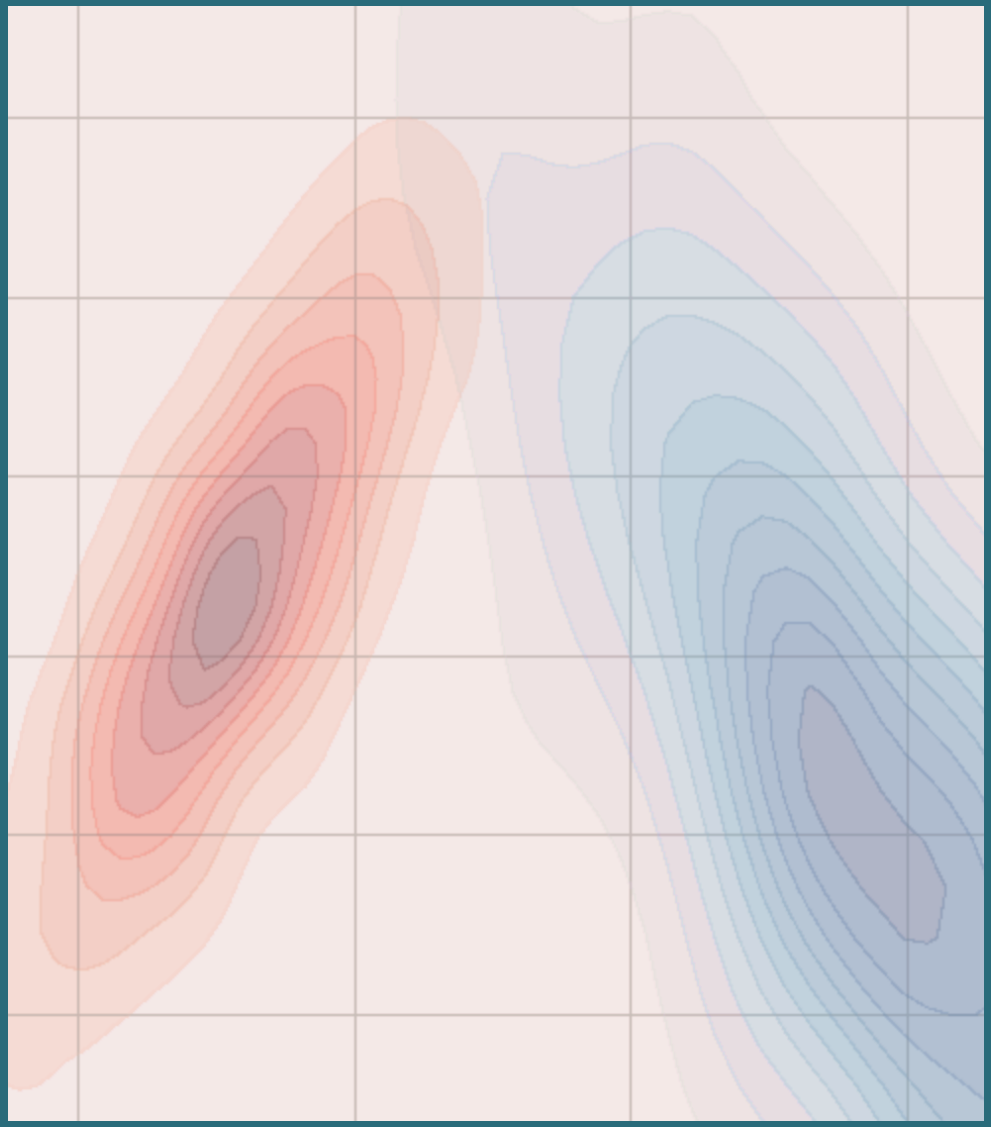
Interrupted Time Series Experiments in Python

