

Detecting Anomalies

D. A. Dickey

SAMSI and NCSU



Statistical Issues in Homeland Security

- Discriminants: Classical Methods

Involve measured inputs

$X_1 = \#$ of key words

$X_2 =$ time sent

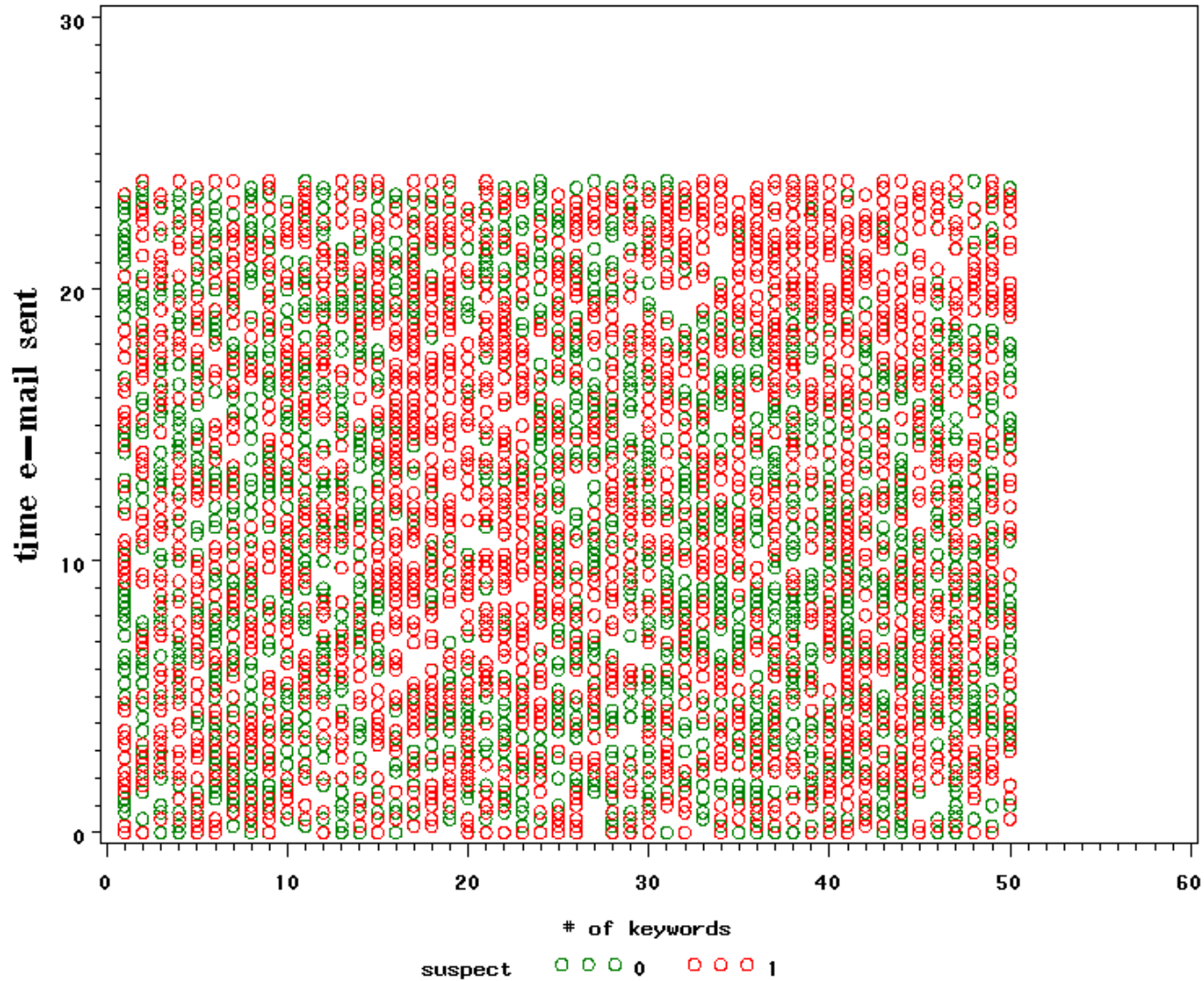
Have test data (known category)

$Y=1$, message is relevant

$Y=0$, message is ignorable

p =proportion of 1's at some (X_1, X_2) setting

Overall, about 25% are “of interest” (red)



