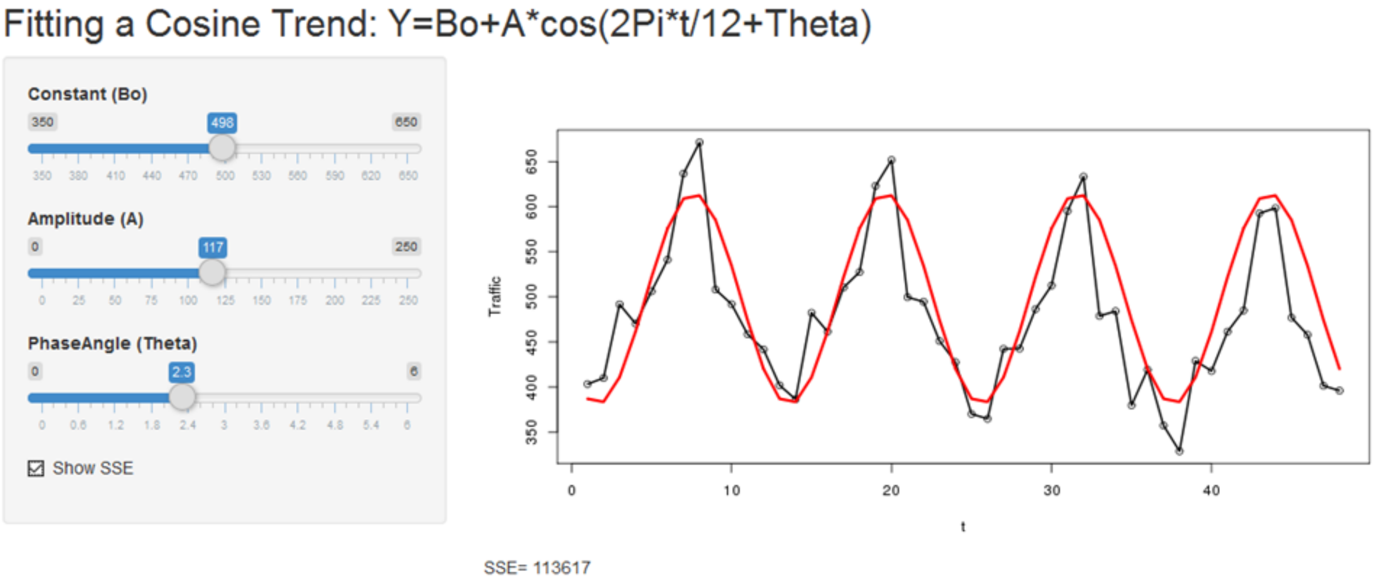
What mathematical model might apply? Functions that oscillate?

Bridge Traffic as a function of time



Fitting Cosine Trend Model: minimize sum of squared residuals

http://shiny.stlawu.edu:3838/sample-apps/CosineTrend/



Best fitting model (i.e. B0, A, Theta) minimizes the SSE = sum of squared residuals, just like linear regression.

Fitting a Cosine Trend

This is a nonlinear model

$$Y = \beta_0 + \alpha \cos\left(\frac{2\pi t}{12} + \theta\right) + \varepsilon$$

But with a little trigonometry…

$$Y = \beta_0 + \alpha\cos(\theta)\boxed{\cos\left(\frac{2\pi t}{12}\right)} - \alpha\sin(\theta)\boxed{\sin\left(\frac{2\pi t}{12}\right)} + \varepsilon$$

Use two predictors: $X_{cos} = \cos\left(\frac{2\pi t}{12}\right)$ and $X_{sin} = \sin\left(\frac{2\pi t}{12}\right)$

$$Y = \beta_0 + \beta_1 X_{cos} + \beta_2 X_{sin} + \varepsilon$$

Cosine Trend for Peace Bridge Traffic

Regression Equation

Traffic = 478.35 - 77.94 Xcos - 62.01 Xsin

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	478.35	6.06	78.92	0.000	
Xcos	-77.94	8.57	-9.09	0.000	1.00
Xsin	-62.01	8.57	-7.23	0.000	1.00

