Automatic Forecasting Systems

Autobox 5.0 is the recipient of the **best** dedicated forecasting package in J. Scott Armstrong's book titled "Principles of Forecasting" (p. 671)

P.O. Box 563
Hatboro, PA 19040
Tel: (215) 675-0652
Fax: (215) 672-2534

sales@Autobox.com

www.Autobox.com

Since 1976

Forecasting History

Forecasting History is Always Easier Than Forecasting The Future

"The Americans have need of the telephone, but we do not. We have plenty of messenger boys."

-- Sir William Preece, chief engineer of the British Post Office, in 1876.

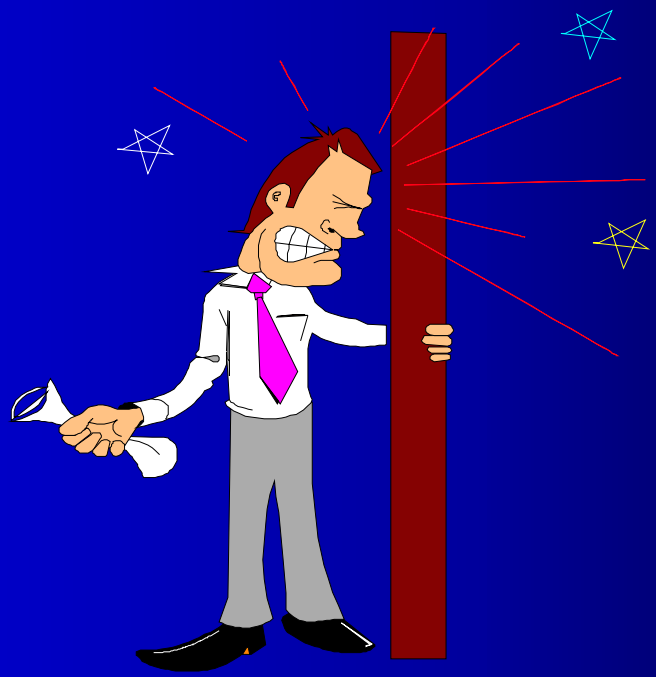
Abraham Lincoln (1809 - 1865) said:

"If we could first know where we are, then whither we are tending, we could then decide what to do and how to do it."

Why Information??

- To understand the nature of the system generating the data.
- To derive optimal forecasts.
- To understand the dynamic effects between a cause and the resulting effect.
- *To derive optimal control policies.*

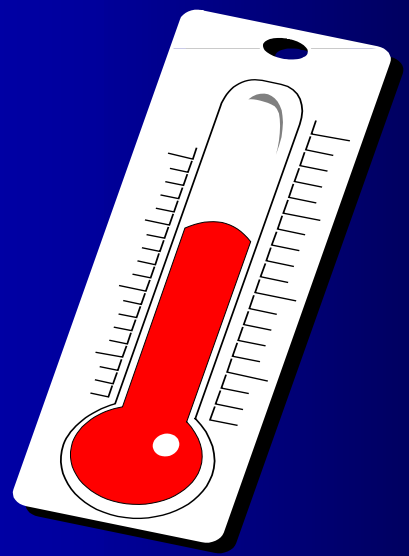
If only we had known sooner….



TYPICAL BANK CUSTOMER							
	Checks	Deposited	ACH	ACH	ACH	Coll.	Activity
Period	Paid	Items	Debits	Credits	Rec'd	Balance	Index
9604	206	115		0	40	43,600	
9605	195	161	0	0	43	39,761	
9606	180	151	0	1	42	35,463	
9607	219	141	247	1	40	32,676	
9608	183	47	474	1	43	20,639	
9609	201	109	0	0	43	16,939	47.19
9610	202	102	0	0	47	20,247	39.40
9611	193	107	233	0	50	19,077	35.99
9612	204	117	0	0	51	24,082	35.50
9701	200	161	282	0	54	34,630	38.56
9702	194	130	271	7	44	30,144	39.59
9703	206	131	268	9	48	24,222	41.57
9810	196	106	335	1	53	79,870	40.98
9811	210	136	352	1	48	59,806	41.32
9812	237	166	349	1	71	60,517	37.38
9901	207	158	354	1	58	68,254	33.86
9902	195	119	364	1	58	83,290	32.48
9903	244	145	348	1	67	98,463	34.05

Each period's value for each activity is standardized by dividing it by its standard deviation.

The *Activity Index* is the sum of each standardized activity value for each period.



Account activity is recorded sequentially through time (daily).

Data recorded sequentially through time is called “Time Series Data”.

The analysis of time series data must use special statistical techniques, called “Time Series Techniques”.

"Time Series Techniques" differ from more commonly known methods:

Regression analysis or 'Ordinary Least Squares Regression' - designed for cross-section data, not time series data.

Providing correct information requires time series techniques - AUTOBOX.

Serious Disconnect between the Teaching and Practice of Statistics

- 99.9% of all Academic presentation of statistical tools REQUIRES independent observations
- In time series data, this is clearly not the case

When analyzing time series data you are concerned with two items:

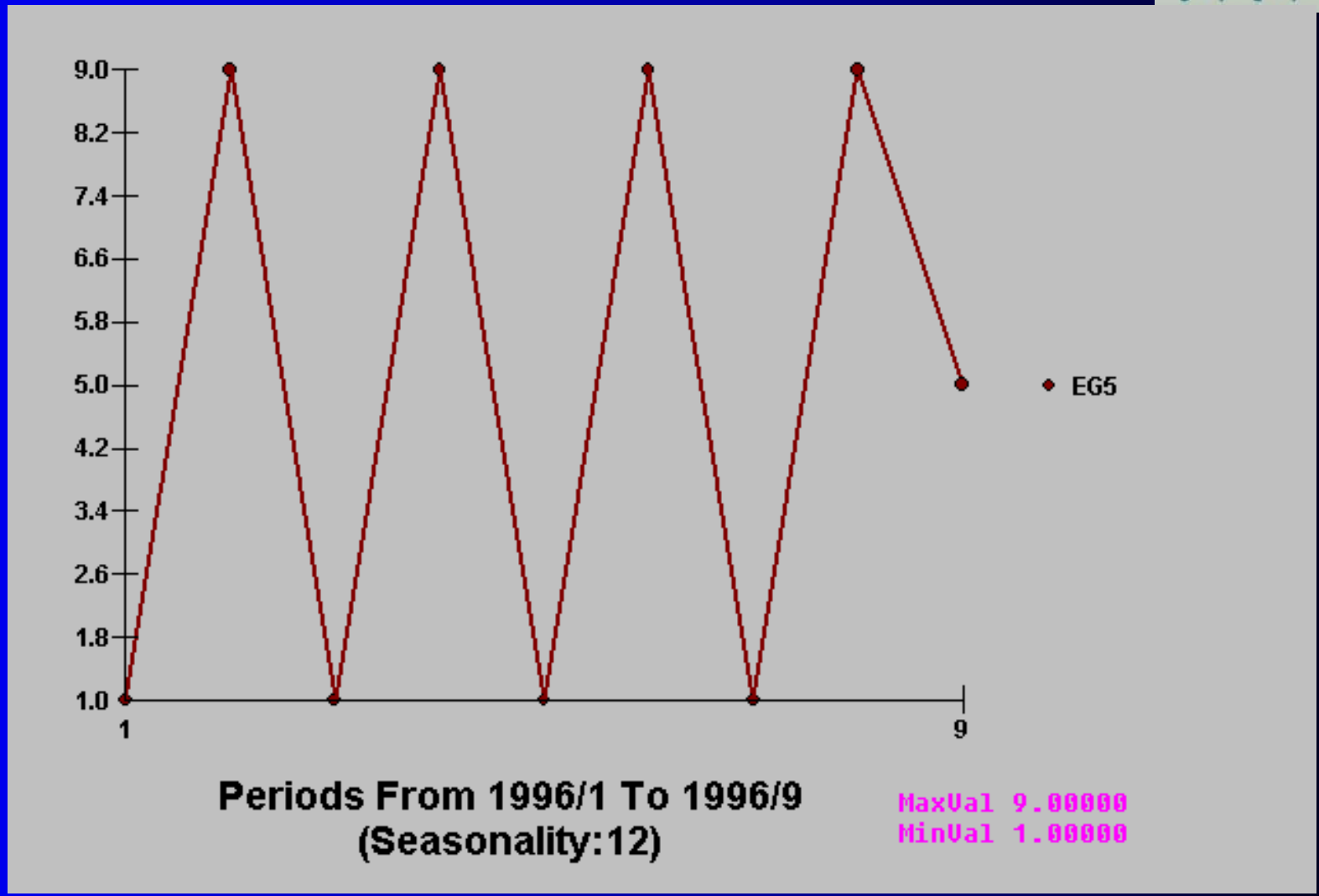
- Early Warning Systems detecting peculiar data
- Forecasting expected Demand for Cash or Bad Debts or Overdraft Revenue; Call Center forecasting for Workforce Management;Marketing/Promotion Effectiveness

Early Warning Systems

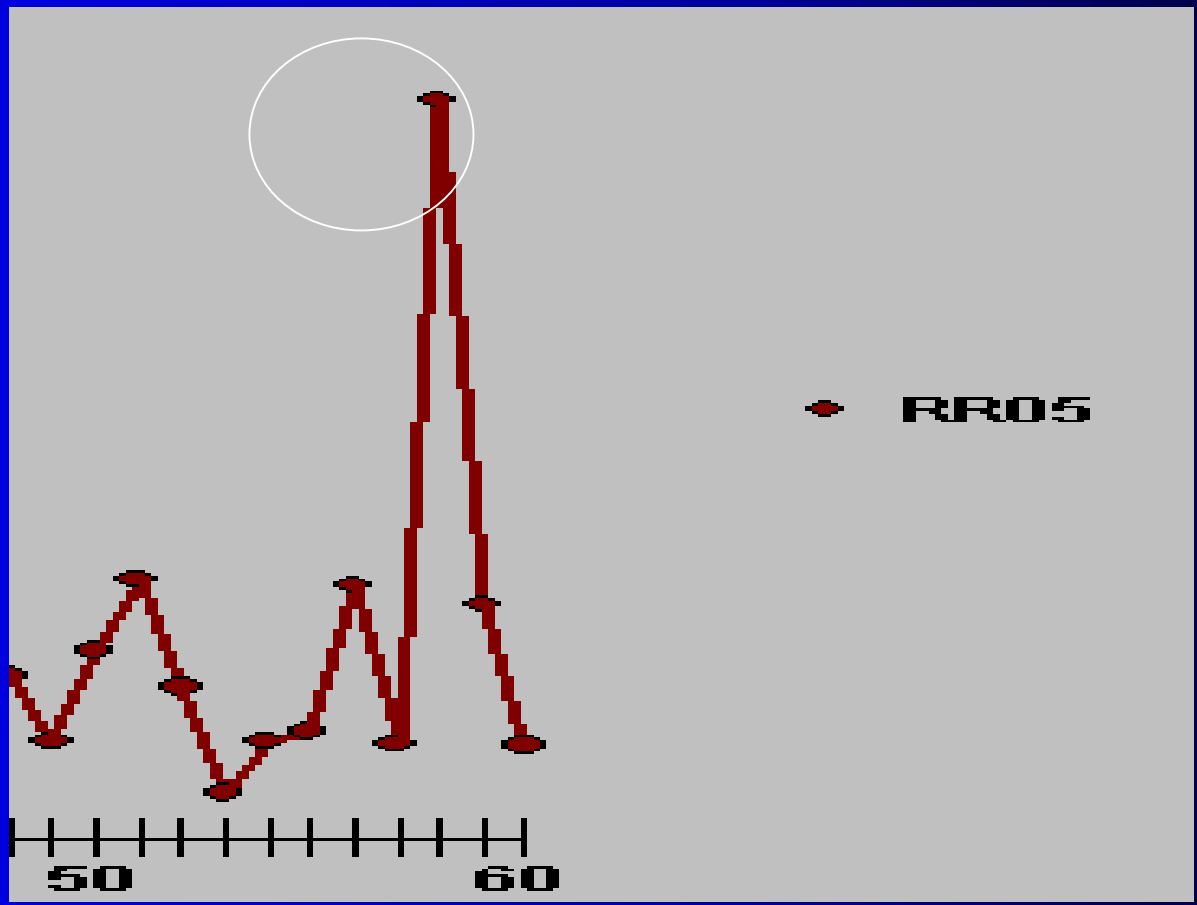
Find out why

- Early warning systems should not simply detect high and low values, but should detect unusual activity based upon history. Find out why
- For example the number of checks or average balance each month could be used as a barometer of customer behavior.