Recursive Splitting

Table of key by suspect

key	suspect		
Frequency			
Expected			
Percent			
Row Pct			
Col Pct	0,	1,	Total

<8.5 keywds	1580	404	1984
	1482.9	501.12	
	13.14	3.36	16.51
	79.64	20.36	
	17.59	13.31	

>8.5 keywds	7404	2632	10036
	7501.1	2534.9	
	61.60	21.90	83.49
	73.77	26.23	
	82.41	86.69	

Total	8984	3036	12020
	74.74	25.26	100.00

$\chi^2 = \sum[(\text{obs}-\text{exp})^2/\text{exp}] = 3.1$
 significant discriminator

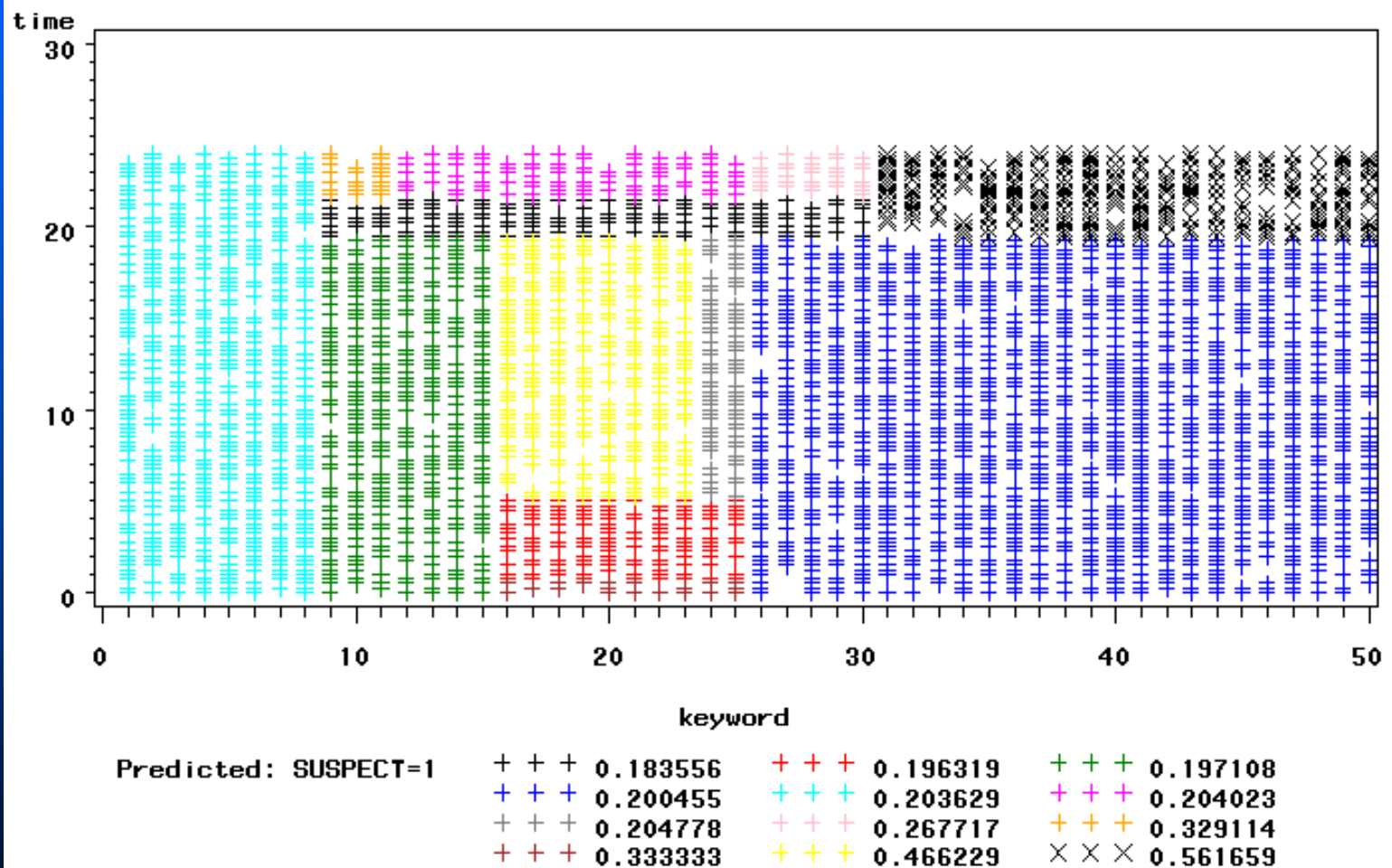
Statistics for Table of key by suspect

Statistic	DF	Value	Prob
Chi-Square	1	30.1598	<.0001

Detection

- Can I get bigger χ^2 with split at 12 or 15 keywords? (no)
- Is time of day better than keywords? (no)
- Data set 1 keywords < 8.5 split this
- Data set 2 keywords > 8.5 split this
- Keep splitting until _____(?)
- Subset described by keyword and time range has 54% suspect. Probability high → investigate future e-mails in this range.