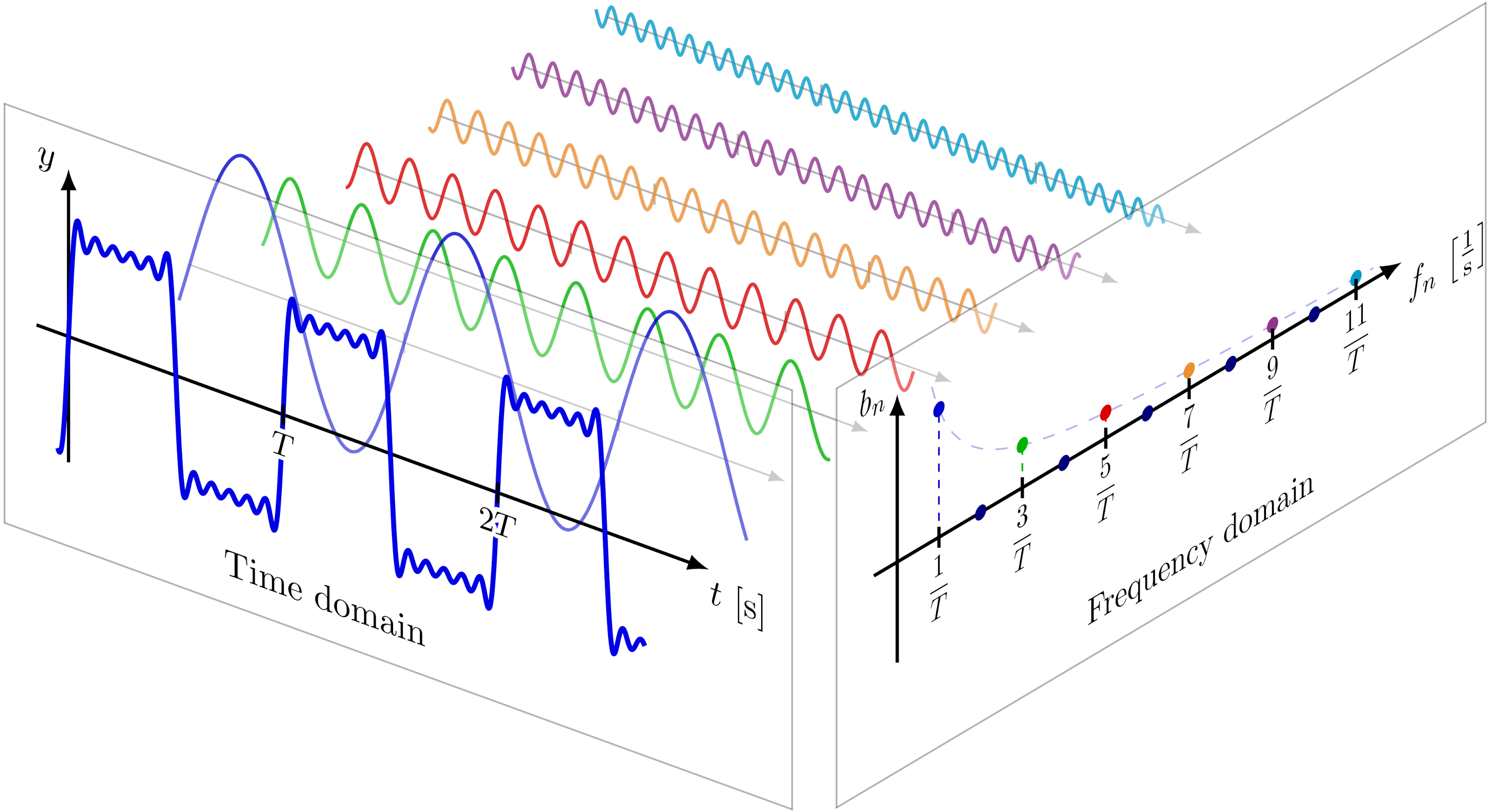
Model Summary

S	R-sq	R-sq(adj)
41.9961	75.00%	73.89%

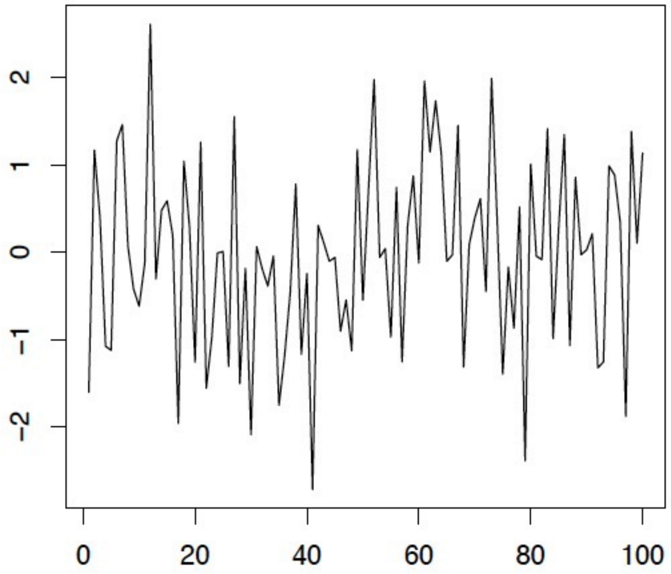
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	238147	119074	67.51	0.000
Xcos	1	145841	145841	82.69	0.000
Xsin	1	92267	92267	52.32	0.000
Error	45	79365	1764		
Total	47	317512			