The mean can be unusual

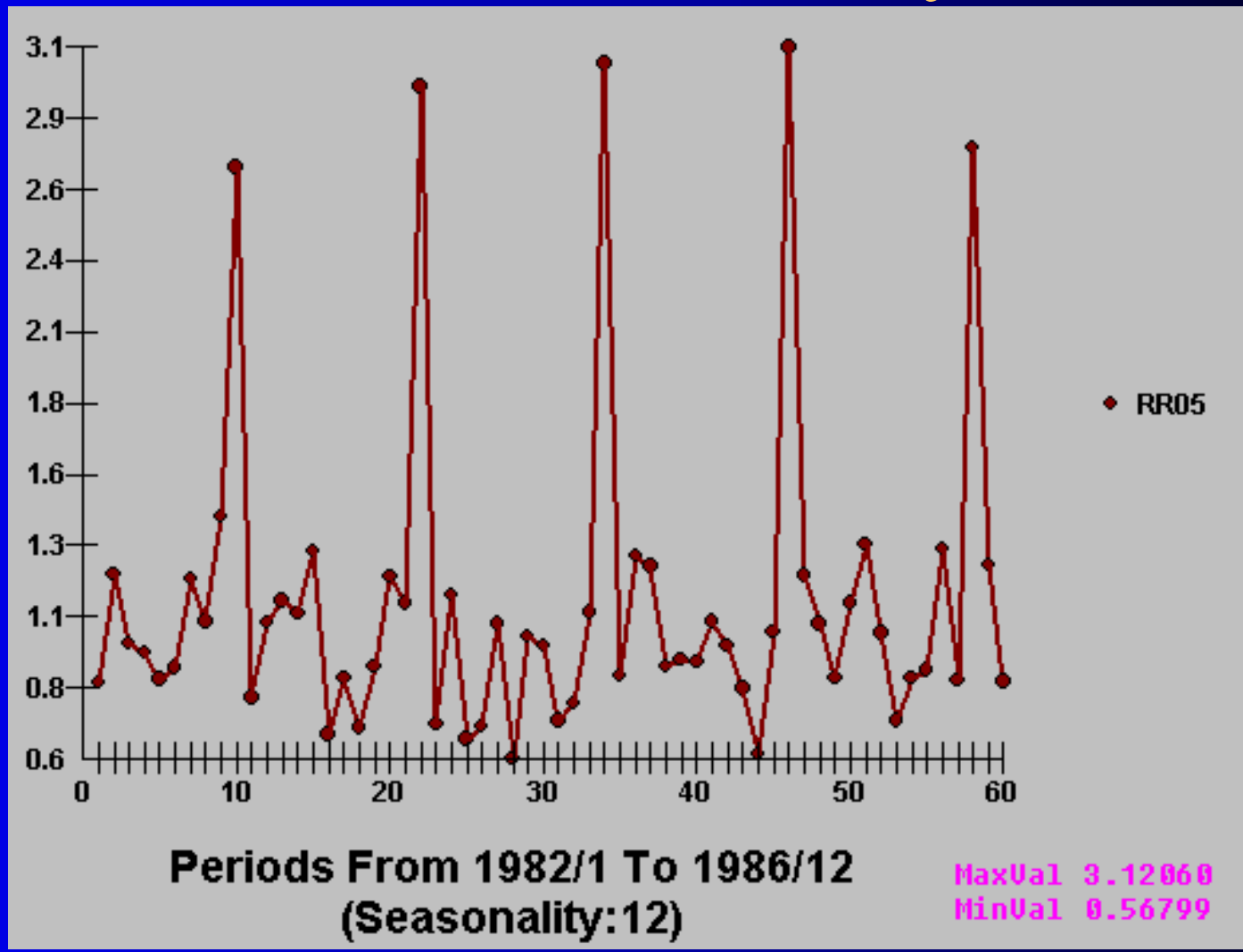


This value is not unusual



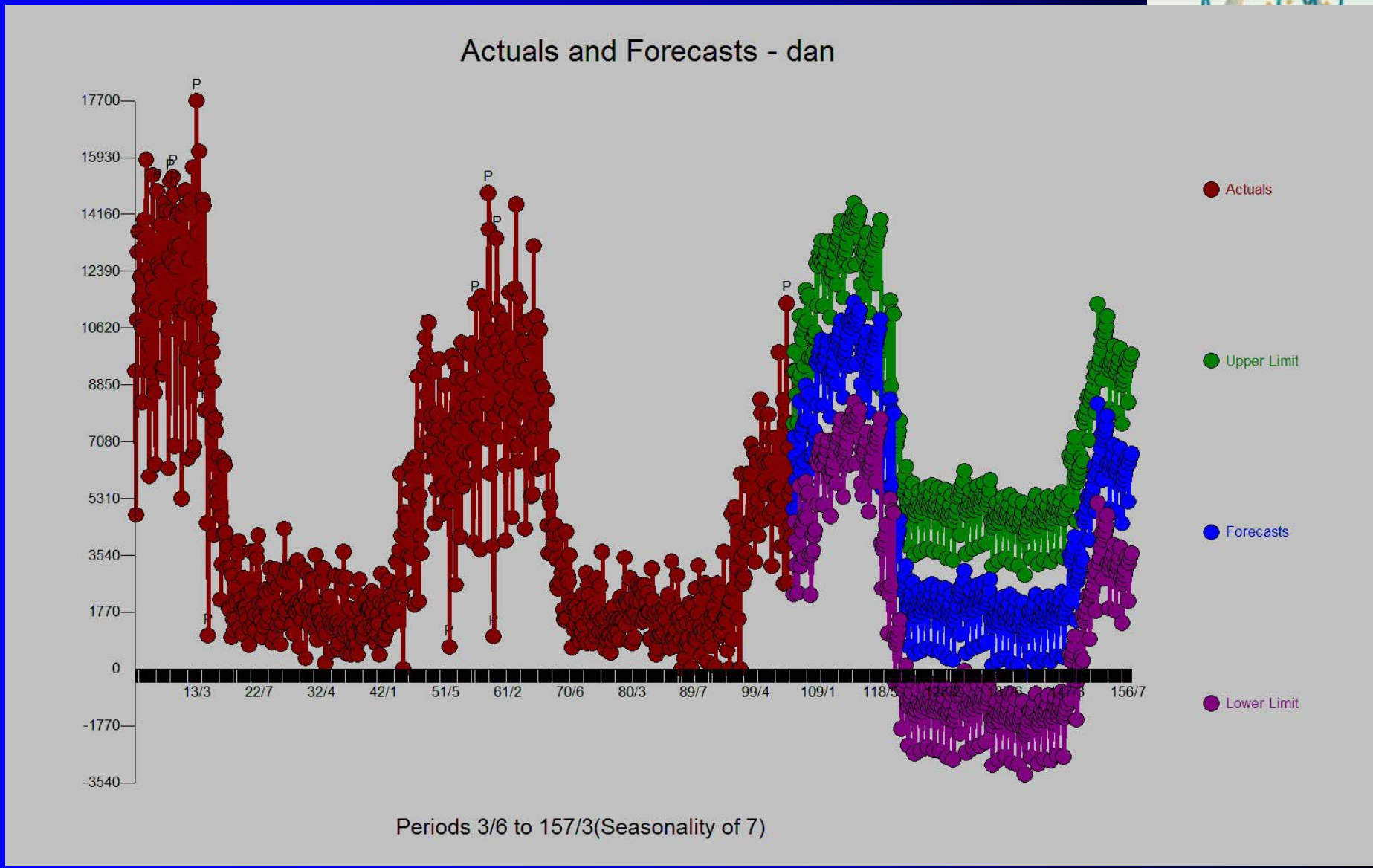
How can this value not be unusual?

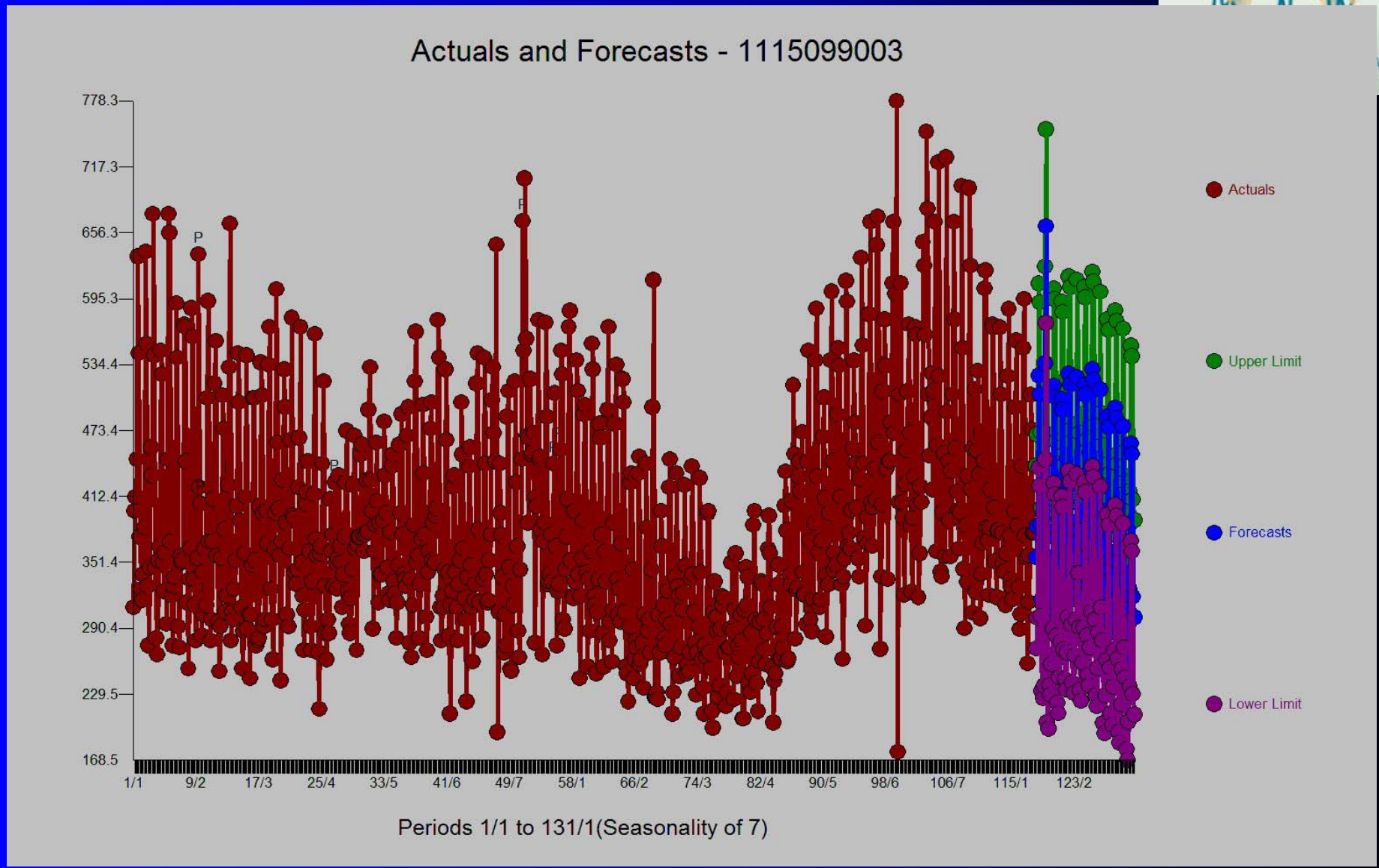
When it's part of a pattern across the history