E-mail hot spots



Statistical Issues in Homeland Security

■ Time Series

Intervention Analysis

Has level changed?

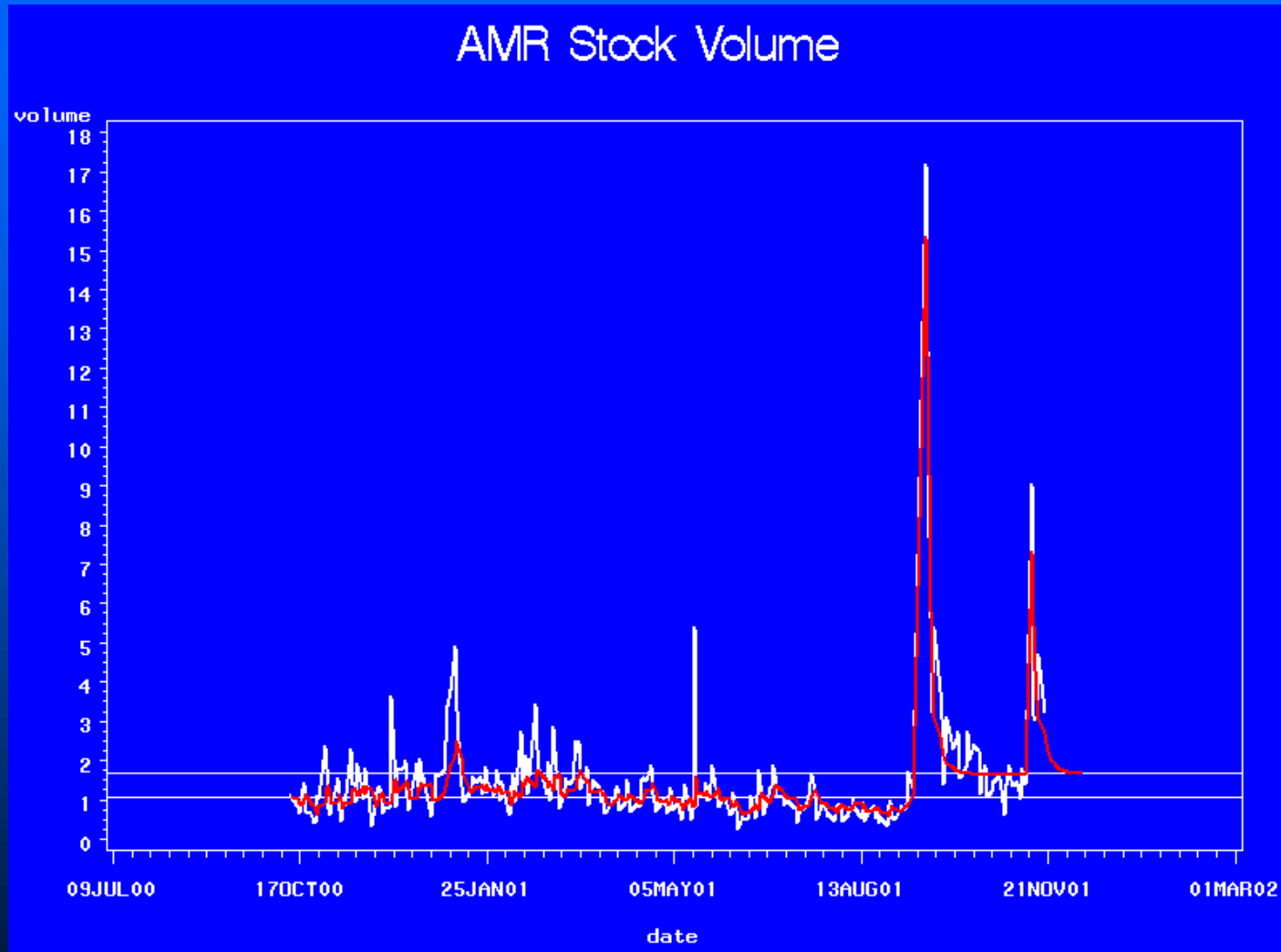