Breaking a signal up into sum of its sinusoidal pieces, like Taylor series



Fourier models



Fit a model like:

$$\sum_{j=1}^{m} \left[A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t)\right]$$

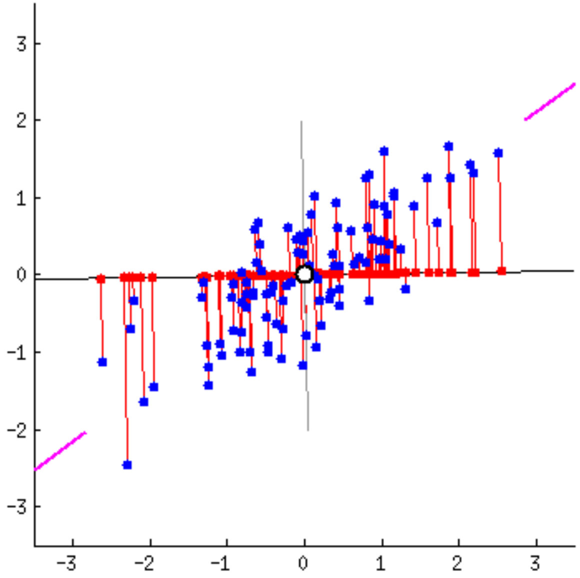
Problem: what's m? What are the periods?

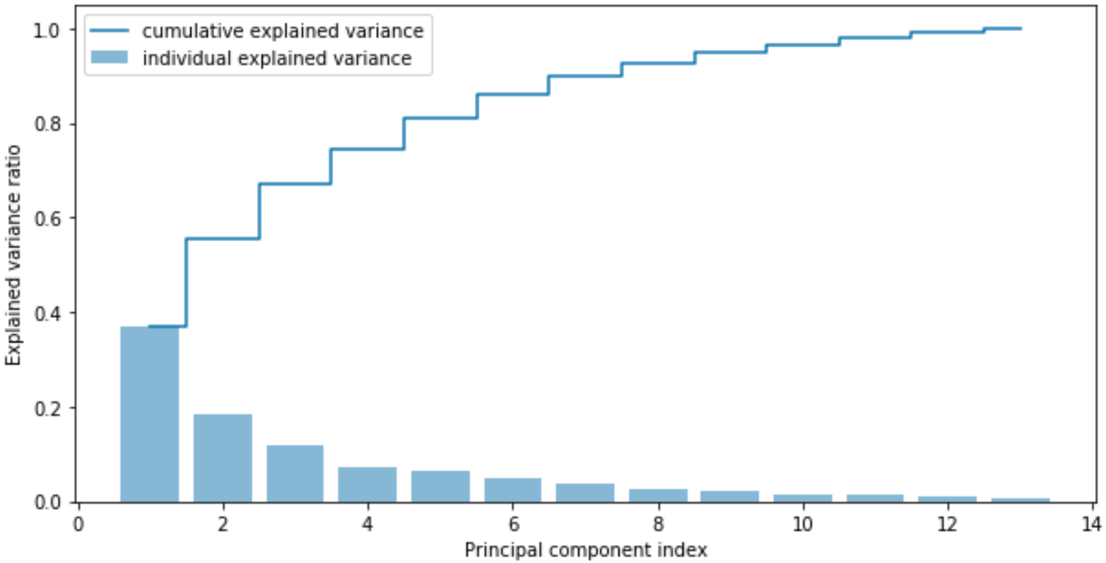
Solution: take the Fourier transform!