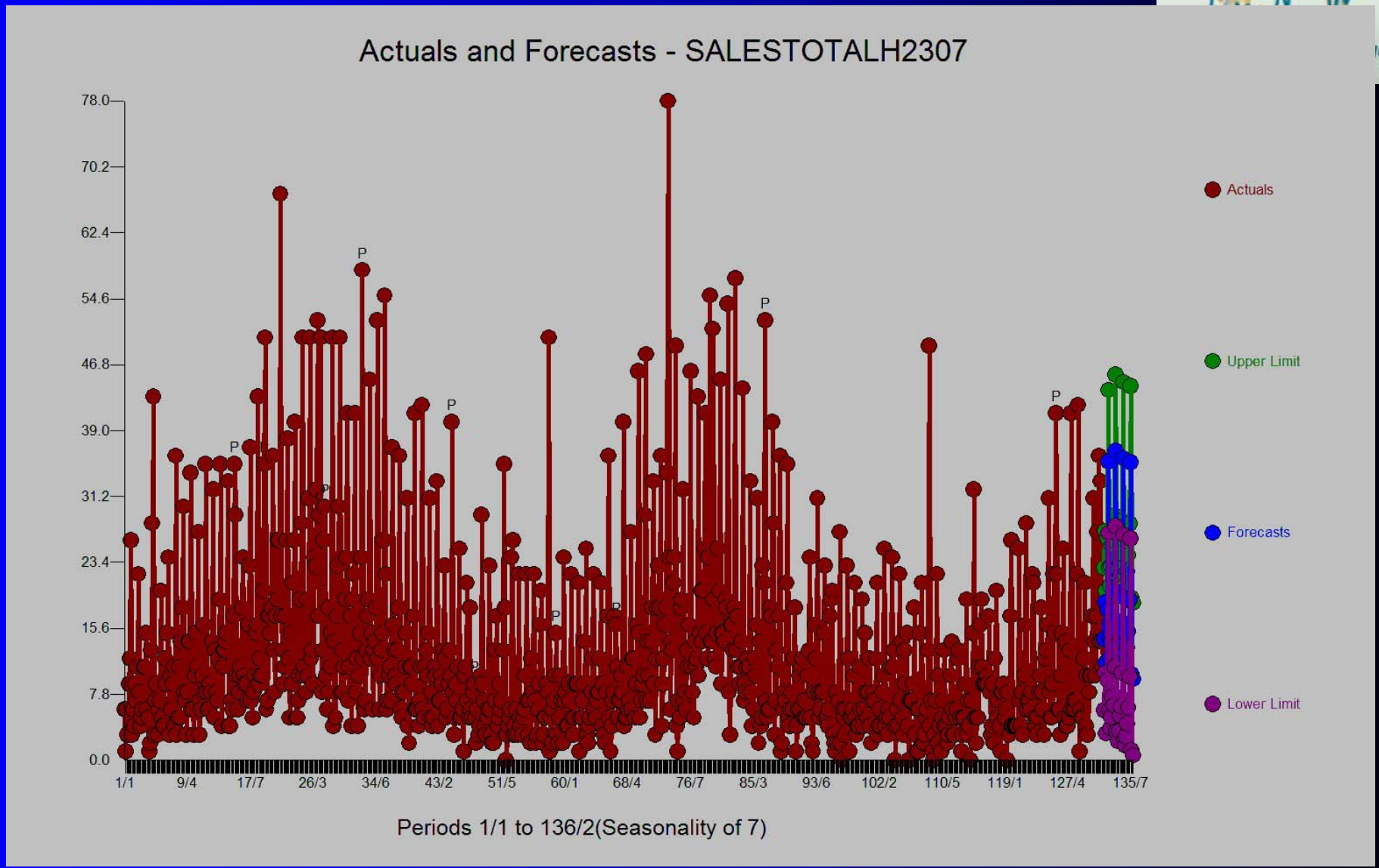


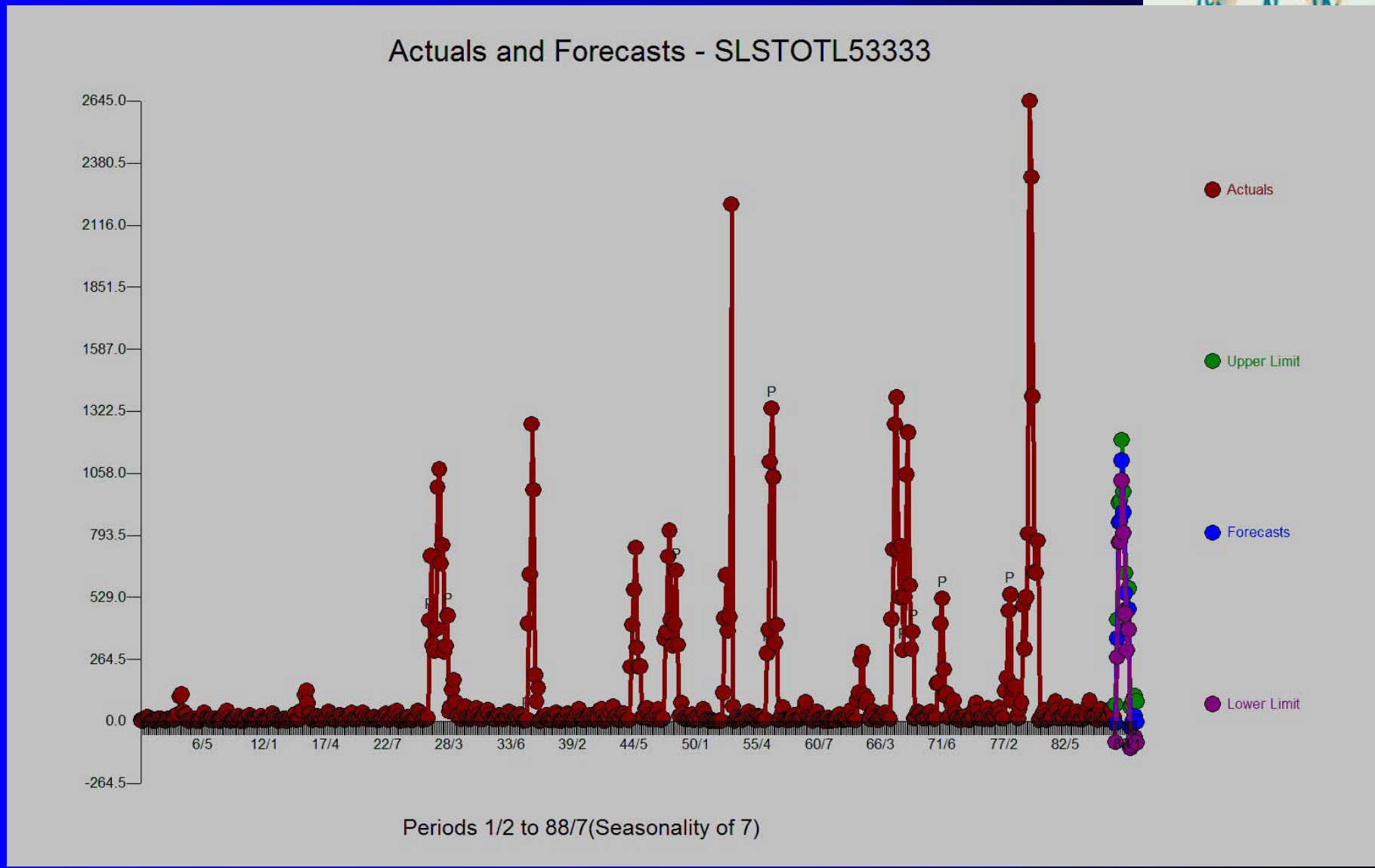
. . .

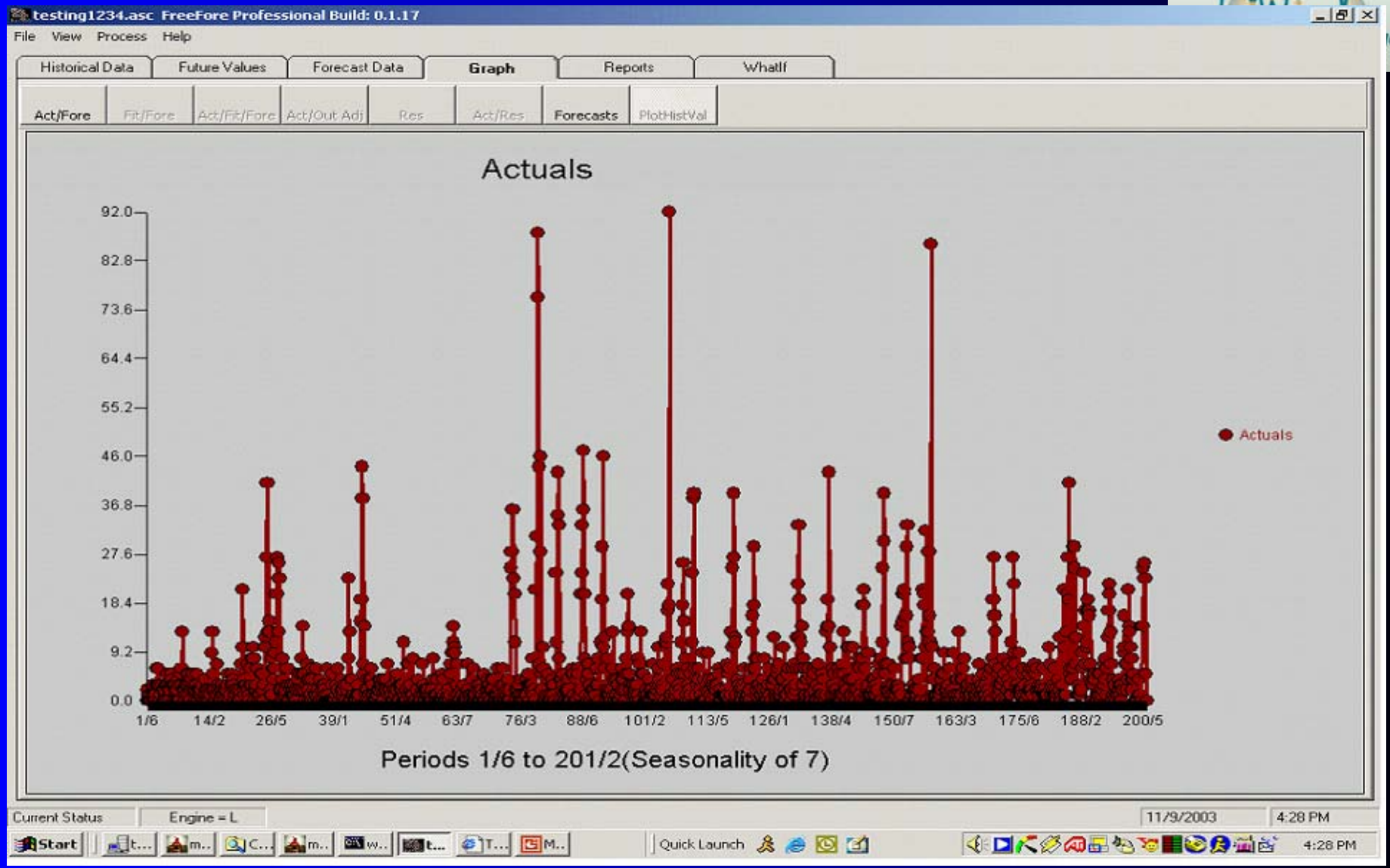
Some Interesting Datasets

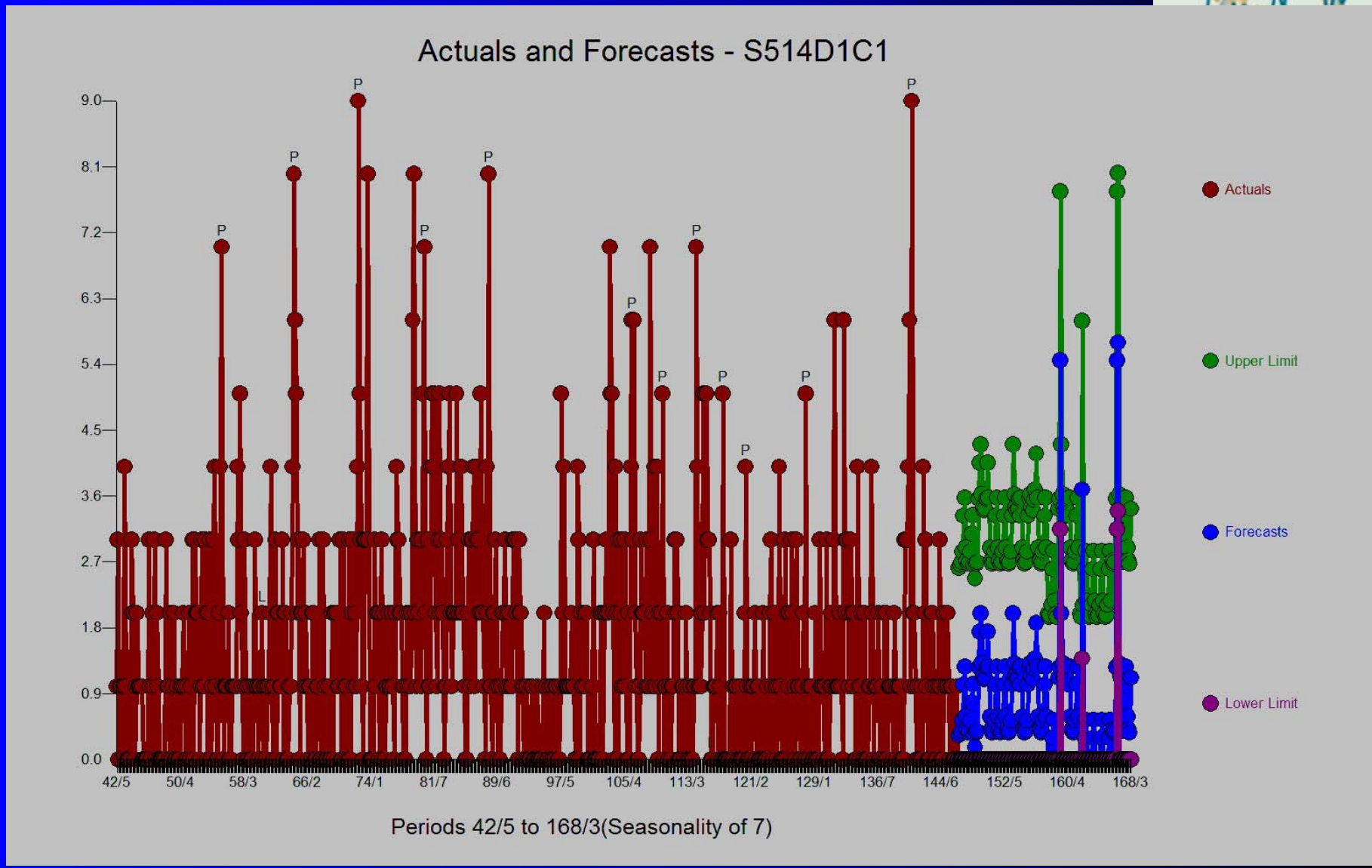

Actuals and Forecasts - dan
17700
15930
14160
12390
10620
8850
7080
5310
3540
1770
0
-1770
-3540
13/3
22/7
32/4
42/1
51/5
61/2
70/6
80/3
89/7
99/4
109/1
156/7
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 3/6 to 157/3(Seasonality of 7)

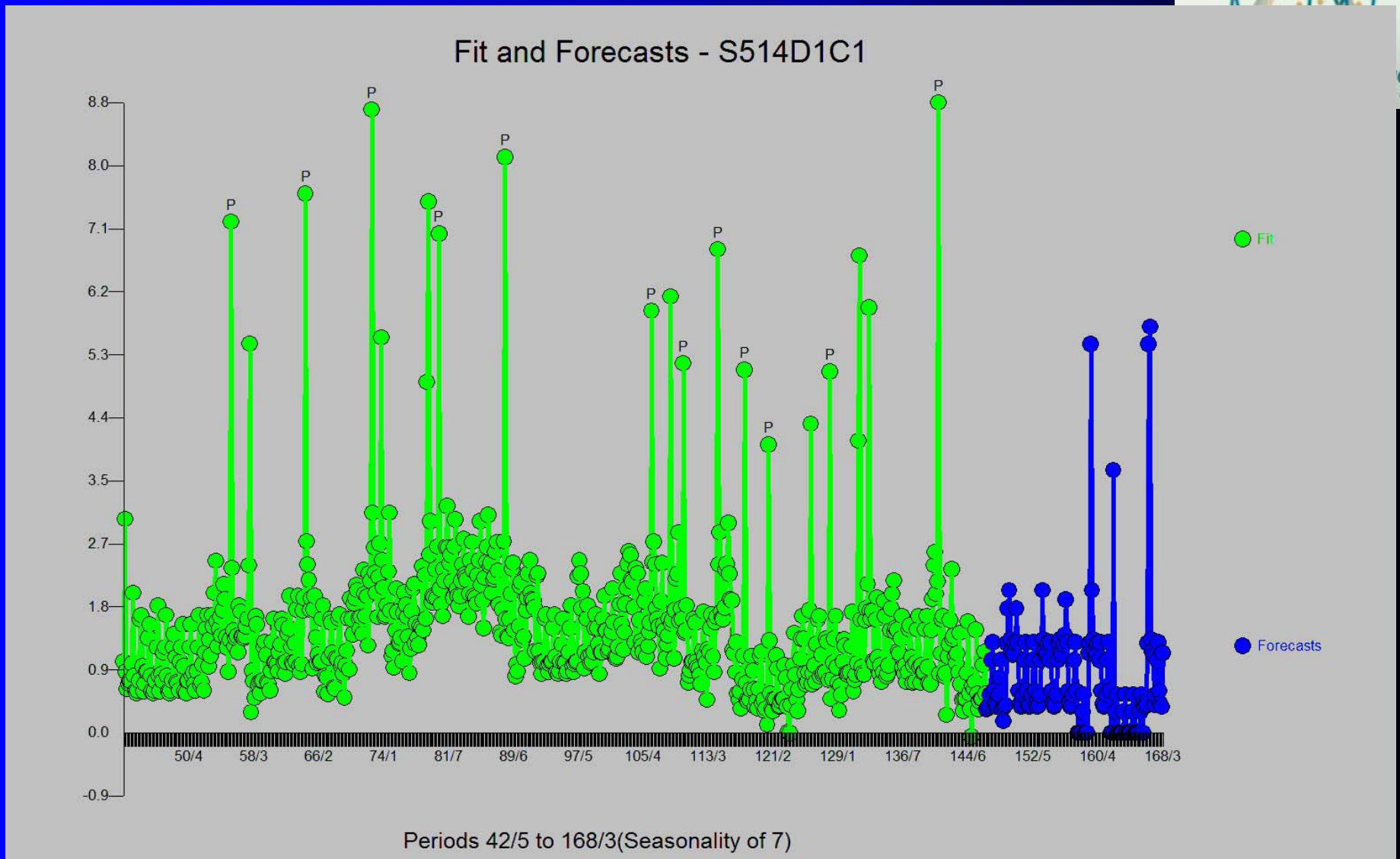

Actuals and Forecasts - 1115099003
778.3
717.3
656.3
595.3
534.4
473.4
412.4
351.4
290.4
229.5
168.5
1/1
9/2
17/3
25/4
33/5
41/6
49/7
58/1
66/2
74/3
82/4
90/5
98/6
106/7
115/1
123/2
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 1/1 to 131/1(Seasonality of 7)