prescriptions/week

threatening calls

\$\$ in suspect account

Duration of effects

Effect of 9/11 on economy

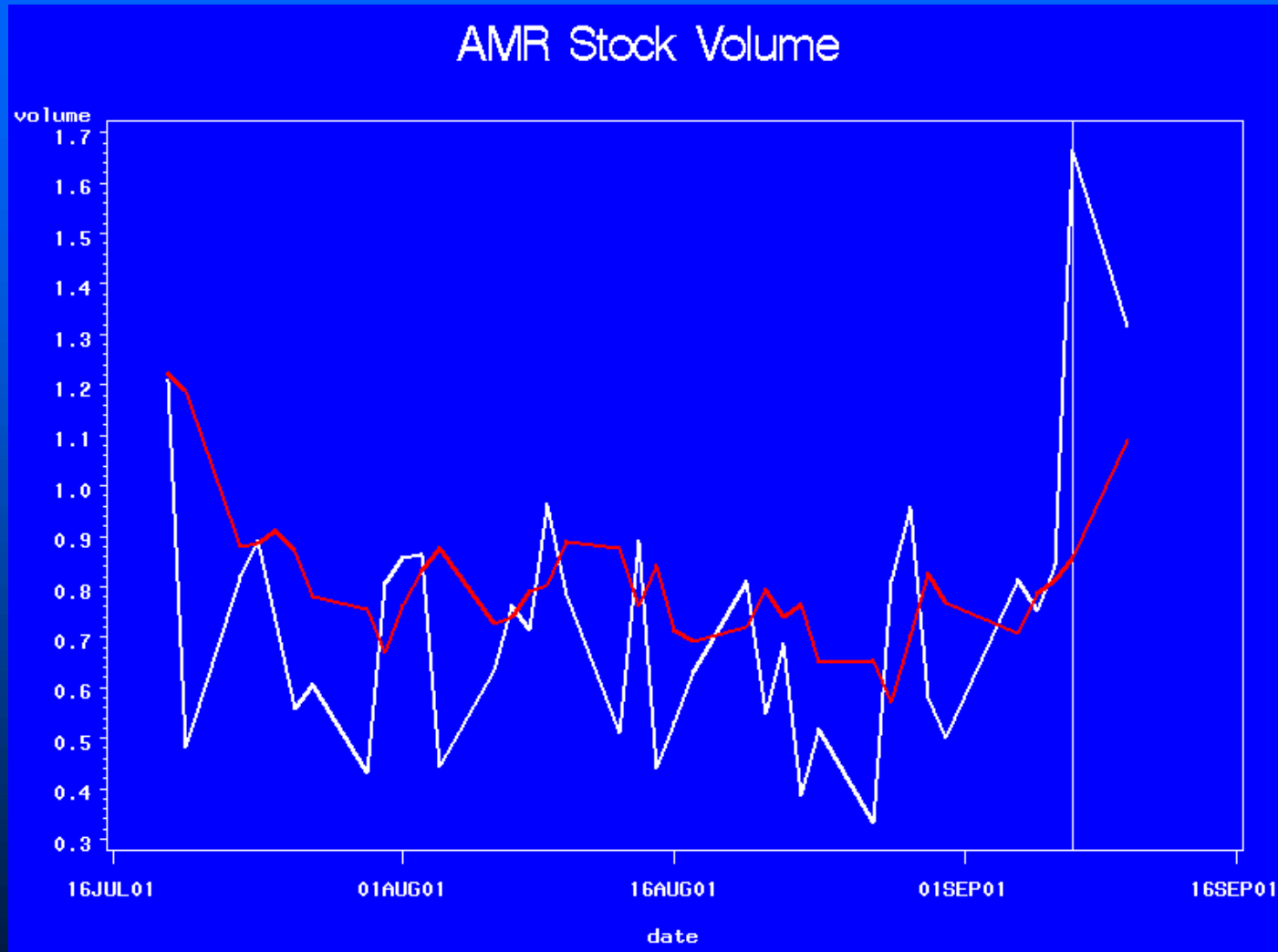


(Volume in millions)