Idea: change to a basis where basis vectors correspond to periods, ordered by how much variability they explain. Then just keep the first few.

Same math as Shazam or Sound Hound!

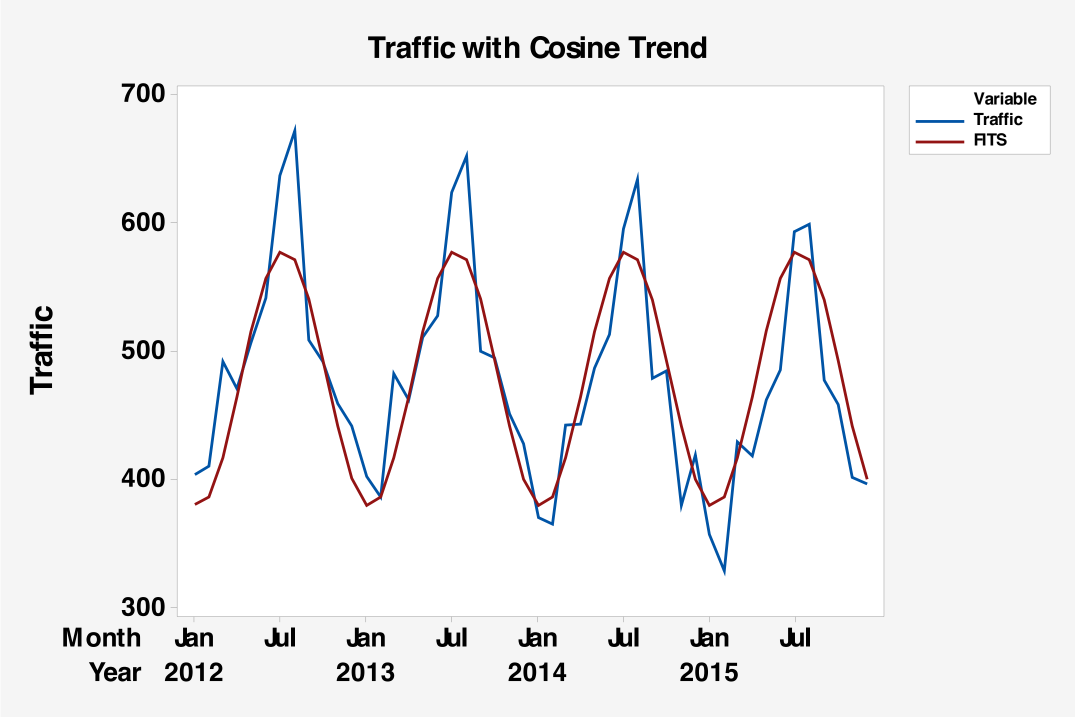
Principal Component Analysis (PCA)





Exact same concept as PCA. Accomplish via Singular Value Decomp.

Bridge Traffic with Cosine Trend



Seasonal Means Model

Basic Idea: Allow a separate value (mean) for each seasonal period (month)

Could find the sample mean for each month OR

Use regression with indicators for the months

$$Month7 = \begin{cases} 1 & \text{if July} \\ 0 & \text{otherwise} \end{cases}$$

$$\boxed{Y = \beta_0 + \beta_1 Month2 + \beta_2 Month3 + \cdots + \beta_{11} Month12 + \varepsilon}$$

Note: Need to leave one month's indicator out. Intercept (β_0) gives mean for that month. Other coefficients measure change to the other months.