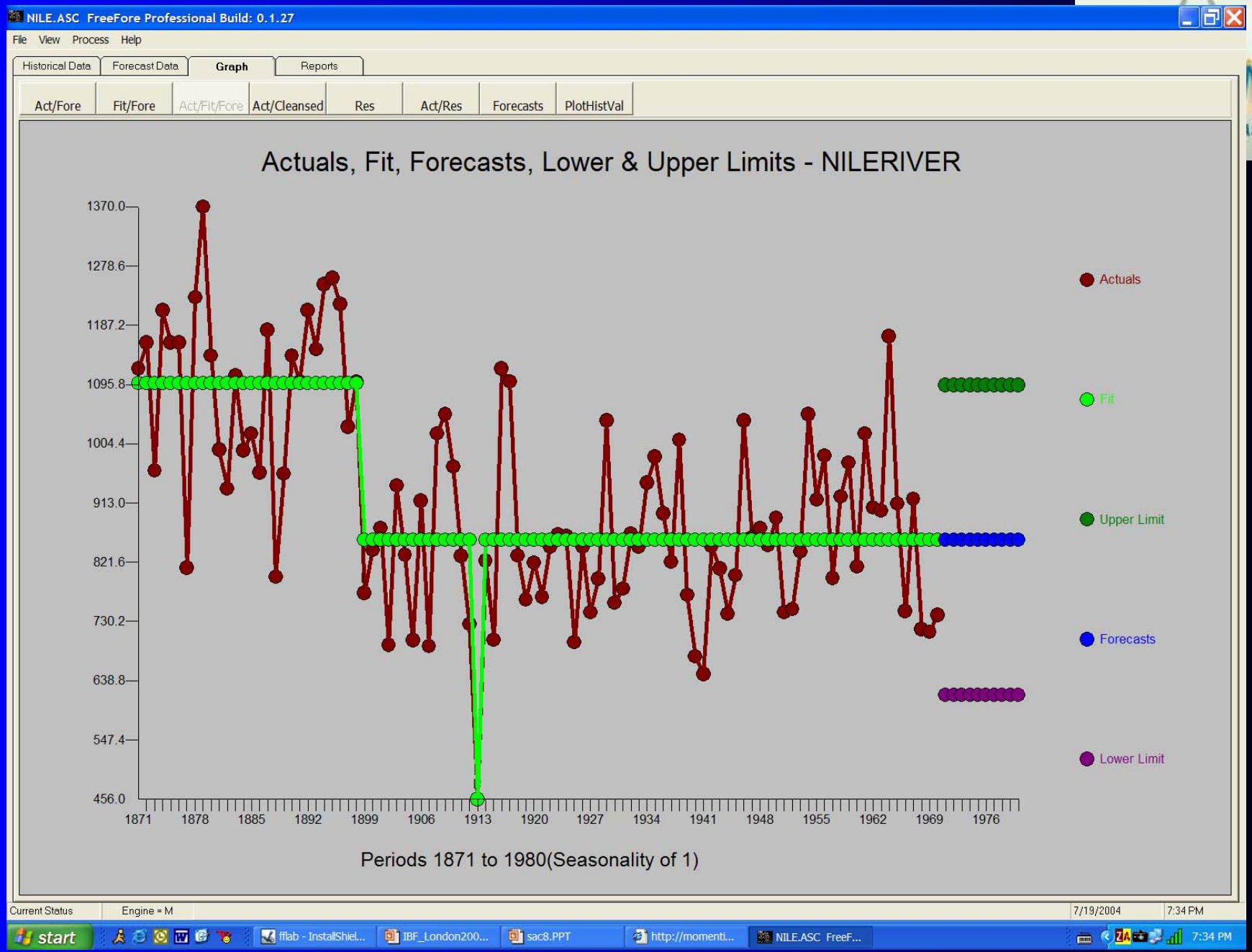

Actuals and Forecasts - SALESTOTALH2307
78.0
70.2
62.4
54.6
46.8
39.0
31.2
23.4
15.6
7.8
0.0
1/1
9/4
17/7
26/3
34/6
43/2
51/5
60/1
68/4
76/7
85/3
93/6
102/2
110/5
119/1
127/4
135/7
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 1/1 to 136/2(Seasonality of 7)


Actuals and Forecasts - SLSTOTL53333
2645.0
2380.5
2116.0
1851.5
1587.0
1322.5
1058.0
793.5
529.0
264.5
0.0
-264.5
6/5
12/1
17/4
22/7
28/3
33/6
39/2
44/5
50/1
55/4
60/7
66/3
71/6
77/2
82/5
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 1/2 to 88/7(Seasonality of 7)

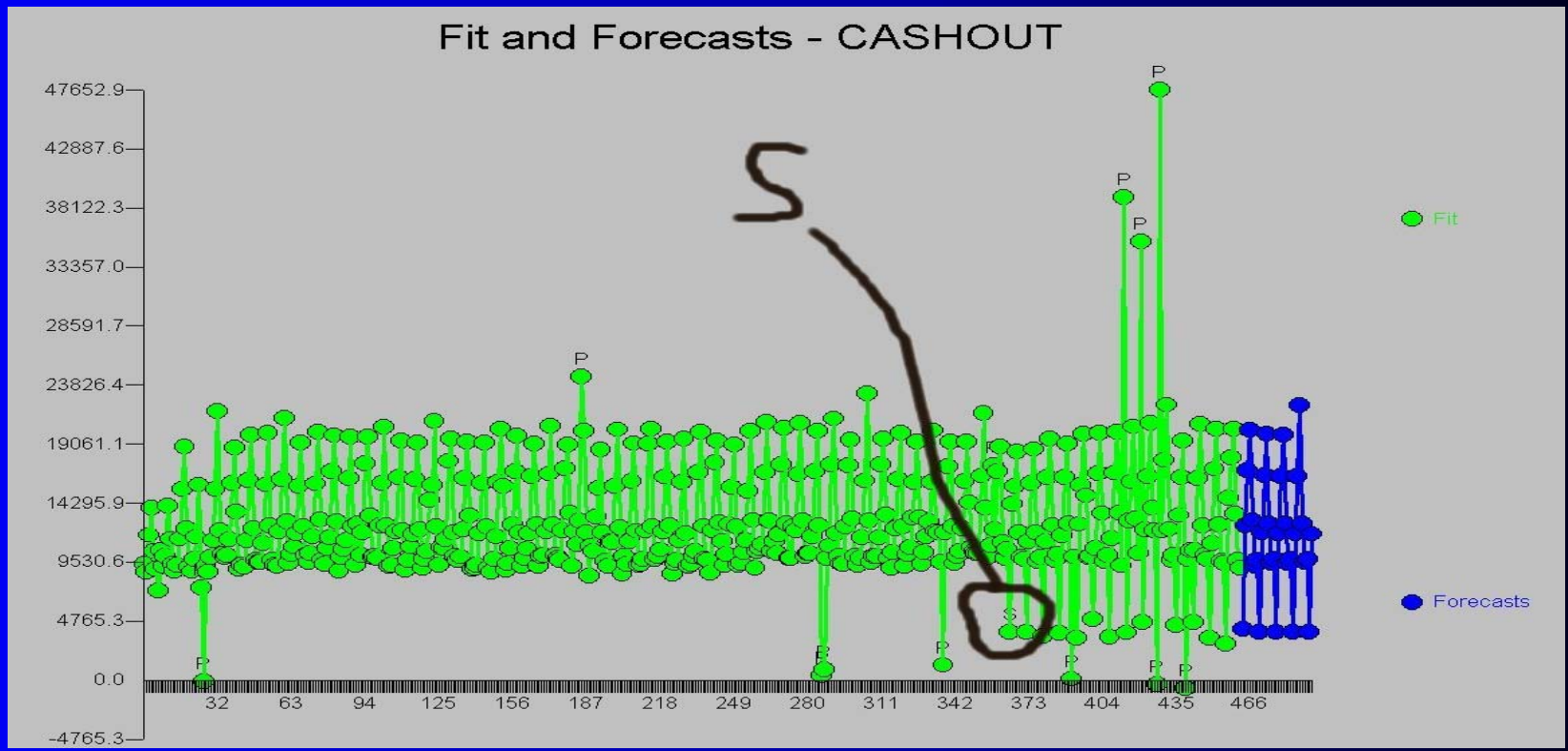

testing1234.asc FreeFore Professional Build: 0.1.17
File View Process Help
Historical Data
Future Values
Forecast Data
Graph
Reports
Whatlf
Act/Fore
Fit/Fore
Act/Fit/Fore
Act/Out Adj
Res
Act/Res
Forecasts
PlotHistVal
Actuals
92.0
82.8
73.6
64.4
55.2
46.0
36.8
27.6
18.4
9.2
0.0
1/6
14/2
26/5
39/1
51/4
63/7
76/3
88/6
101/2
113/5
126/1
138/4
150/7
163/3
175/6
188/2
200/5
Actuals
Periods 1/6 to 201/2(Seasonality of 7)
Current Status
Engine = L
11/9/2003
4:28 PM
Start
Quick Launch
4:28 PM


Actuals and Forecasts - S514D1C1
9.0
8.1
7.2
6.3
5.4
4.5
3.6
2.7
1.8
0.9
0.0
42/5
50/4
58/3
66/2
74/1
81/7
89/6
97/5
105/4
113/3
121/2
129/1
136/7
144/6
152/5
160/4
168/3
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 42/5 to 168/3(Seasonality of 7)


Fit and Forecasts - S514D1C1
8.8
8.0
7.1
6.2
5.3
4.4
3.5
2.7
1.8
0.9
0.0
-0.9
50/4
58/3
66/2
74/1
81/7
89/6
97/5
105/4
113/3
121/2
129/1
136/7
144/6
152/5
160/4
168/3
Fit
Forecasts
Periods 42/5 to 168/3(Seasonality of 7)


NILE.ASC FreeFore Professional Build: 0.1.27
File View Process Help
Historical Data
Forecast Data
Graph
Reports
Act/Fore
Fit/Fore
Act/Fit/Fore
Act/Cleansed
Res
Act/Res
Forecasts
PlotHistVal
Actuals, Fit, Forecasts, Lower & Upper Limits - NILERIVER
1370.0
1278.6
1187.2
1095.8
1004.4
913.0
821.6
730.2
638.8
547.4
456.0
1871
1878
1885
1892
1899
1906
1913
1920
1927
1934
1941
1948
1955
1962
1969
1976
Actuals
Fit
Upper Limit
Forecasts
Lower Limit
Periods 1871 to 1980(Seasonality of 1)
Current Status
Engine = M
7/19/2004
7:34 PM
start
fflab - InstallShiel...
IBF_London200...
sac8.PPT
http://momenti...
NILE.ASC FreeF...
7:34 PM

Daily Demand For Cash



The source of the spurious correlation is a common cause acting on the variables. In the recent *spurious regression* literature in time series econometrics (Granger & Newbold, *Journal of Econometrics,* 1974) the misleading inference comes about through applying the regression theory for stationary series to non-stationary series. The dangers of doing this were pointed out by G. U. Yule in his 1926 "Why Do We Sometimes Get Nonsense Correlations between Time-series? A Study in Sampling and the Nature of Time-series," *Journal of the Royal Statistical Society,* **89,** 1-69.

Three Components of a Model

Combination of Three Kinds of Structures

Y_t = Causal + Memory + Dummy

CAUSAL

- Using expected events (i.e. holidays, weather,day-of-the-week, week-of-the-year, 1st/15th of the month effects, etc.)

MEMORY

- Using historical values such as Demand Last Week, Yesterday etc.

DUMMY

- Using Day-of-the Week Profiles, Growth Patterns over Time (Level Shifts and/or Local Time Trends) , Week-of-the-Year Profiles.

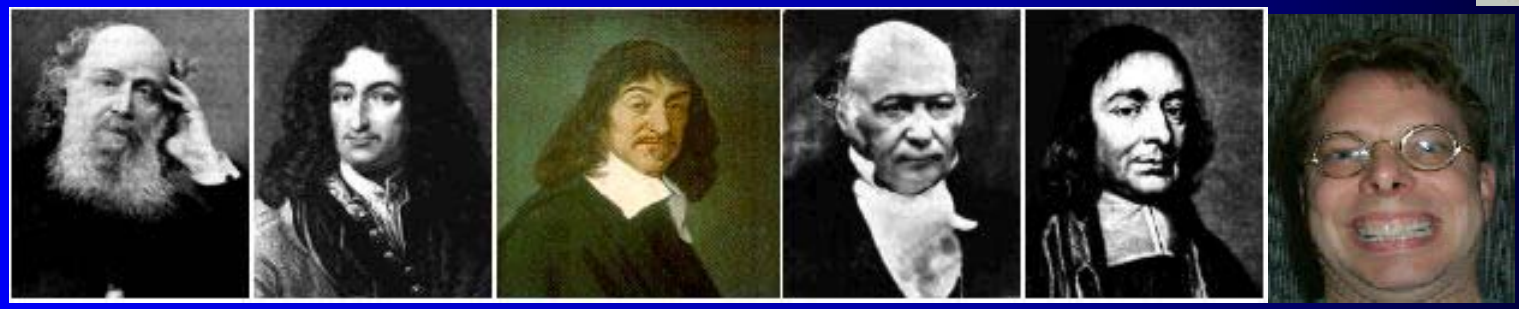
What does Autobox do?

- Autobox does not select a model from a user or system-defined set of models.
- To produce more-accurate forecasts, Autobox automatically tailors the forecast model to each problem. and the best weightings.
- It corrects for omitted variables e.g.,competitive activity that have had historical effects by identifying pulses, seasonal pulses, level shifts and local time trends, and then enhances the forecast model through dummy variables and/or autoregressive memory schemes.