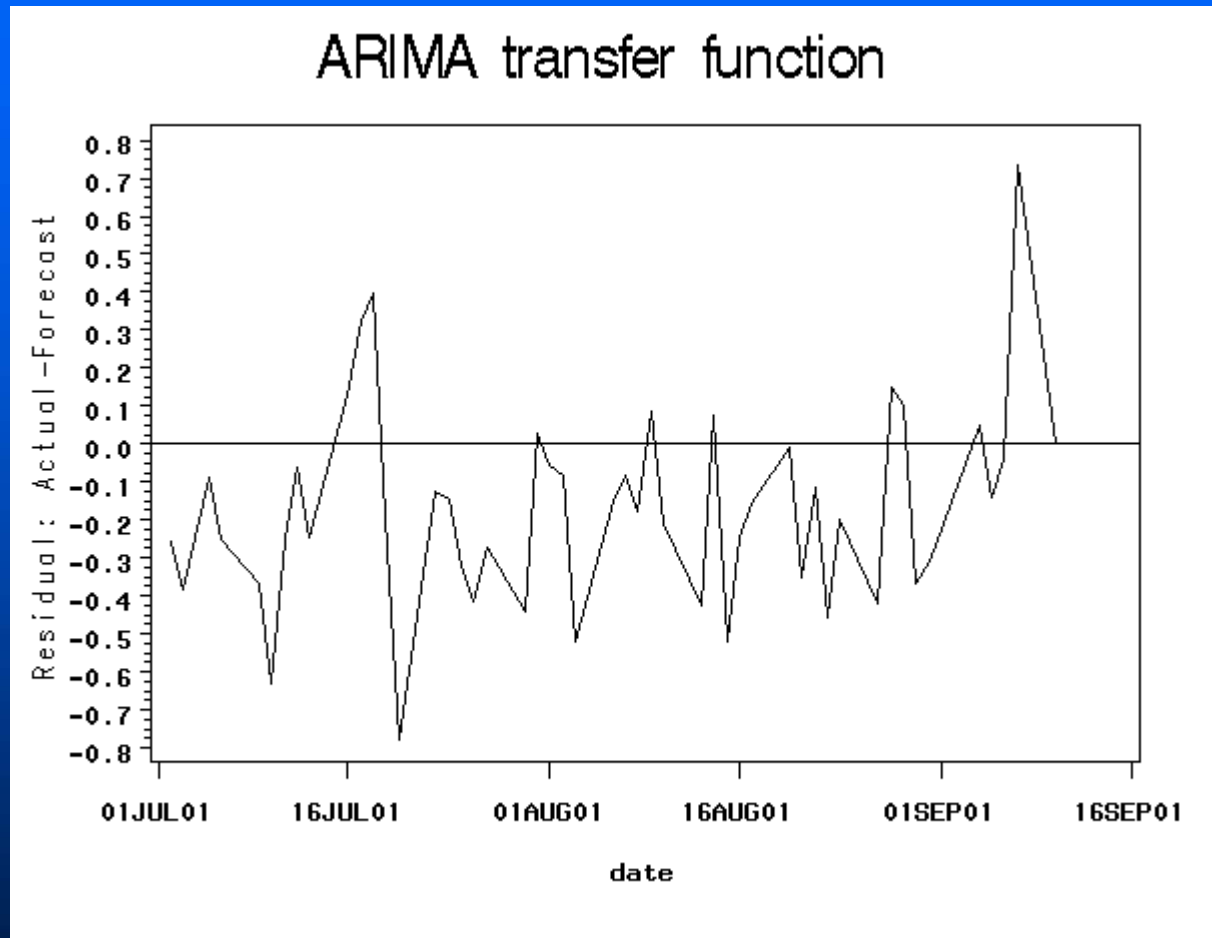
- Red line is prediction
- Idea – relate to past
- $Y_t - \mu = \rho (Y_{t-1} - \mu) + e_t$
- Today's deviation = proportion of yesterday's plus “shock” (or surprise)
- $\mu=200$, $\rho=.8$, today we see 250
- Tomorrow, predict _____
- Two days ahead, predict _____

Now add “Intervention”

- $X = 0, 0, 0, 1, 0, 0, 0\dots$
- $Y_t - 100 = .5(Y_{t-1} - 100) + 8 X_t + e_t$
- Expected Y's: 100, 100, 100, 108, 104, 102,
- For 9/11 data (volume in thousands)
 - » $Y_t - 1.6630 = .76(Y_{t-1} - 1.6630) + 15.46718 X_t + e_t$
- Half-life $(0.76)^j = 0.50 \rightarrow j = 2.5$ days