Drew Fustin
Automaton Data
drewfustin@gmail.com
@drewfustin



Me

Automaton Data [contracting]
drew@automatondata.com

PhD, Physics

Data Scientist

SPOT HERO

GRUBHUB

Experiments are Hard

Some are Straightforward

Some are Straightforward

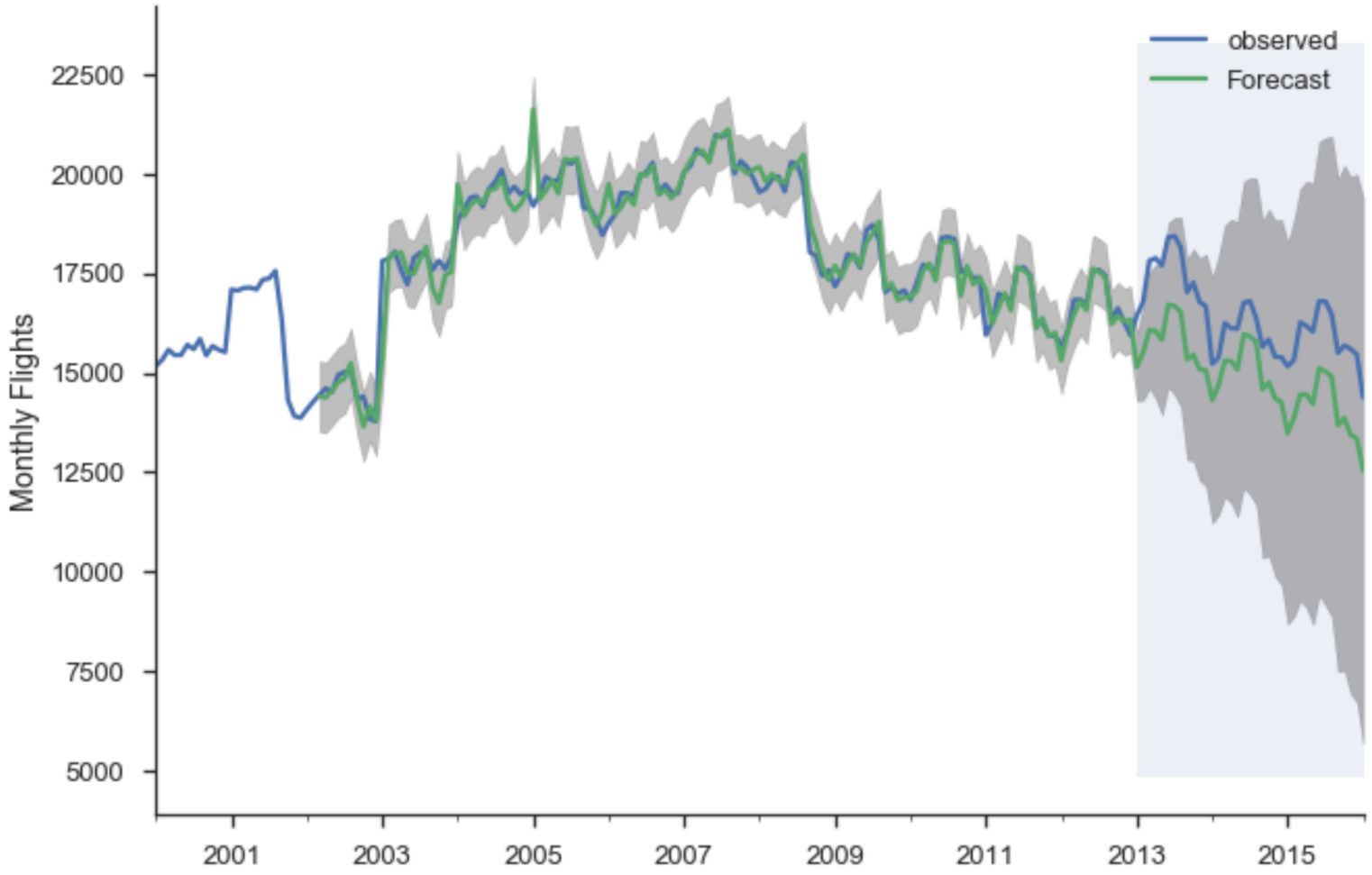
$$\Pr(p_B > p_A) = \sum_{i=0}^{\alpha_B - 1} \frac{B(\alpha_A + i, \beta_B + \beta_A)}{(\beta_B + i) B(1 + i, \beta_B) B(\alpha_A, \beta_A)}$$

Where:

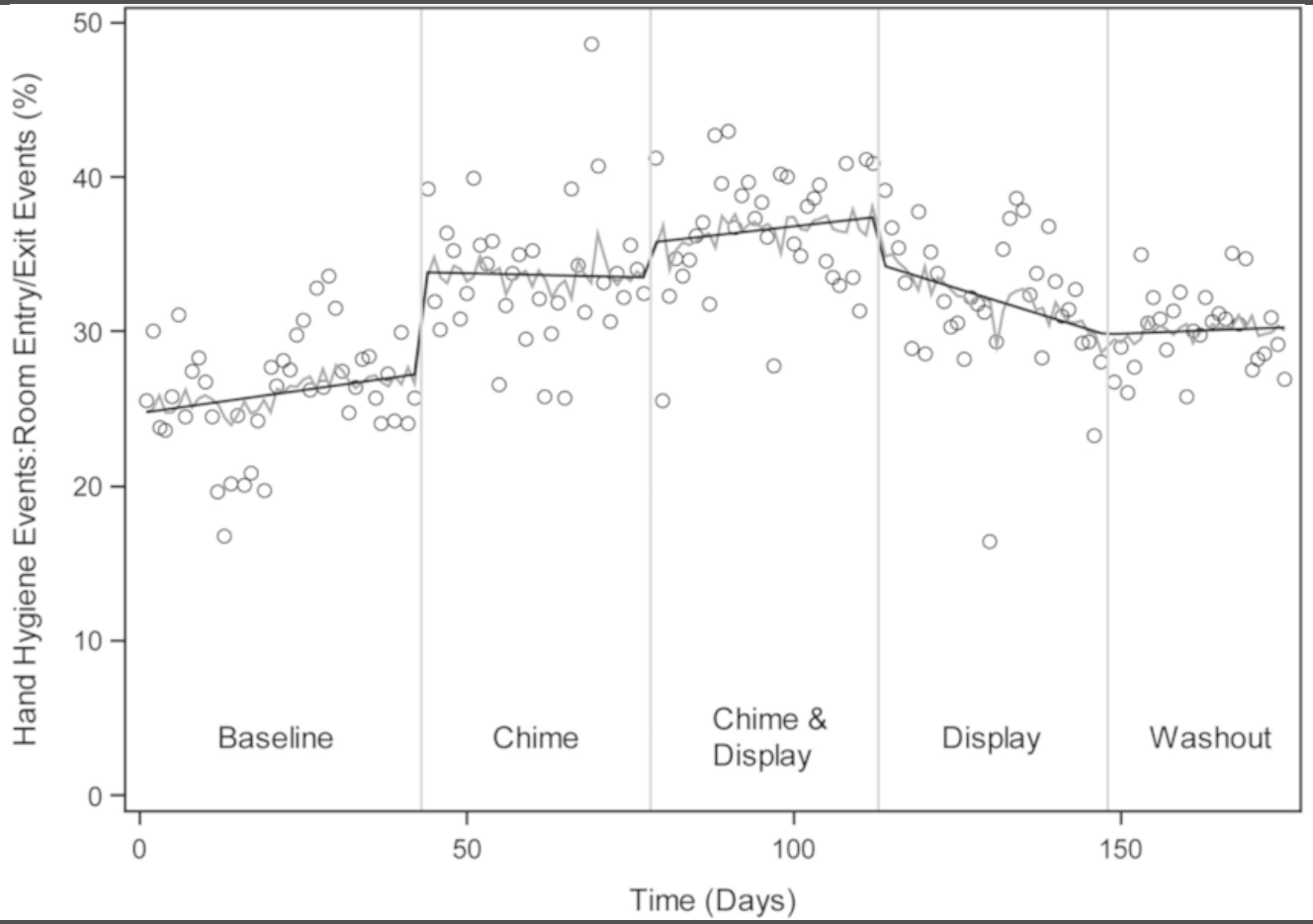
- α_A is one plus the number of successes for A
- β_A is one plus the number of failures for A
- α_B is one plus the number of successes for B
- β_B is one plus the number of failures for B
- B is the beta function