Forecasting History (CAUSAL)

Historical development of regression and correlation

Earliest Known Uses of Mathematical Expressions



LEFT TO RIGHT: **James Joseph Sylvester,** who introduced the words *matrix, discriminant, invariant, totient,* and *Jacobian;* **Gottfried Wilhelm Leibniz,** who introduced the words *variable, constant, function, abscissa, parameter, coordinate* and perhaps *derivative;* **René Descartes,** who introduced the terms *real number* and *imaginary number;* **Sir William Rowan Hamilton,** who introduced the terms *vector, scalar, tensor, associative;* **John Wallis,** who introduced the terms *induction, interpolation* and *hyper geometric series; and Mark Frost who first coined the term "AUTOBOX is Great" or "AUTOBOX-o-Akbar" and for maintaining a database of statistics for Playboy Bunnies.*

The story…

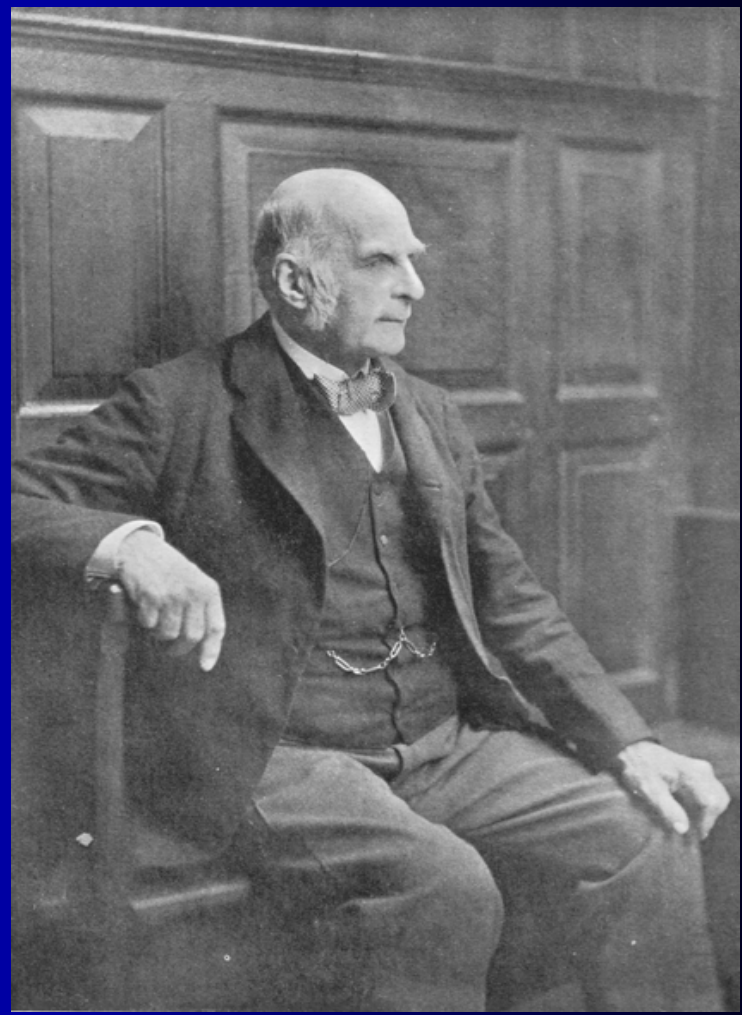
- The complete name of the correlation coefficient leads many students to believe that Karl Pearson developed the statistical measure himself.
- Sir Francis Galton originally conceived the modern notions of regression and correlation.
- Pearson developed rigorous treatment of mathematics of Pearson Product Moment Correlation

Historical Outline

- **Galton**: Heredity experiments lead to initial concepts of regression and correlation
- **Edgeworth**: Estimating Correlation Coefficient. Involves Pearson in the subject
- **Pearson**: "Rigorously" derives best value for correlation coefficient
- **Fisher**: Combines the components into one discipline. Intraclass correlation and Analysis of Variance

Sir Francis Galton

- Tropical Explorer
- Eugenicist
- Statistician
- Anthropologist
- Criminologist
- Hereditarian
- Half-cousin of Charles Darwin
- Psychologist

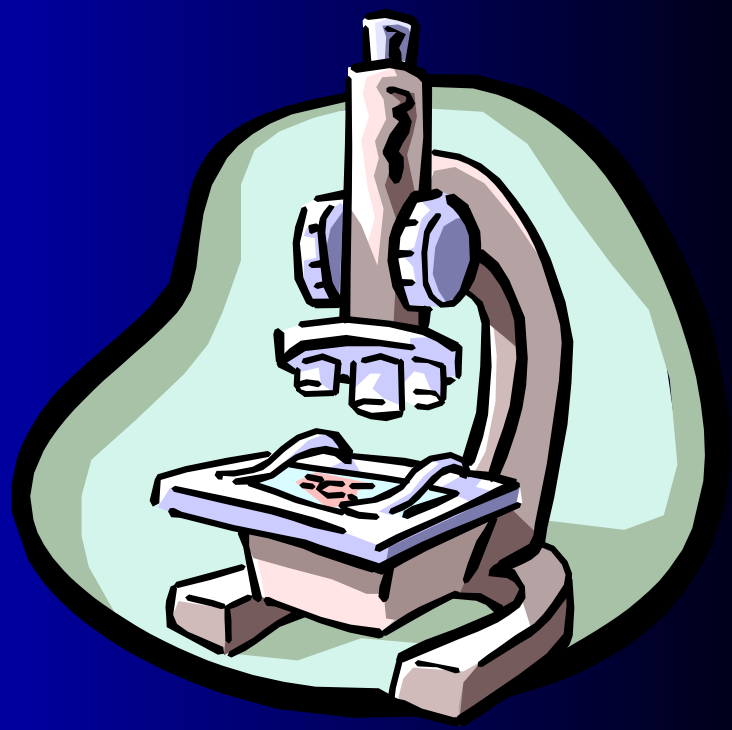


Sir Francis Galton

In 1884, Galton wrote to the distinguished botanist, Alphonse de Candolle: "It strikes me that the Jews are specialized for a parasitical existence upon other nations, and that there is need of evidence that they are capable of fulfilling the varied duties of a civilized nation by themselves." Karl Pearson, Galton's disciple and biographer, echoed this opinion 40 years later during his attempt to prove the undesirability of Jewish immigration into Britain: "...for such men as religion, social habits, or language keep as a caste apart, there should be no place. they will not be absorbed by, and at the same time strengthen the existing population; they will develop into a parasitic race..." (Hirsch, 1970/1976, p. 161).

Initial conceptualization

- Galton's experiment with sweet peas (1875) led to the development of initial concepts of linear regression.



“Sweet peas” experiment

- Experiment conducted in 1875
- Sweet peas could self-fertilize: “daughter plants express genetic variations from mother plants without contribution from a second parent.”
- Eliminated the problem of statistically assessing genetic contributions from multiple sources.

“Sweet peas” experiment (2)

- Distributed packets of seeds to 7 friends
- Uniformly distributed sizes, split into 7 size groups with 10 seeds per size.
- There was substantial variation among packets.
- 7 sizes 10 seeds per size 7 friends = 490 seeds
- Friends were to harvest seeds from the new generation of friends and return them to Galton.

Birth of regression

- Plotted the weights of daughter seeds against weights of mother seeds.
- Hand fitted a line to the data
- Slope of the line connecting the means of different columns is equivalent to regression slope.

Reversion

- Dispersion among the progeny seeds didn't lead to populations increasingly variable from generation to generation. Why?
- Galton's answer: Reversion
- Daughter weights were distributed closer to the average population weight than that of the parent.
- "The mean progeny reverted to type and … variation was just sufficient to maintain population variability"
- AKA regression toward the mean.

Reversion

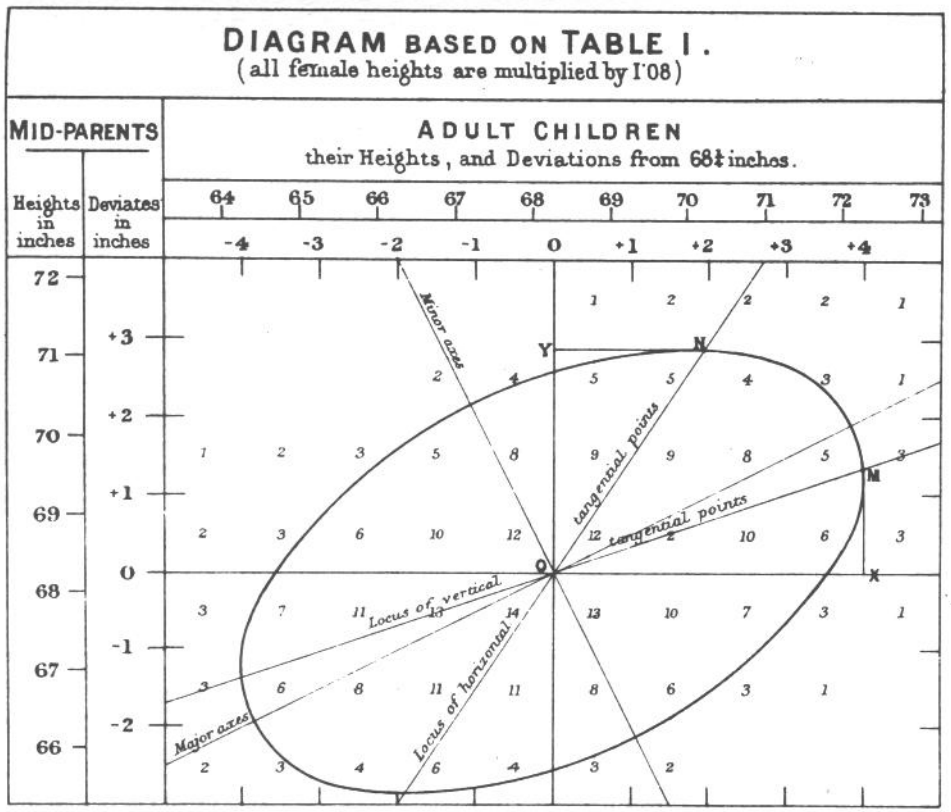
- Galton's model appears in the Appendix (p. 532) to his "Typical laws of heredity," *Nature* 15 (1877), 492-495, 512-514, 532-533. Galton here focused on the inheritance of measurable characteristics; his observations are on the weight of peas. The key idea is that the offspring does not inherit all the peculiarities of the parents but is pulled back to the average of its ancestors. The idea is expressed in what would now be called a stable first-order normal autoregressive process where "time" is measured in generations. The process is stable because the *reversion coefficient* is the fraction of the parental deviation that is inherited.
- .

Moving on…

Symmetric studies of Stature

- In 1885 Galton was observing relationship between midparent (average of the two parents) and offspring heights.
- He normalized the heights of females and males by multiplying heights of females by 1.08 (estimated from data)

Graphical Representation



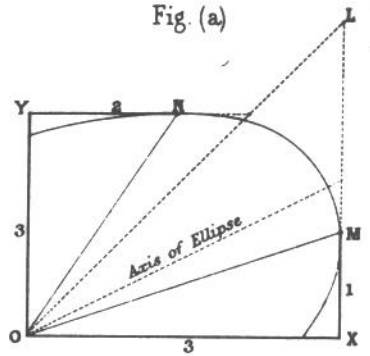


Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)

Graphical Representation (2)

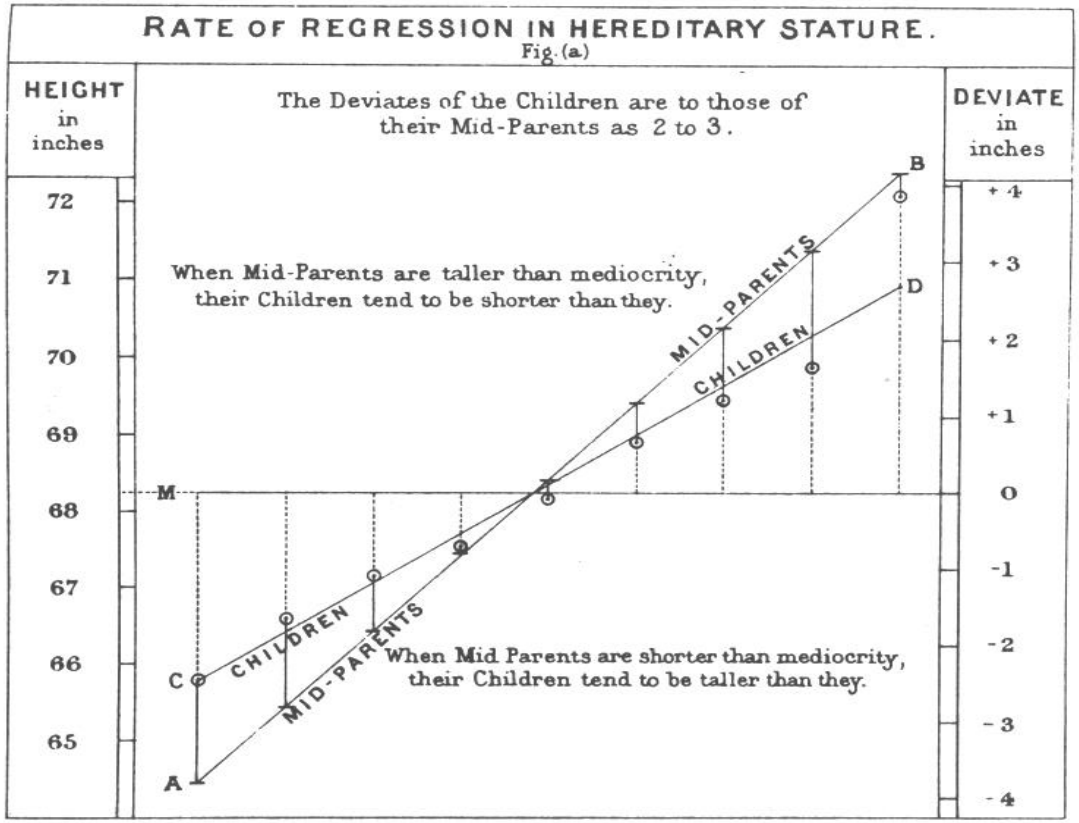


Figure 8.8. Galton's graphical illustration of regression; the circles give the average heights for groups of children whose midparental heights can be read from the line AB. The difference between the line CD (drawn by eye to approximate the circles) and AB represents regression toward mediocrity. (From Galton, 1886a.)