Bridge Traffic: Seasonal Means

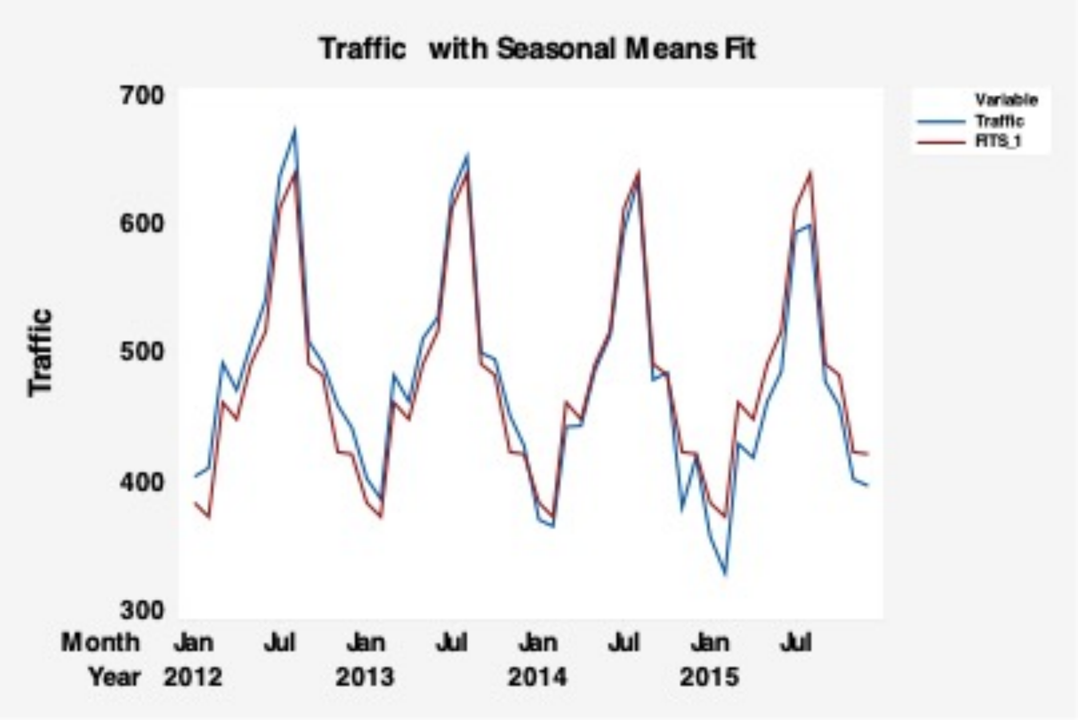
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	383.1	13.0	29.58	0.000	
Month					
2	-10.7	18.3	-0.58	0.564	1.83
3	78.2	18.3	4.27	0.000	1.83
4	64.9	18.3	3.54	0.001	1.83
5	108.0	18.3	5.90	0.000	1.83
6	133.5	18.3	7.29	0.000	1.83
7	229.0	18.3	12.50	0.000	1.83
8	255.8	18.3	13.96	0.000	1.83
9	107.8	18.3	5.88	0.000	1.83
10	99.0	18.3	5.41	0.000	1.83
11	39.6	18.3	2.16	0.038	1.83
12	37.9	18.3	2.07	0.046	1.83