evanmiller.org/bayesian-ab-testing.html

Some are Convoluted

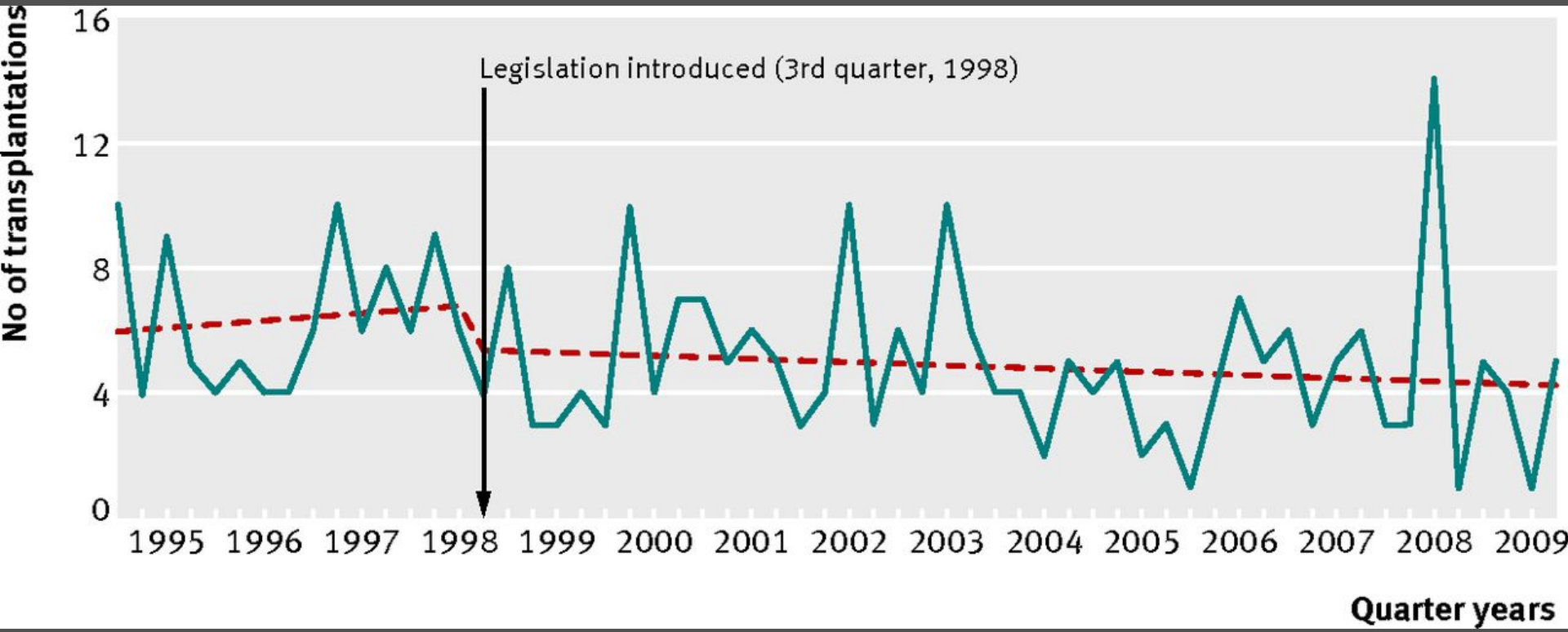


Interrupted Time Series



Ellison, et al
A Prospective Controlled Trial of an Electronic Hand Hygiene Reminder System

Interrupted Time Series



Hawton, et al
Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in England and Wales: interrupted time series analyses

Interrupted Time Series

So shaky, they're called
quasi-experiments

Interrupted Time Series

Related to
regression discontinuity

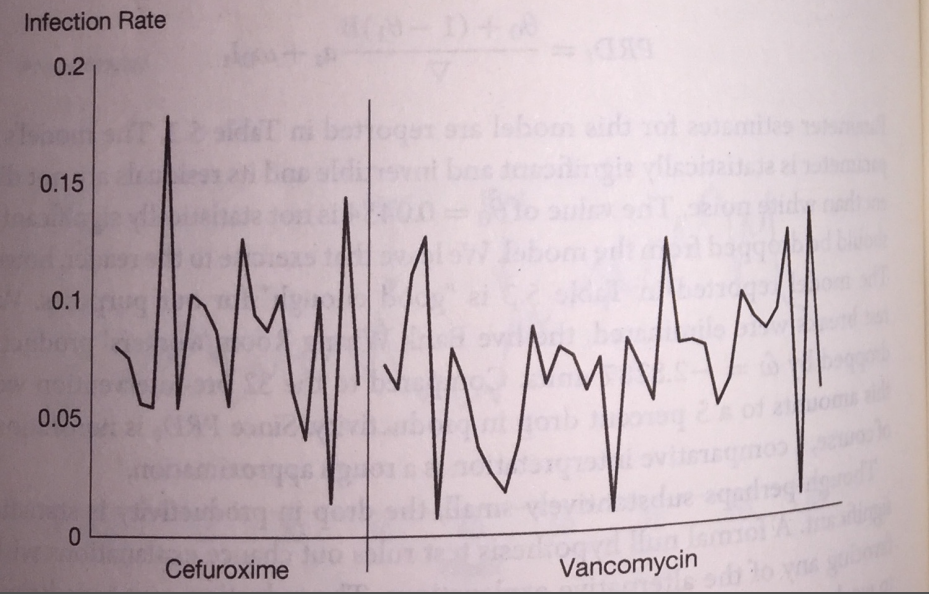
as time is the running variable,
it is not randomly assigned
(required of RD design)