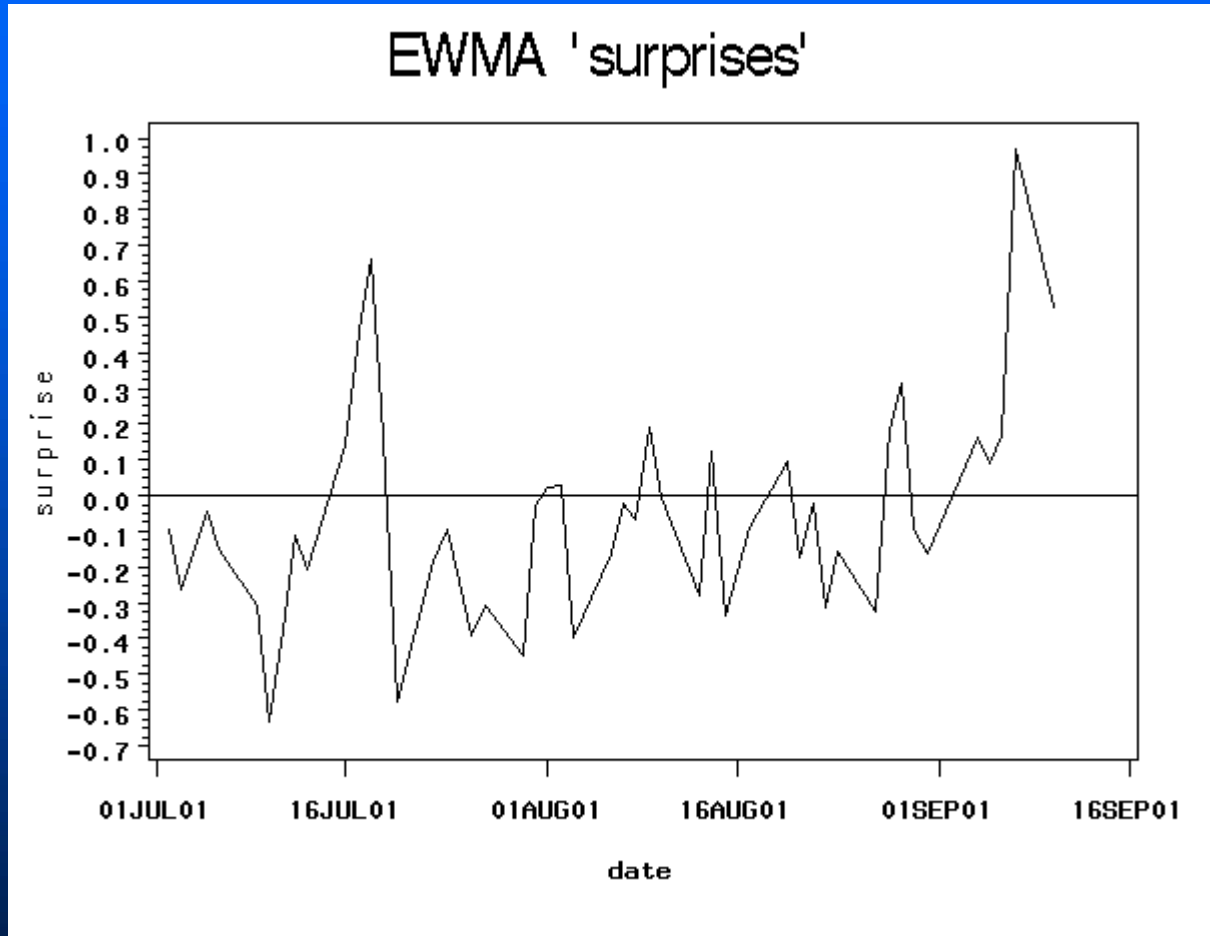


Forecast and Observed



Deviations from forecast

- Lots of work & time for model
- Easier way?
- Exponential smoothing (EWMA)
- Predict Y_t as weighted sum of past Y
- Use exponentially declining weights
- Example:
 - » Predict Y_t as $(0.1)(Y_{t-1} + .9Y_{t-2} + .81Y_{t-3} + \dots)$
- Can estimate weight!
- Weights sum to 1.