Galton on Correlation

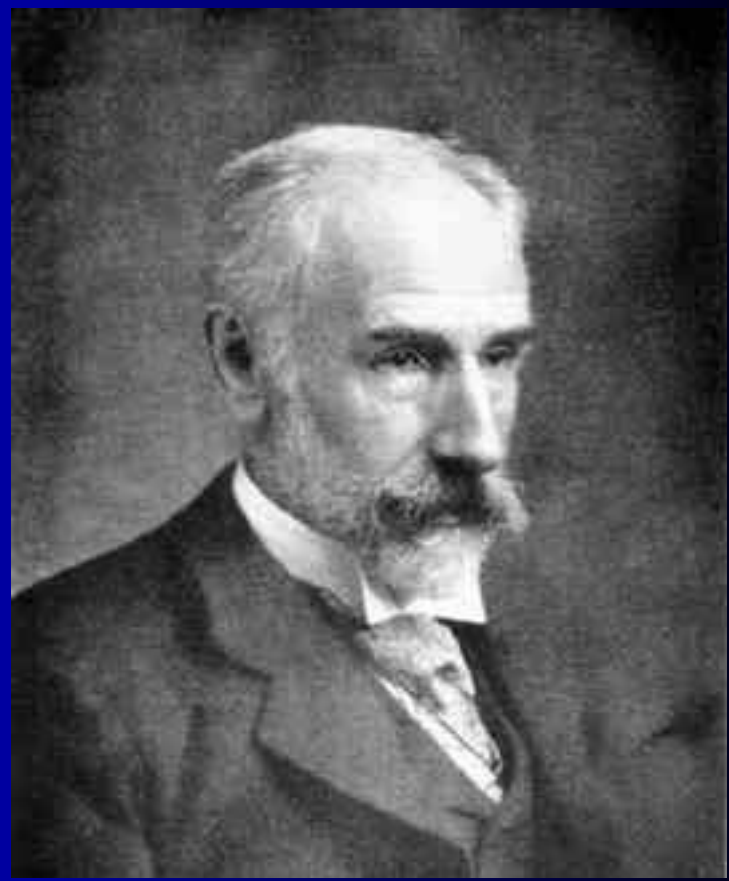
- December 1888, Galton's "Co-relations and their measurement, chiefly from anthropometric data"
- If both measurements (midparent and child's height) were expressed in terms of their probable errors, then both regression lines had same slope r (closeness of co-relation).
- In addition, "co-relation" was originally used because "correlation" was taken and had different meaning.

Galton on Correlation

- December 1988, Galton's "Co-relations and their measurement, chiefly from anthropometric data"
- If both measurements (midparent and child's height) were expressed in terms of their probable errors, then both regression lines had same slope r (closeness of co-relation).
- In addition, "co-relation" was originally used because "correlation" was taken and had different meaning.

Francis Ysidro Edgeworth

- Born in Dublin, Ireland.
- Distant cousin of Galton.
- Education in Classical Literature(1869)
- Passed the bar (1877)
- Got interested in statistics, self-taught, had better grasp of mathematical statistics than most of his contemporaries
- Wanted to extend implementation of mathematics to social sciences.



Estimating Correlation Coefficients (Galton)

- Start with table showing cross classified grouped data (in terms of deviations)
- Find mean value for each row (stature)
- Plot it against the row value.
- Interchange rows and columns and repeat

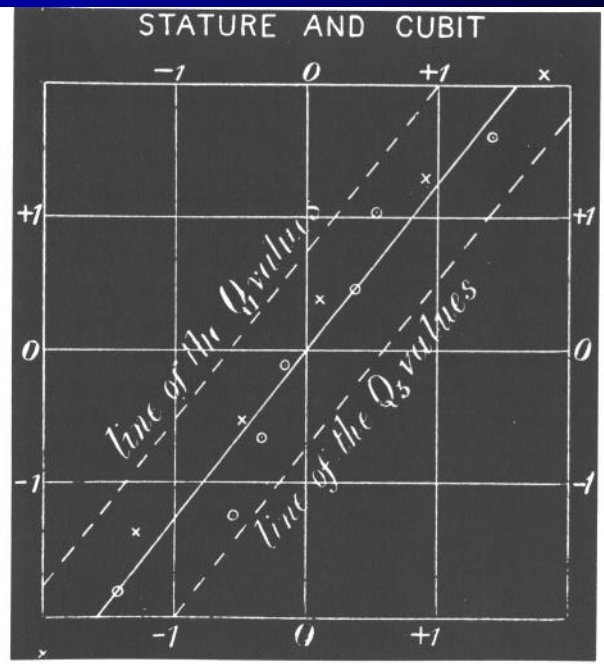


Figure 9.2. Galton's graphical determination of a correlation coefficient, based upon the data in Table 9.1. Galton converted both stature and cubit measures to standard units (number of probable errors from the median), and then for each row plotted median cubit versus stature (the circles), and for each column plotted median stature versus cubit (the crosses). In both cases the medians were plotted along the x axis. He found the solid line by eye, and took the correlation coefficient as the inclination of that line to the vertical, which he judged to be r = 0.8. Dotted lines approximating within row (and within column) quartiles are also shown. (From Galton, 1888.)

Edgeworth's estimation (attempt #1)

- Focused on individuals with above average values but excluded those with extreme values for which "the law of error is liable to break down"

1. Find average cubit length and average stature
2. Convert both figures to deviations from the population averages in standard units.
3. Estimate correlation coefficient as the ratio of standardized average stature divided by standardized average cubit length.

Criticism of Edgeworth

- He made a series of arithmetic errors and came up with the same correlation coefficient as Galton!
- In reality this estimation tended to be strongly biased toward 0.
- Actual value was $\rho = 0.68$, not the computed 0.80.

Edgeworth's estimation (attempt #2)

Suppose that we are dealing with standardized variables: "each measured from the corresponding average, in units of the proper modulus"

$$\rho = \frac{\sum \left(x^2 \cdot y / x \right)}{\sum \left(x^2 \right)} = \frac{\sum \left(xy \right)}{\sum \left(x^2 \right)}$$

This amounts to finding least squares estimate of ρ, regressing y on x, with error having modulus.

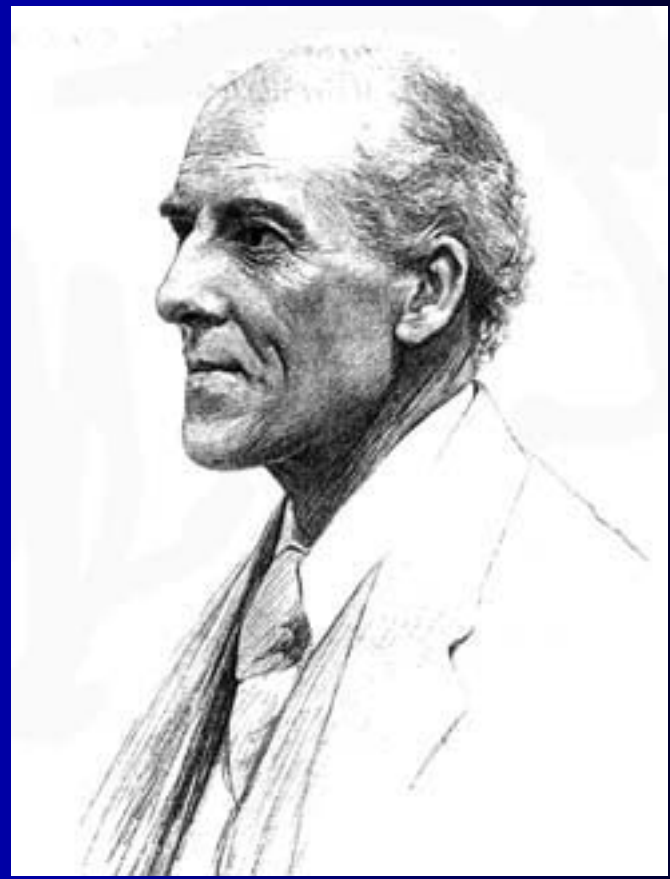
$$\frac{\sqrt{\sum \left[x^2 (1 - \rho)^2 \right]}}{\sum \left(x^2 \right)} = \frac{\sqrt{1 - \rho^2}}{\sqrt{\sum \left(x^2 \right)}} = \frac{\sqrt{1 - \rho^2}}{\sqrt{n/2}}$$

Edgeworth's Estimation (cntd.)

- This is similar to Pearson's product moment estimate of the correlation coefficient.
- Depending on standardization modulus it may even be exactly same as Pearson's.
- Since Edgeworth didn't explicitly specify the modulus until after Pearson came up with his estimation, the matter remains under doubt.

Karl Pearson

- Was skeptical of Galton's work until 1892, after corresponding with Edgeworth
- Coined the term standard deviation
- Credited with "the best value for correlation coefficient (Pearson's coefficient of correlation)



Sir Ronald A. Fisher

- 1921 introduced concept of likelihood.
- 1922 gave new definition of statistics (reduction of data).
- Had long-standing dispute with Pearson.
- Not a cousin of Darwin

Regression Function

- Regression function takes form:

$$Y = a + b(x - \bar{x})$$

- Where a and b can be computed by:

$$a = \bar{y}$$

$$b = \frac{\sum \{y(x - \bar{x})\}}{\sum \{(x - \bar{x})^2\}}$$

Estimating Correlation Coefficient

- "The method of calculation might have been derived from the consideration that the correlation is the geometric mean of the two regression coefficients"

$$\frac{nrs_1 s_2}{ns_1^2} \text{ and } \frac{nrs_1 s_2}{ns_2^2}$$

,where s_1,s_2 are estimates of deviations from the average

- Thus we can estimate ρ (in modern notation) as

$$\rho = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Flawed When Applied To Time Series Data

A source for spurious correlation is a common cause acting on the variables. In the recent *spurious regression* literature in time series econometrics (Granger & Newbold, *Journal of Econometrics,* 1974) the misleading inference comes about through applying the regression theory for stationary series to non-stationary series. The dangers of doing this were pointed out by G. U. Yule in his 1926 "Why Do We Sometimes Get Nonsense Correlations between Time-series? A Study in Sampling and the Nature of Time-series," *Journal of the Royal Statistical Society,* **89,** 1-69.

Flawed When Applied To Time Series Data (2)

More generally the misleading inference comes about through applying the regression theory for stationary series to series that have auto-regressive structure. Recognizing this , early researchers attempted to extract the within relationship (autoregressive structure) and then proceed to examine cross-correlative (among) relationships.

Flawed When Applied To Time Series Data (2)

Initial attempts to adjust for within relationships included “de-trending” and/or differencing. Both of which were usually presumptive and often lead to “Model Specification Bias”. Box and Jenkins codified this process by recognizing that an ARIMA filter was the optimum transform to extract the “within structure” prior to identifying the “among structure” . They pointed out that both “de-trending” and “differencing” were particular cases of a filter, whose optimized form was an ARMAX model potentially containing both ARIMA and Dummy Variables such as Trends.

How to Identify the Relationship

The first step to this process is to develop an ARIMA model for each of the user-specified input time series in the equation. Each series must then be made stationary by applying the appropriate differencing and transformation parameters from its ARIMA model. Each input series is prewhitened by its own ARIMA model AR (autoregressive) and MA (moving average) factors. The output series is also prewhitened by the input series AR and MA factors.. The cross correlations between the prewhitened input and output reveal the extent of this interrelationship.

SPURIOUS CORRELATION. The term was introduced by Karl Pearson in "Mathematical Contributions to the Theory of Evolution - On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs," *Proc. Royal Society,* **60,** (1897), 489-498. Pearson showed that correlation between indices u (= x/z) and v (= y/z) was a misleading guide to correlation between x and y.

The term has been applied to other correlation scenarios with potential for misleading inferences. In Student's "The Elimination of Spurious Correlation due to Position in Time or Space" (*Biometrika,* **10,** (1914), 179-180) the source of the spurious correlation is the common trends in the series

The dangers of doing this were pointed out by G. U. Yule in his 1926 "Why Do We Sometimes Get Nonsense Correlations between Time-series? A Study in Sampling and the Nature of Time-series," *Journal of the Royal Statistical Society,* **89,** 1-69.

In 1951 Durbin-Watson Tests were developed to test the hypothesis that the residuals from an OLS model were uncorrelated for LAG 1. (N.B. that ONLY lag 1 was being tested)

Residual Analysis for Independence: The Durbin-watson Statistic

- Used when data is collected over time to detect autocorrelation (Residuals in one time period are related to residuals in another period)
- Measures Violation of independence assumption

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Should be close to 2.

If not, examine the model for autocorrelation.

Durbin-Watson Tests are Only Valid When The Model Being Tested has No Lags of The Output (Y) Series

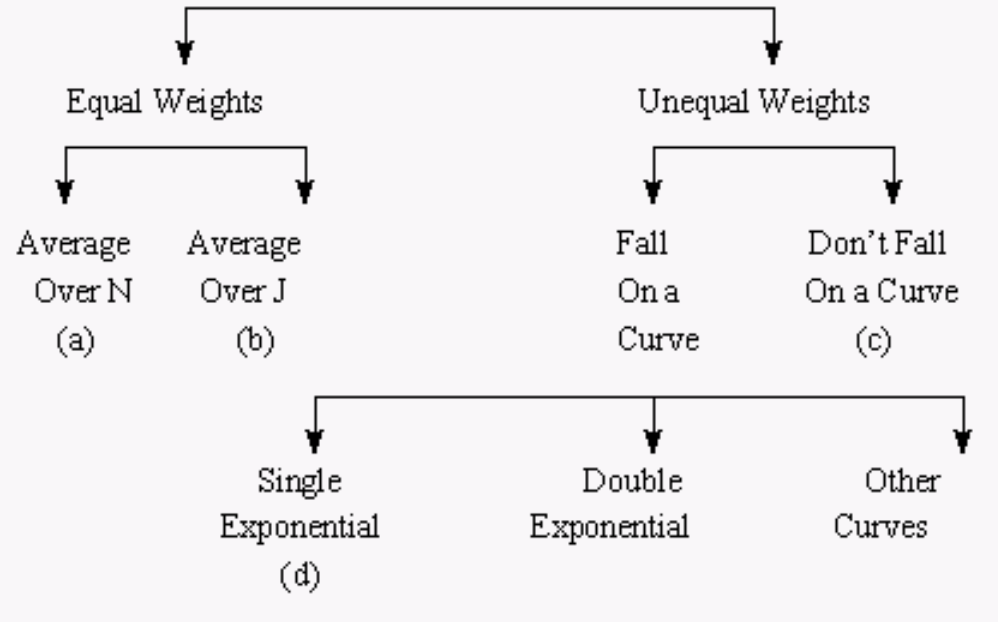
Combination of Three Kinds of Structures

$$Y_t = \text{Causal} + \text{Memory} + \text{Dummy}$$

Historical development of Memory

AUTOMATIC FORECASTING SYSTEMS

Memory Model



Auto-Projective Equations

a) $Y_{N+1} = (1/N)*Y_1 + (1/N)*Y_2 + (1/N)*Y_3 + \ldots\ldots\ldots\ldots (1/N)*Y_N$

b) $Y_{N+1} = (1/J)*Y_N + (1/J)*Y_{N-1} + (1/J)*Y_{N-2}$ where J=3

c) $Y_{N+1} = .6*Y_N + .3*Y_{N-1} + .1*Y_{N-2}$ where .6,.3,.1 are the weights

a) $Y_{N+1} = (1/N)*Y_1 + (1/N)*Y_2 + (1/N)*Y_3 + \ldots\ldots\ldots\ldots (1/N)*Y_N$