Interrupted Time Series

basic idea #1: fit ARIMA + effect size

for the next $N_{post} = 35$ months, with prophylactic vancomycin. Modeling time series will be especially difficult. The inherent difficulty of identifying $_t$) from a short time series is aggravated in this instance by the unruly earance of this time series and the possibility of a large impact. We start the tive *ARIMA* modeling strategy with the simplest possible model:

$$SSI_t = \theta_0 + a_t + \omega_0 I_t$$



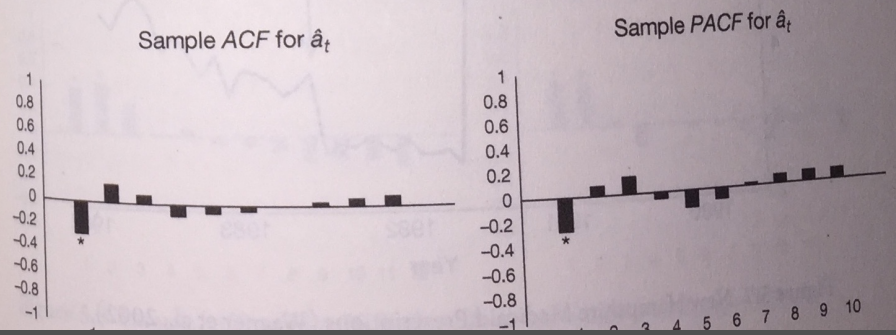
$$SSI_t = 0.084 - 0.014I_t + \frac{a_t}{1 + .27B}$$

$$= 0.084 - 0.014I_t + \sum_{k=0}^{\infty} 0.27^k a_t$$

$$= 0.084 - 0.014I_t + a_t + 0.27a_{t-1} + 0.07a_{t-2} + 0.02a_{t-3} + \cdots$$

Table 5.4. Surgical Site Infection Rates, Estimation and Diagnosis

	Estimate	SE	t	Diagnosis	
Model 1	$\hat{\theta}_0$	.0833	.0075	11.04	RSE = 0.03373
	$\hat{\omega}_0$	−.0141	.0095	−1.49	$Q_{10} = 7.4$
					$Z_{KS} = 0.805$
Model 2	$\hat{\theta}_0$	.0836	.0059	14.11	RSE = 0.03277
	$\hat{\phi}_1$	−.2710	.1312	−2.07	$Q_{10} = 3.9$
	$\hat{\omega}_0$	−.0144	.0074	−1.95	$Z_{KS} = 0.805$