Deviations from EWMA forecast

- Not every signal is sudden pulse
- How about slower change?
- Next example – detecting trend change

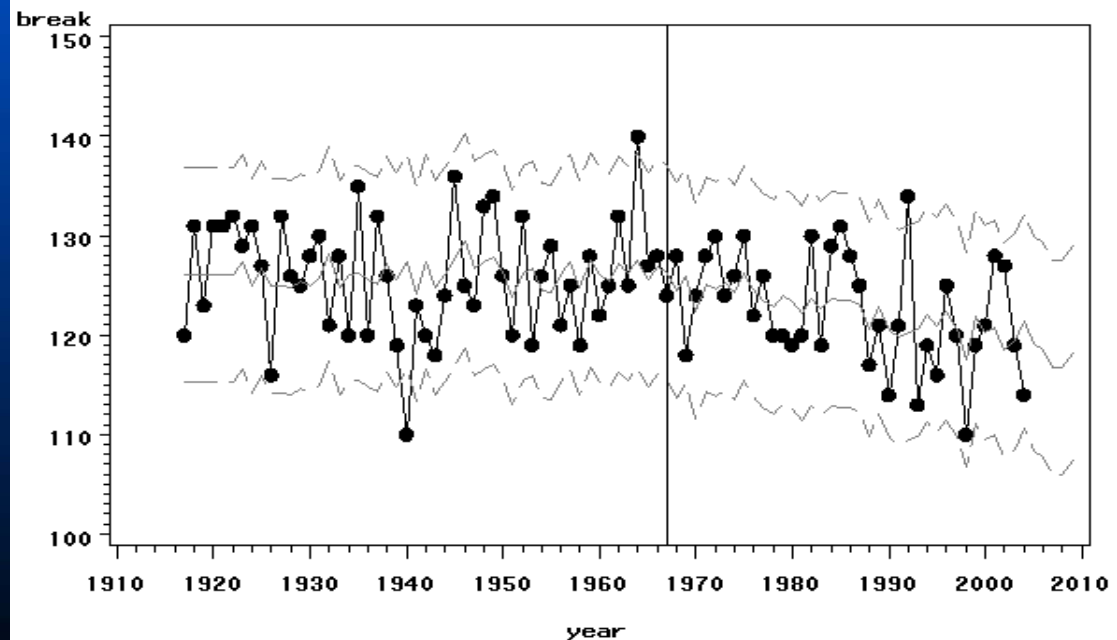


Example 2: “Nenana Ice Classic” Trend or no trend?

Start = 1917

pot is now \$285,000

Tanana River Ice Breakup



- First idea – pretend we know year of potential change
- $X = 0, 0, 0, 1, 2, 3, 4, 5$
- $Y_t = 125 - 0.5 X_t$
- $Y_t = 125, 125, 125, 124.5, 124, 123.5, 123, \dots$
- Is this significant?
- Probability of -0.5 or less if slope is really 0 in the long run (this is the “P-value”)

Add autocorrelated error to model (AR1,1)

X variable called “ramp”

Set break at 1967

Use SAS, P-value for ramp <0.0001 !!!

The ARIMA Procedure

Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Variable
Shift					
MU	126.01962	0.59155	213.03	<.0001	break
AR1,1	-0.21784	0.10929	-1.99	0.0494	break
NUM1	-0.18686	0.04326	-4.32	<.0001	ramp

Not fair! Year not known

- (1) Try lots of years
- (2) Use nonlinear regression
 - Gives estimate of join point
 - Gives confidence intervals (likely regions) for results
 - Model $Y = A + B (X - C)$ if $X > C$, $Y = A$ if $X < C$.
 - Minimize sum of squared errors $(\text{obs} - \text{pred})^2$

Iter	C	a	b	Sum of Squares
0	1960.0	126.0	-0.2000	2850.8
	(more iterates)			
5	1967.6	126.0	-0.1873	2702.1

NOTE: Convergence criterion met.

Output

■ Source	DF	Sum of Squares	Mean Square	Approx F Value	Pr > F
■ Model	2	403.5	201.7	6.35	0.0027
■ Error	85	2702.1	31.7894		
■ Corrected Total	87	3105.6			

■ Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
■ C	1967.6	10.7354	1946.2	1988.9
■ a	126.0	0.7895	124.4	127.6
■ b	-0.1873	0.0868	-0.3599	-0.0147