b) $Y_{N+1} = (1/J)*Y_N + (1/J)*Y_{N-1} + (1/J)*Y_{N-2}$ where J=3

c) $Y_{N+1} = .6*Y_N + .3*Y_{N-1} + .1*Y_{N-2}$ where .6,.3,.1 are the weights

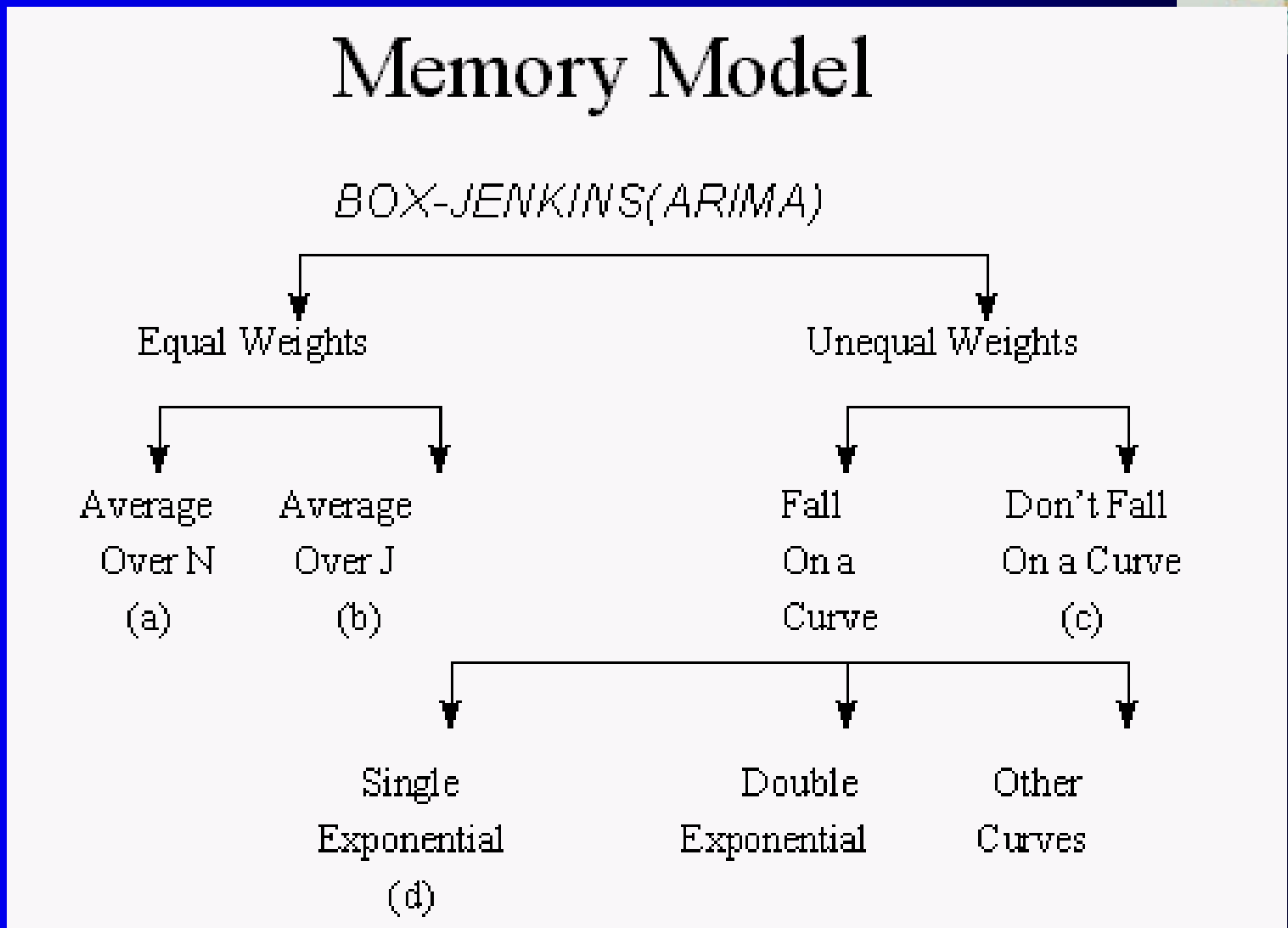
d) $Y_{N+1} = C1*Y_N + C2*Y_{N-1} + C3*Y_{N-2} + CK*Y_{N-K}$ where C1,C2,C3

are the weights for example:

$C1 = .8, C2 = .2*.8, C3 = .2*.2*.8$, etc. $CK = .2^{(K-1)}.8$

$C1 = .8 \quad C2 = .16 \quad C3 = .032$

The Family Tree



Consider an "N Period" Equally Weighted Model

$$Y_{N+1} = (1/N)*Y_1 + (1/N)*Y_2 + (1/N)*Y_3 + \ldots\ldots\ldots\ldots (1/N)*Y_N$$

$$Y_{N+1} = (1/N)*Y_1 + (1/N)*Y_2 + (1/N)*Y_3 + \ldots\ldots\ldots\ldots (1/N)*Y_N$$

The Mechanics of a 60 day Weighted Average

If you wished to use a 60 period equal weighted average you would need to have available the most recent 60 values. In the early days of computing storage was a major problem thus Statistical Innovation was in order.

Relationship Between Number of Observations in an Equally Weighted Average and The Exponential Model Smoothing Coefficient in terms of Average Age of the Data

Number of Observations	Variance of Estimate	Smoothing Constant
3	0.333	0.5
4	0.25	0.4
5	0.2	0.333
5.67	0.177	0.3
6	0.167	0.286
9	0.111	0.2
12	0.083	0.154
18	0.056	0.105
19	0.053	0.1
24	0.042	0.08
39	0.026	0.05
52	0.019	0.038
199	0.005	0.01

R.G. Brown in 1961 developed the concept of capturing historical data in a forecast and then using that forecast and an adjustment for the last error to get a new forecast.

$$Y(new)=(1-a)*Y(old)+a*error$$

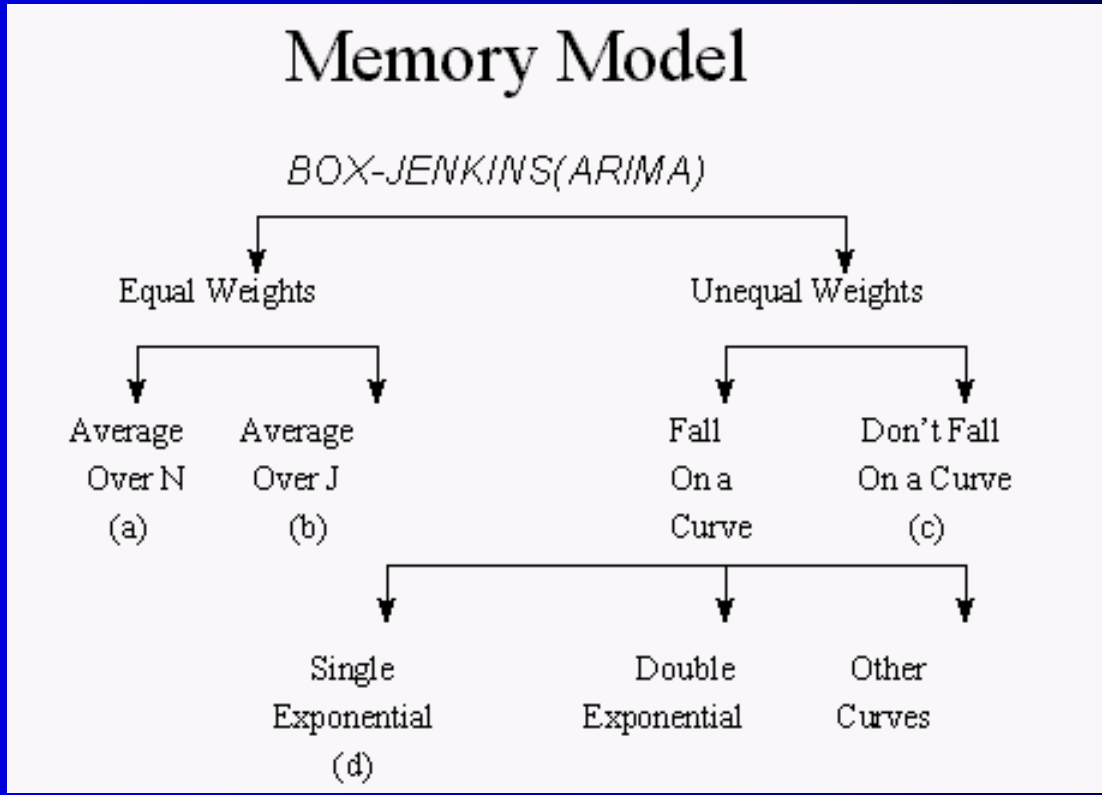
There was no theoretical development used just the idea that one could quickly compute an updated forecast and only two values were required to be stored.

1. The Previous Forecast
2. The Smoothing Coefficient(a)

In terms of selecting the appropriate Smoothing Coefficient, one was told to try different values between 0. and 1.0 and see which one you like best. Failing that you could call NYC and find out what they liked !

This method had an intuitive appeal as it was equivalent to exponentially forgetting the past or equivalently equally weighting a recent set without having to store all the data. The IT folks just loved it as it was fast and efficient if not as accurate as could be developed

The Family Tree



Box and Jenkins in 1963 suggested using autoregressive coefficients to IDENTIFY the nature of the required memory structure rather than assuming it as Brown had done.

This lead rather naturally into pattern recognition schemes to automatically identify the form of the model ….thus AUTOBOX was introduced in the early 70's

Combination of Three Kinds of Structures

$$Y_t = \text{Causal} + \text{Memory} + \text{Dummy}$$

Historical development of Dummy

Early researchers assumed Trend Models and Additive Seasonal Factors like the Holt-Winters Class of Models. Again identification was bypassed and Estimation was conducted based upon an assumed model.

No thought was given to distinguishing between Level and Trend Changes or the detection of break points in trends. No consideration was given to detecting the onset of “seasonal factors”

Intervention Detection schemes introduced in the early 1980's suggested the empirical construct of Dummy Variables. The literature sometimes refers to Outliers (a one-time Pulse).

AUTOMATIC FORECASTING
SYSTEMS

Intervention Analysis/AIA References

Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, 2nd ed. San Francisco: Holden Day.

Box, G.E.P., and Tiao, G. (1975). "Intervention Analysis with Applications to Economic and Environmental Problems," Journal of the American Statistical Association, Vol 70, pp. 70-79.

Chang, I., and Tiao, G.C. (1983). "Estimation of Time Series Parameters in the Presence of Outliers," Technical Report #8, Statistics Research Center, Graduate School of Business, University of Chicago, Chicago.

McCleary, R., and Hay, R. (1980). Applied Time Series Analysis for the Social Sciences. Los Angeles: Sage.

Reilly, D.P. (1980). "Experiences with an Automatic Box-Jenkins Modeling Algorithm," in Time Series Analysis, ed. O.D. Anderson. (Amsterdam: North-Holland), pp. 493-508.

Reilly, D.P. (1987). "Experiences with an Automatic Transfer Function Algorithm," in Computer Science and Statistics Proceedings of the 19th Symposium on the Interface, ed. R.M. Heiberger, (Alexandria, VI: American Statistical Association), pp. 128-135.

Tsay, R.S. (1986). "Time Series Model Specification in the Presence of Outliers," Journal of the American Statistical Society, Vol. 81, pp. 132-141.

Wei, W. (1989). Time Series Analysis Univariate and Multivariate Methods. Redwood City: Addison Wesley.

Outliers

- One time events that need to be “corrected for” in order to properly identify the general term or model
- Consistent events (i.e. holidays, events) that should be included in the model so that the future expected demand can be tweaked to anticipate a pre-spike, post spike or at the moment of the event spike.
- If you can’t identify the reason for the outlier than you will not get to the root of the process relationship and be relegated to the passenger instead of the driver

OUTLIERS: WHAT TO DO ABOUT THEM?

- OLS procedures are INFLUENCED strongly by outliers. This means that a single observation can have excessive influence on the fitted model, the significance tests, the prediction intervals, etc.

- Outliers are troublesome because we want our statistical models to reflect the MAIN BODY of the data, not just single observations.

Outliers

- Working definition
  - An outlier x_k is an element of a data sequence S that is inconsistent with out expectations, based on the majority of other elements of S.
- Sources of outliers
  - Measurement errors
  - Other uninteresting anomalous data
    - valid data observations made under anomalous conditions
  - *Surprising observations that may be important*

Peculiar Data

- Zhong, Ohshima, and Ohsuga (2001):
  - Hypotheses (knowledge) generated from databases can be divided into three categories
    - Incorrect hypotheses
    - Useless hypotheses
    - New, surprising, interesting hypotheses
- To find last class, authors suggest looking for *peculiar data*
  - A data is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in the data set.

Why 3-sigma fails

AUTOMATIC FORECASTING SYSTEMS

- Outlier sensitivity of mean and standard deviation
  - mean moves towards outliers
  - standard deviation is inflated
- Too few outliers detected (e.g., none)

Types of Outliers

- Pulse
- Seasonal Pulse
- Level Shift (changes in intercepts)
- Time Trends (changes in slopes)

Some Notation

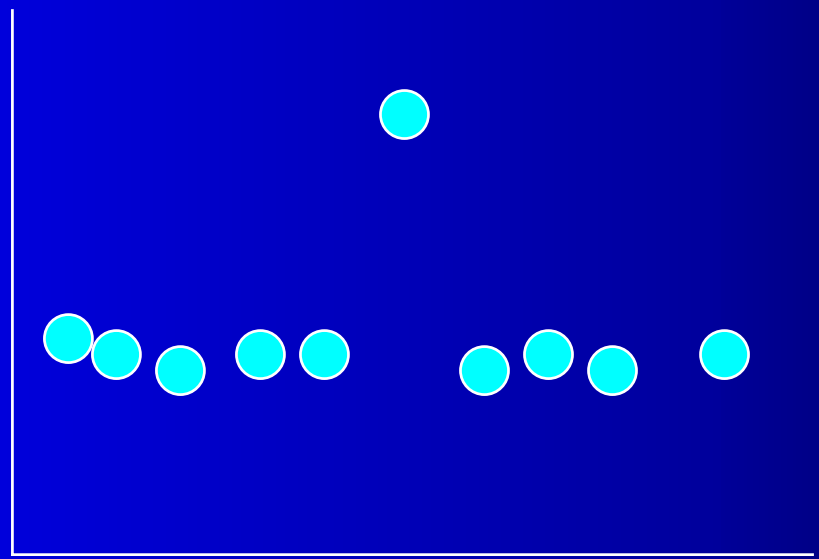
Y_t The series we are trying to predict/analyze

X_t The input or causal series that we think might be important

Example of a Pulse Intervention

Z_t represents a pulse or a one-time intervention at time period 6.