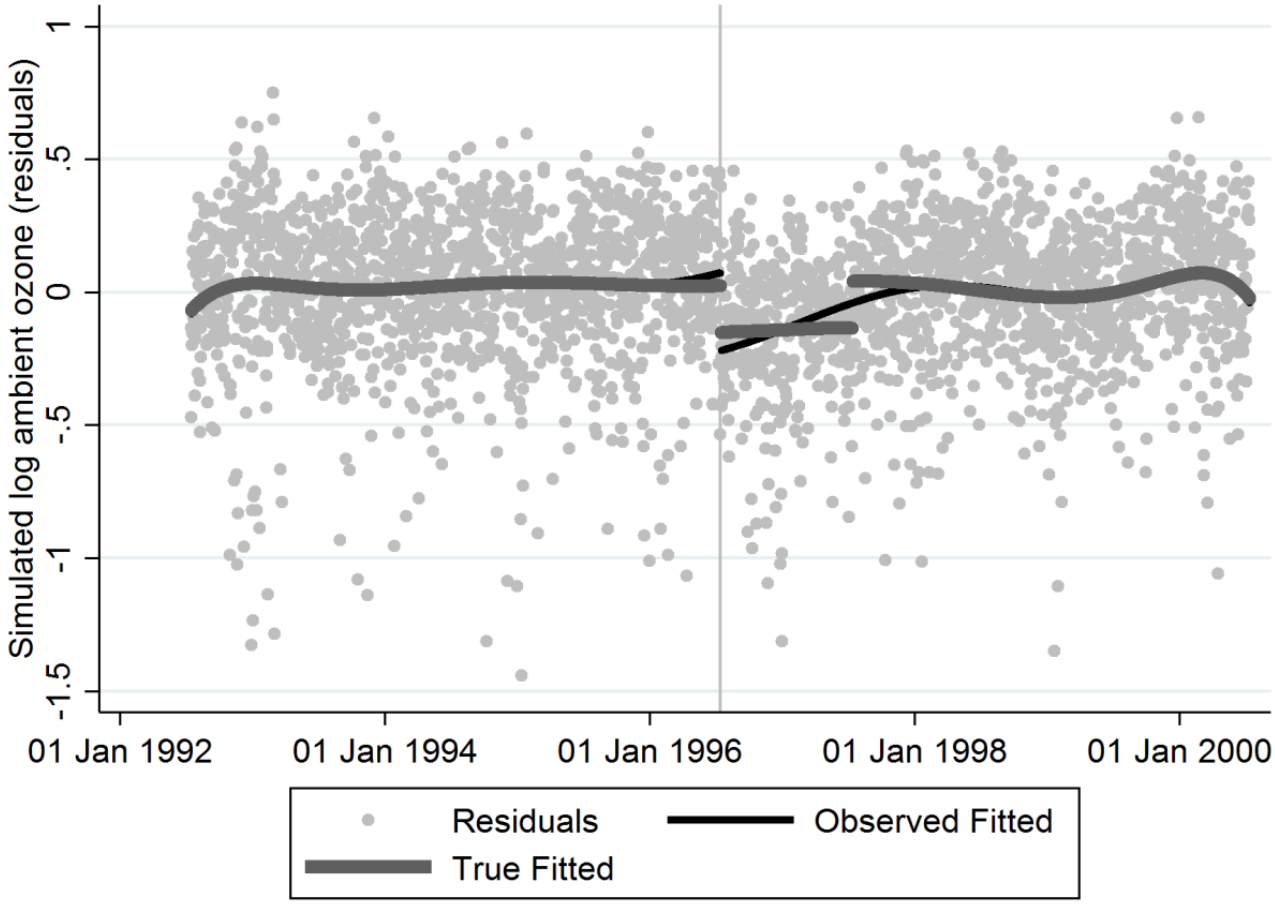


McCleary, et al
Design and Analysis of Time Series Experiments

@drewfustin | 2017.12.04

Interrupted Time Series

basic idea #2: fit curves at discontinuity



Hausman and Rapson
Regression Discontinuity in Time: Considerations for Empirical Applications

Let's Simulate Some Data

Homogeneous Poisson Point Process λ is fixed in time

- $N(0) = 0$;
- has independent increments; and
- the number of events (or points) in any interval of length t is a Poisson random variable with parameter (or mean) λt.

The last property implies:

$$E[N(t)] = \lambda t.$$

In other words, the probability of the random variable $N(t)$ being equal to n is given by:

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

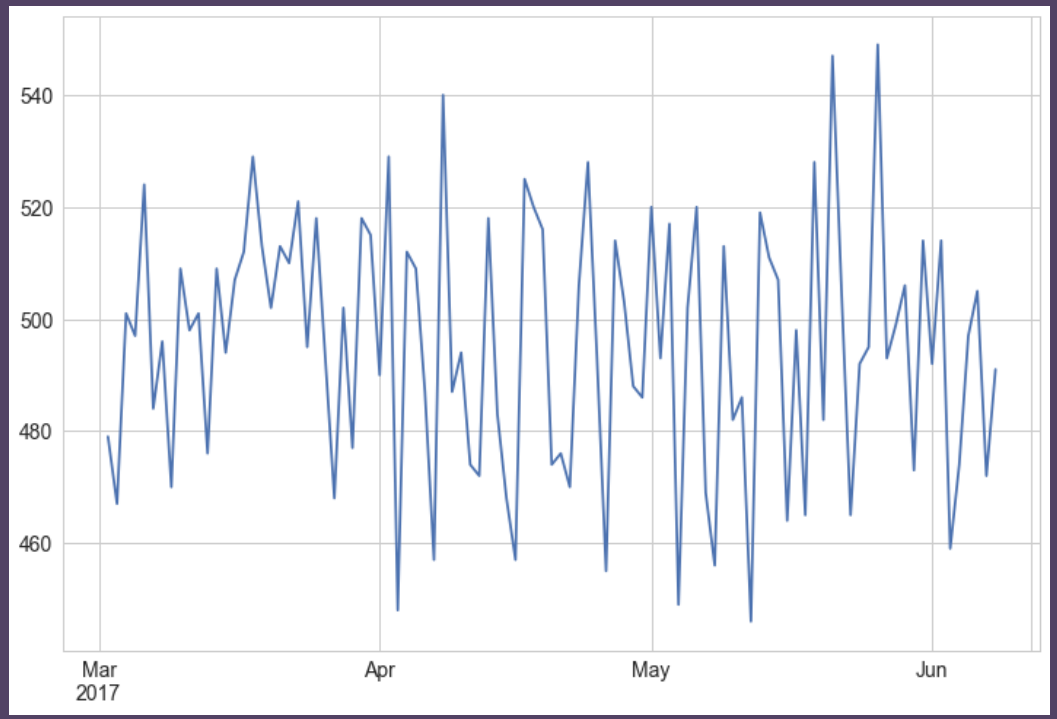
The Poisson counting process can also be defined by stating that the time differences between events of the counting process are exponential variables with mean $1/\lambda$.[52] The time differences between the events or arrivals are known as **interarrival** [53] or **interoccurence** times.[52]