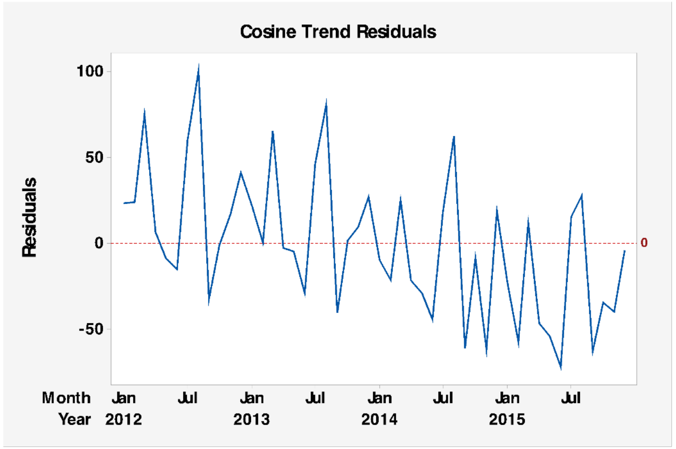
Model Summary

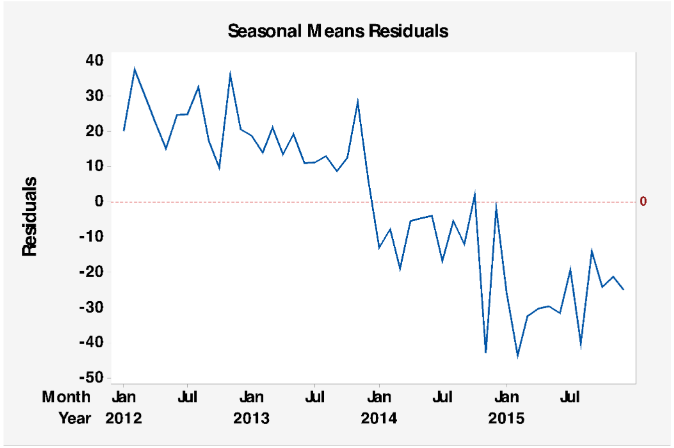
S	R-sq	R-sq(adj)	R-sq(pred)
25.9048	92.39%	90.07%	86.47%

Bridge Traffic with Seasonal Means



Residuals for Bridge Traffic Cosine Trend and Seasonal Means





Looks like a decreasing trend in both

$\Rightarrow$ Try adding a linear term to either seasonal model

Seasonal Means + Linear Trend

$$Y = \beta_0 + \beta_1 t + \beta_2 Month2 + \beta_3 Month3 + \cdots + \beta_{12} Month12 + \varepsilon$$

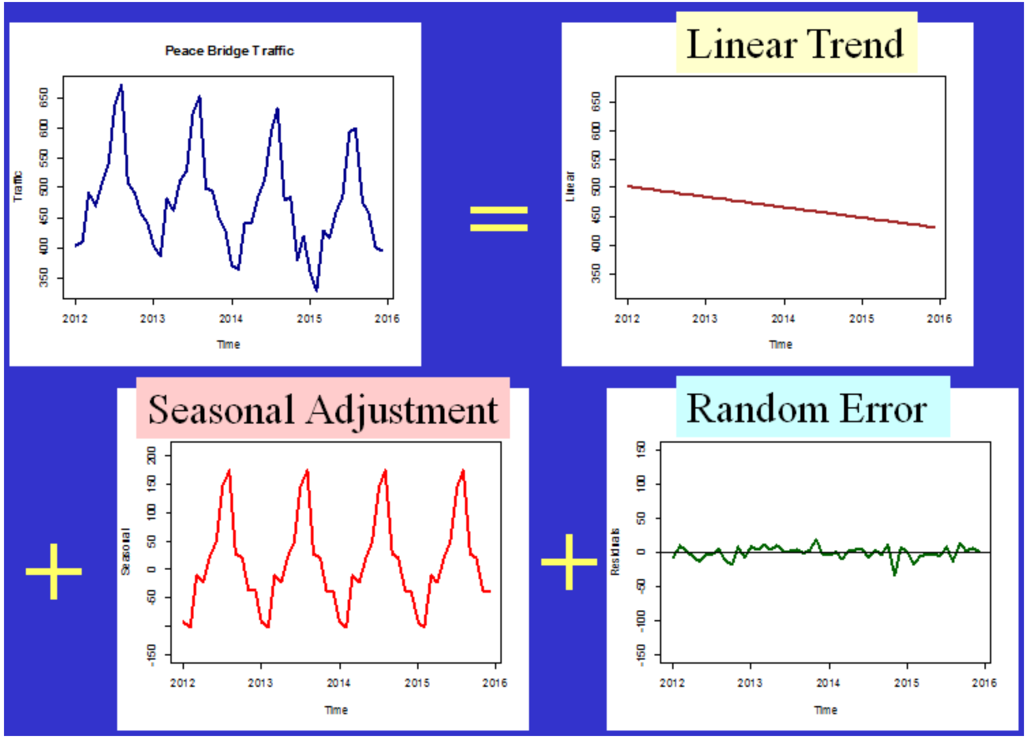
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	412.04	5.83	70.69	0.000	
t	-1.525	0.116	-13.13	0.000	1.07
Month					
2	-9.15	7.63	-1.20	0.239	1.83
3	81.27	7.63	10.65	0.000	1.84
4	69.47	7.64	9.10	0.000	1.84
5	114.12	7.64	14.93	0.000	1.84
6	141.10	7.65	18.44	0.000	1.84
7	238.12	7.66	31.08	0.000	1.85
8	266.45	7.67	34.72	0.000	1.85
9	119.97	7.69	15.61	0.000	1.86
10	112.77	7.70	14.64	0.000	1.87
11	54.80	7.72	7.10	0.000	1.88
12	54.70	7.74	7.07	0.000	1.88