Let's Simulate Some Data

Homogeneous Poisson Point Process

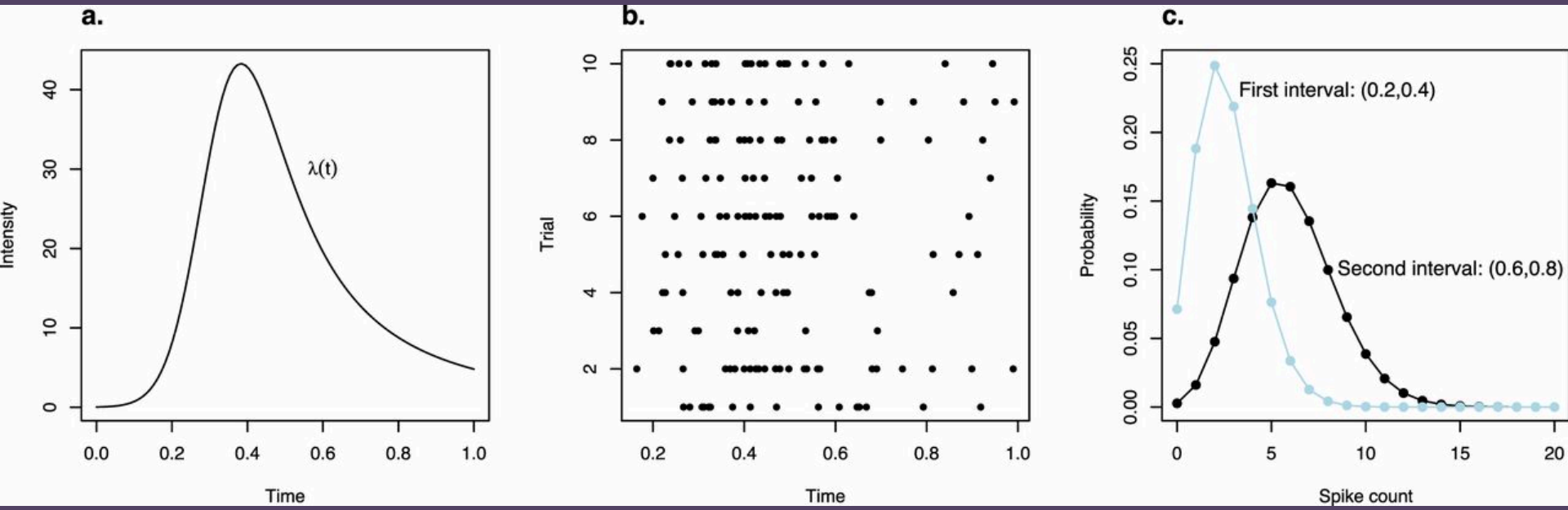
λ is fixed in time

cumulatively add events distributed exponentially with rate $1/\lambda$



Let's Simulate Some Data

Nonhomogeneous Poisson Point Process $\lambda = \lambda(t)$

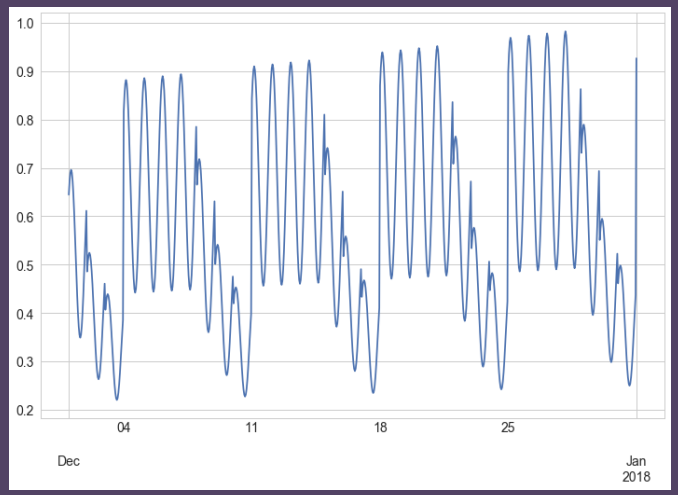


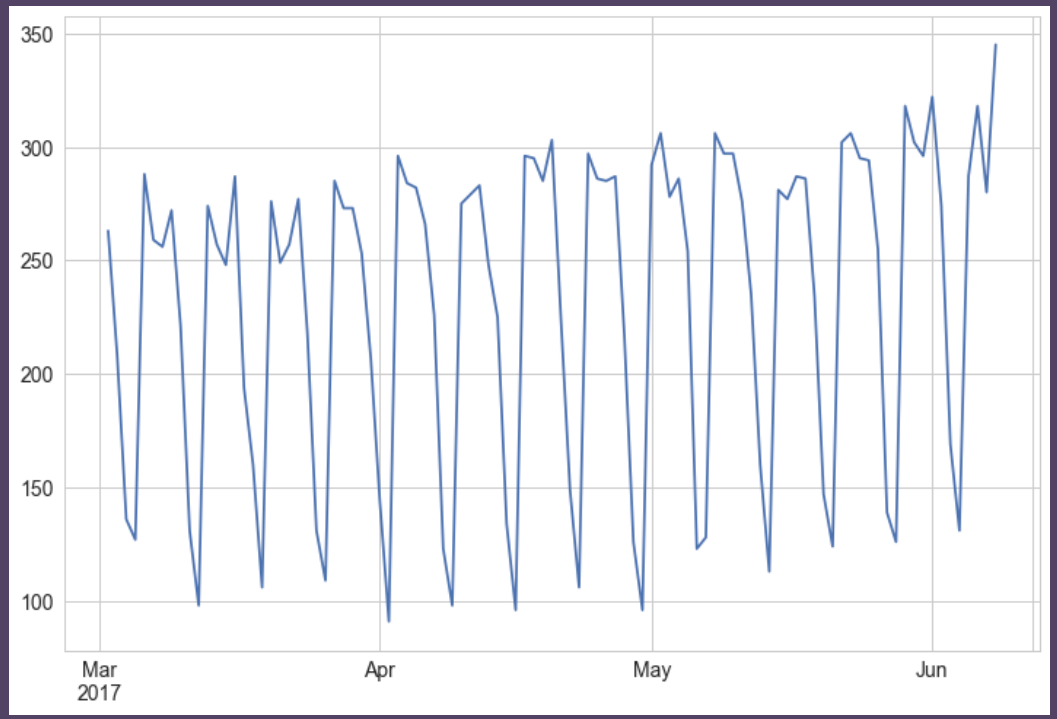
Barthelmé, et al
Modeling Fixation locations using spatial point processes

Sim HPP, cut some p=p(t)

Let's Simulate Some Data

Nonhomogeneous Poisson Point Process $\lambda = \lambda(t)$





ARIMA Forecasting

Autoregressive (AR)

Integrated (I)

Moving Average (MA)

The above can be generalized as follows.

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t.$$

This defines an ARIMA(p,d,q) process with **drift** $\delta/(1 - \Sigma\phi_i)$.

ARIMA Forecasting

ARIMA Forecasting

Autoregressive (AR)

$$\phi_p(B)=(1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)$$

Integrated (I)

$$\theta_q(B)=(1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q)$$

Moving Average (MA)

$$\nabla^d(B)=(1-B)^d$$

e.g. ARIMA (1, 1, 0):

$$(1-\phi_1 B)(1-B)=1-(1+\phi_1)B+\phi_1 B^2$$

ARIMA Forecasting

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}}\quad \underbrace{(P,D,Q)_m}_{\text{Seasonal part of the model}}$$