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.7910	98.72%	98.28%	97.60%

Decomposing a Time Series



Cosine Trend vs. Seasonal Means

Cosine trend Fewer parameters (3 vs. 12)

Model Summary

S	R-sq	R-sq(adj)
41.9961	75.00%	73.89%

Seasonal means Better R^2, adjusted R^2, and $\hat{\sigma}_\varepsilon$

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
25.9048	92.39%	90.07%	86.47%

Why stop at cosine model with only 3 parameters? Fourier analysis gives functions that can capture more cycles in the data.

Characteristic	n	%
Age group		
18-24	725	7.79
25-44	5029	54.00
45-65	3366	36.14
Other	193	2.07
Sex		
Male	6348	68.16
Female	2965	31.84
Race/ethnicity		
Black, non-Hispanic	1069	11.48
White, non-Hispanic	8003	85.93
Hispanic	186	2
Other	55	0.59
Education level		
Less than high school	2071	22.61
High school Graduate/ GED completed	5111	54.88
Some college/ technical school or more	1974	21.55

Marital Status		
Never married / single	4902	53.01
Divorced/ separated	2480	26.8
Widowed	330	3.57
Married/ partnered	1535	16.69
Occupation		
Employed/ self-employed	7526	80.81
Unemployed	322	3.46
Unknown	1465	15.73
City/ town of residence		
Urbanized	5279	64.43
Urban clusters	478	5.83
Rural	2461	30.03
Out of state		
Injury location		
House or apartment	7551	81.08
Other	1336	14.35
Injured at victim home		
Yes	5696	35.5
No	3135	64.5

Precipitating Circumstance	**n**	**%**
Life stressor		
Physical health problem	565	6.55
Recent criminal legal problem	70	0.81
History of child abuse/ neglect	17	0.2
Job problem	54	0.63
Interpersonal		
Intimate partner problem	210	2.43
Family relationship problem	55	0.64
Suicide event		
History of suicide attempt	207	2.4
Precipitating circumstance known	8529	98.89

Mental health/Substance abuse		
Other substance abuse problem (excludes alcohol)	7363	86.55
Alcohol problem	1144	13.45
Current diagnosed mental health problem	3497	41.11
Depression/ dysthymia	174	-
Anxiety disorder	270	-
Bipolar disorder	83	-
Post-traumatic stress disorder	64	-
ADD or ADHD	61	-
Other mental problem	102	-
Current mental health treatment	2775	32.17
History of ever being treated for a mental health problem	4190	48.58

Toxicology test positive	**n**	**%**
Substance class and drug cause of death		
Opioid	8230	88.37
Fentanyl	7446	81.01
Heroine and/or Morphine	2662	28.96
Oxycodon	970	10.55
Methadone	262	2.85
Buprenorphine	216	2.35
Hydrocodone	446	4.85
Cocaine	3387	36.37
Alcohol	1901	20.41
BAC >= 0.08	42	2.68
BAC < 0.08	1525	97.32
Benzodiazepine	2433	26.12
Alprazolam	1204	13.10
Clonazepam	562	6.11
Marijuana	2156	23.15
Antidepressant	1331	14.29
Amphetamine	1178	12.65
Anticonvulsant	1015	10.90
Antipsychotic	274	2.94
Number of substance causing death		
1	781	8.53
2	945	10.32
3	1072	11.71
4	1022	11.16
5 or more	5337	58.28

Substance abuse	**n**	**%**
Previous drug overdose		
No previous overdose reported	7839	86.8
Previous OD within the past month	335	3.71
Previous OD between a month and 1 year prior	341	3.78
Previous OD that occurred more than 1 year prior	149	1.65
Previous OD, timing unknown	367	4.06
Recent opioid use relapse		
No evidence	8334	92.4
Relapse mentioned, timing unclear	290	3.22
Relapse occurred < 2 weeks of fatal overdose	363	4.02
Relapse occurred > 2 weeks < 3 months of fatal overdose	32	0.35
Treatment for substance abuse		
No treatment	7280	80.61
No current treatment, but treated in the past	1076	11.91
Current treatment	675	7.47

History of opioid abuse		
None	3045	33.72
Substance unknown	2259	25.01
Current or past abuse of heroin	3024	33.48
Prescription opioids	362	4.01
Both prescription opioids & heroin	341	3.78
Scene indications of drug use		
Any evidence of drug use	5884	65.05
Evidence of rapid overdose	797	8.81
Needle close to the body	681	7.53
Route of drug administration		
Evidence of injection	3033	33.53
Needle/syringe	2042	22.57
Track marks on victim	1712	18.93
Cooker	1097	12.13
Filter report	292	3.23
Touniquet report	337	3.73
Witness report	90	0.99
Evidence of ingestion	1745	19.29
Evidence of snorting	1139	12.59
Evidence of smoking	709	7.84

Drug type and response to drug overdose	n	%
Illicit drug	3413	37.73
Evidence of illicit drug: powder	1127	12.46
Evidence of illicit drug: crystal	108	1.19
Evidence of illicit drug: witness report	591	6.53
Prescription drug	2153	23.8
Prescribed to the victim	1445	15.97
Unknown who prescribed	782	8.64
Not prescribed to the victim	264	2.92
Form of prescription drug		
Pills/tablets	644	7.12
Bottle	1596	17.64
Patch	47	0.52
Response to drug overdose		
Bystander present at time of overdose		
No bystander present	173	1.92
One bystander present	1623	17.97
Multiple bystander present	690	7.64
Bystanders present, unknown num	1003	11.11

No person witnessed drug overdose	626	6.93
1+ person withnessed drug overdose	564	6.25
Naloxone administered		
Yes	1174	12.98
No	1802	19.92
Unknown	355	3.92
Who administered naloxone?		
By EMS/firefighter	642	7.1
By law enforcement	111	1.23
By hospital (ED/Inpatient)	117	1.29
By family member	32	0.35
By intimate partner	24	0.27
Medical history		
Yes, treated for pain	1954	21.63
No/Unknown	7080	78.37
Prescription information		
Prescribed buprenorphine/methadone	309	3.42

Discussion of the Data for Ukraine project

- Data from the Ukrainian Center for Social and Labor Research; academic researchers; unbiased data collection.
- Gathered from 190 newspapers (local and national). 6627 rows, each an “event” i.e., a rally, riot, or protest.
- Exploratory data analysis: columns for oblast, “negative response”, and “Euromaidan”.
- Missing data on arrests, injuries, deaths, and number of protesters. Small events unreported in the news.
- Wrangle the data to focus on number of protests per day. Now each row is a day, and a column tells how many events occurred.

Data wrangling

- Wrangle the data to focus on number of protests per day. Now each row is a day, and a column tells how many events occurred.
- Extract new time series:
  - p_t is the number of events on day t (that is, all events where t is between the start and end date, inclusive)
  - nr_t is the number of events with a “negative response” on day t
  - e_t is the number of events associated with Euromaidan on day t
  - i_t is the number of civilians injured on day t
- Which of these leads/lags the others? Do negative responses lead to more or fewer protests in subsequent days?

Cross-correlation analysis

- Relationship between p_t and nr_t? Does a negative response action today by the government predict more or fewer protests tomorrow?
- Can't naively fit a regression model $p_t = a + b\ nr_{t-1} + r_t$ (indep. fails)
- For every shift h, compute correlation of p_t and nr_{t-h} and take biggest
- To remove the effect of exogenous variables, prewhiten p_t to get the SARIMA residuals r_t, then filter nr_t the same way to get s_t, then compare r_t and s_{t-h} for all h.
- You can say when nr_{t-h} has a *statistically significant* effect on p_t