period. For example, an ARIMA(1,1,1)(1,1,1)$_4$ model (without a constant) is for quarterly data ($m = 4$) and can be written as

$$\underbrace{(1-\phi_1 B)}_{\substack{\text{Non-seasonal}\\ \text{AR(1)}}}\ \underbrace{(1-\Phi_1 B^4)}_{\substack{\text{Seasonal}\\ \text{AR(1)}}}\ \underbrace{(1-B)}_{\substack{\text{Non-seasonal}\\ \text{difference}}}\ \underbrace{(1-B^4)}_{\substack{\text{Seasonal}\\ \text{difference}}} y_t = \underbrace{(1+\theta_1 B)}_{\substack{\text{Non-seasonal}\\ \text{MA(1)}}}\ \underbrace{(1+\Theta_1 B^4)}_{\substack{\text{Seasonal}\\ \text{MA(1)}}} e_t.$$

ARIMA Forecasting

ARIMA Forecasting

Other Topics

- Stationarity: Adjusted Dickey-Fuller Test
- Autocorrelation Functions: ACF/PACF
  - reading *those* tea leaves
- “Memory” differences between AR, I, MA

PyMC3 for ITS Analysis

Generate NHPP
Fit ARIMA to get (p, d, q) order
Add response ~ $\mathcal{N}(\mu_{\beta}, \sigma_{\beta})$ at t_exp
Use ARIMA ops to iteratively solve y, ε

PyMC3 for ITS Analysis

Generate NHPP
Fit ARIMA to get (p, d, q) order
Add response ~ $\mathcal{N}(\mu_{\beta}, \sigma_{\beta})$ at t_exp
Use ARIMA ops to iteratively solve y, ε

Potential next step:

- Use PyMC3 to check regression discontinuity in across t_exp bound