Z_t = 0,0,0,0,0,1,0,0,0



Number Crunching to find if there is a Significant Intervention

- We create an iterative computer based experiment where we establish a base case model(no intervention) and then compare the base case to models with an intervention.

- We then choose the model with smallest variance. If none of the intervention models has a significantly lower variance then the base model, then we keep the base case model.

$$Y_t = BO + B1X_t + B2L_t + B3Z_t + U_t$$

For simplicity purposes, we will drop the X_t and any L_t 's thus

$$Y_t = BO + B3Z_t + U_t$$

if there are no significant interventions it becomes

$$Y_t = BO + U_t$$

Base Case

$$Y_t = BO + U_t$$

We will estimate this model using a standard regression model with only an intercept to

get $\hat{BO}$ and σ^2_U

Modeling Interventions - Pulse

We will first try $Y_t = BO + B3Z_t + U_t$

where $Z_t = 1,0,0,0,0,0,0,0,0,,,,,,,,,0$

or $Z_t = 1 \quad t = 1$

$Z_t = 0 \quad t > 1$

We run our regression with a pulse at time period = 1.

σ^2_U is an indicator of how just good our candidate intervention model is.

Modeling Interventions - Pulse

It's clear we can create a second candidate intervention model which has Z_t = 0,1,0,0,0,0,0,0,0,0,,,,,,,,,,0

We run our regression with a pulse at time period = 2.

We can continue this path for all possible time periods.

Table of Summary Variances

(1)σ^2_U Base Case (No Pulse)

(2)σ^2_U Pulse at time period=1

(3)σ^2_U Pulse at time period=2

•

•

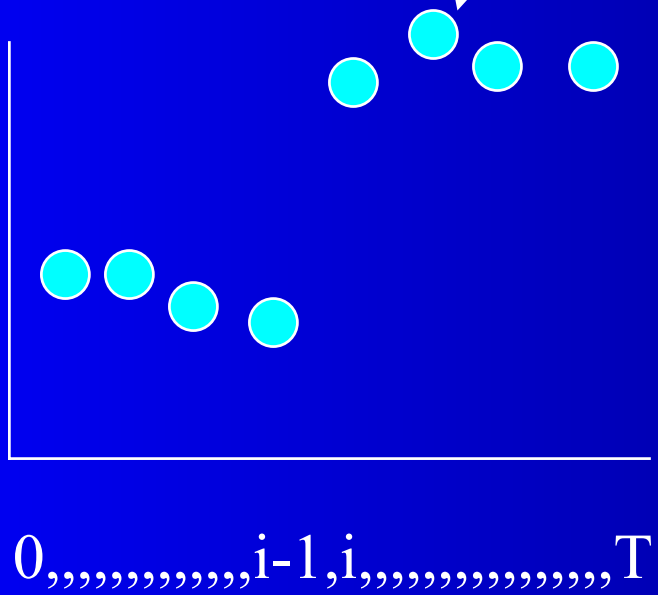
•

(60)σ^2_U Pulse at time period=T

If we had 60 observations then we would have run 61 regressions which yield 61 estimates of the variance.

Modeling Interventions - Level Shift

If there was a level shift and not a pulse then it is clear that a single pulse model would be inadequate thus $Y_t = BO + B3Z_t + U_t$



Assume the appropriate Z_t is

$Z_t = 0,0,0,0,1,1,1,1,1,1,,,,,,,T$

or $Z_t = 0 \quad t < i$

$Z_t = 1 \quad t > i\text{-}1$

Modeling Interventions -Level Shift

- Similar to how we approached pulse interventions, we will try the various possible level shifts at the same time that we are also evaluating our base case and the pulse models
- .
- So our tournament of models is now up to 120; One base case model, 60 models for pulses and 59 models with level shifts.

Modeling Interventions – Level Shift

Our first level shift model would be

$Z_t = 0,1,1,1,1,1,1,1,,,,,,1$

$Z_t = 0 \quad i = 1$

$Z_t = 1 \quad i > 1$

We can continue this path for all possible time periods.

Table of Summary Variances

(1)σ^2_U Base Case (No Pulse)

(2)σ^2_U Pulse at time period=1

•

(61)σ^2_U Pulse at time period=T

(62)σ^2_U Level shift starting at time period=2

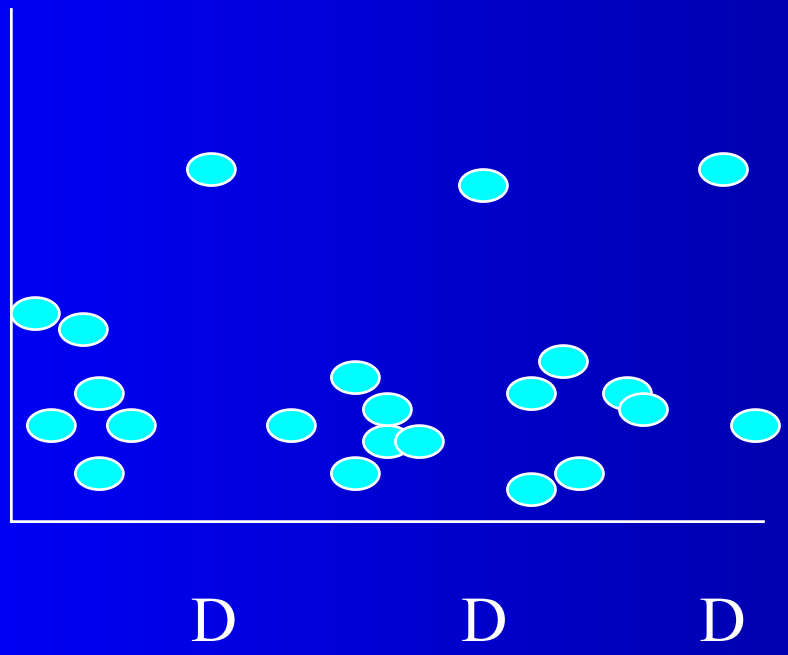
•

(120)σ^2_U Level shift starting at time period=T

Here are the 120 regressions which yield 120 estimates of the variance.

Modeling Interventions - Seasonal Pulses

- There are other kinds of pulses that might need to be considered otherwise our model may be insufficient. For example, December sales are high.



The data suggest this model

$$Y_t = BO + B3Z_t + U_t$$

$Z_t = 0 \quad i <> 12,24,36,48,60$

$Z_t = 1 \quad i = 12,24,36,48,60$

Modeling Interventions - Seasonal Pulses

- In the case of 60 monthly observations, we would have 48 candidate regressions to consider. We will try the various possible seasonal pulses at the same time that we are also evaluating our base case, pulse and level shift models.
- So our tournament of models is now up to 168; One base case model, 60 models for pulses and 59 models with level shifts, 48 models for seasonal pulses. The first seasonal model:

$$Z_t = 1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,,,,,,,T$$

Modeling Interventions - Seasonal Pulses

Our second seasonal pulse model would be

Z_t = 0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,,,,,

Z_t = 0 i <> 2,14,26,38,50

Z_t = 1 i = 2,14,26,38,50

We can continue this path for all possible time periods.

Table of Summary Variances

(1)σ^2_U Base Case (No Pulse)

(2)σ^2_U Pulse at time period=1

(60)σ^2_U Pulse at time period=T

(61)σ^2_U Level shift starting at time period=2

(120)σ^2_U Level shift starting at time period=T

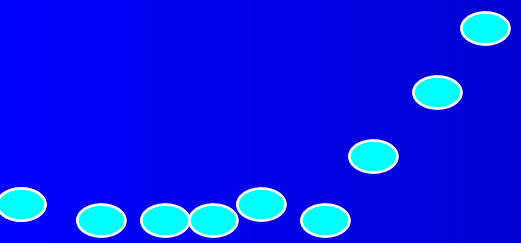
(121)σ^2_U Seasonal pulse starting at time period=1

(168)σ^2_U Seasonal pulse starting at time period=T

Here are the 168 regressions which yield 168 estimates of the variance.

Modeling Interventions - Local Time Trend

The fourth and final form of a deterministic variable is the the local time trend. For example,



1…… i-1, I,,, T

The appropriate form of Z_t is

$Z_t = 0 \quad t < i$

$Z_t = 1 \quad (t-(i-1)) * 1 >= i$

Z_t = 0,0,0,0,0,0,1,2,3,4,5,,,,,

Modeling Interventions - Local Time Trend

Our first local time trend model is

$Z_t = 1,2,3,4,5,6,7,,,,$

$Z_t = Z_t + 1 \quad i >= 1$

Our second local time trend model is

$Z_t = 0,1,2,3,4,5,6,7,,,,$

$Z_t = Z_t + 1 \quad i >= 2$

We can continue this path for all possible time periods.

Table of Summary Variances

(1)σ^2_U Base Case (No Pulse)

(2)σ^2_U Pulse at time period=1

(60)σ^2_U Pulse at time period=T

(61)σ^2_U Level shift starting at time period=2

(120)σ^2_U Level shift starting at time period=T

(121)σ^2_U Seasonal pulse starting at time period=1

(168)σ^2_U Seasonal pulse starting at time period=T

(169)σ^2_U Local time trend starting at time period=1

(228)σ^2_U Local time trend starting at time period=T

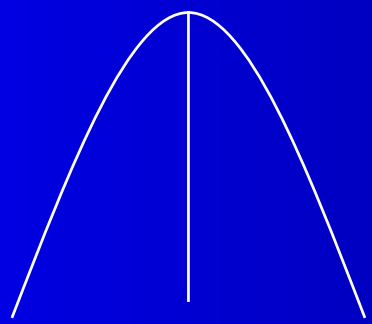
Here are the 228 regressions which yield 228 estimates of the variance.

- The intervention variable that generated the smallest error variance is the winner of the tournament. We now must test if this winner is statistically significant. In other words, has the winner created a reduction in the variance that is significantly different from zero?

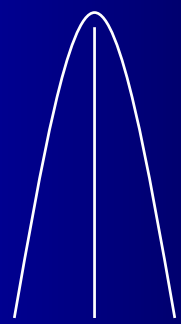
We employ a standard F test to measure whether the reduction is statistically significant.

$$F_{1,\,60} \cong \frac{(9*60) - (4*60)}{(4/60)}$$

where 60 is the # of observations



Base Case $\sigma^2_U = 9$

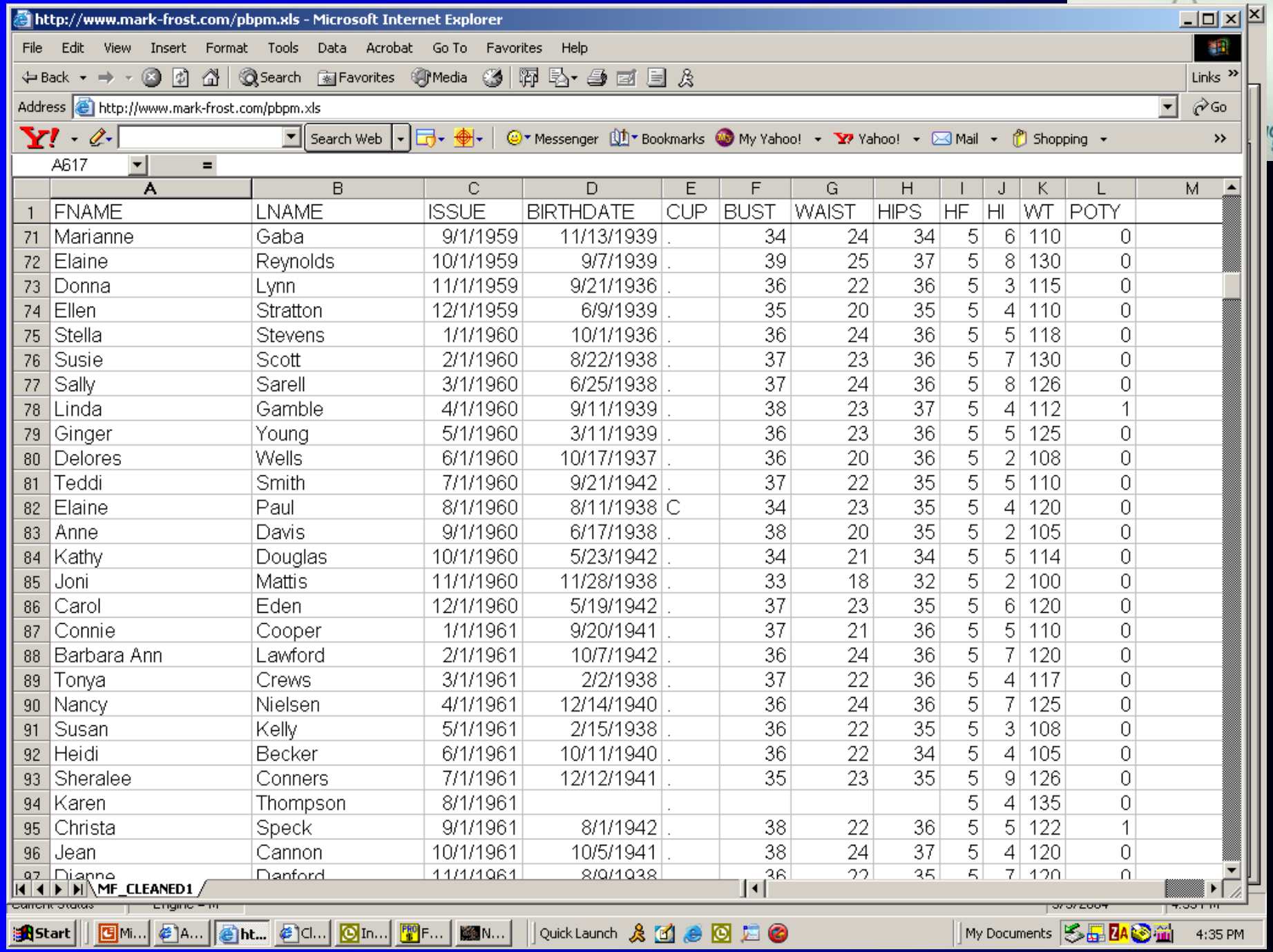


Winner $\sigma^2_U = 4$

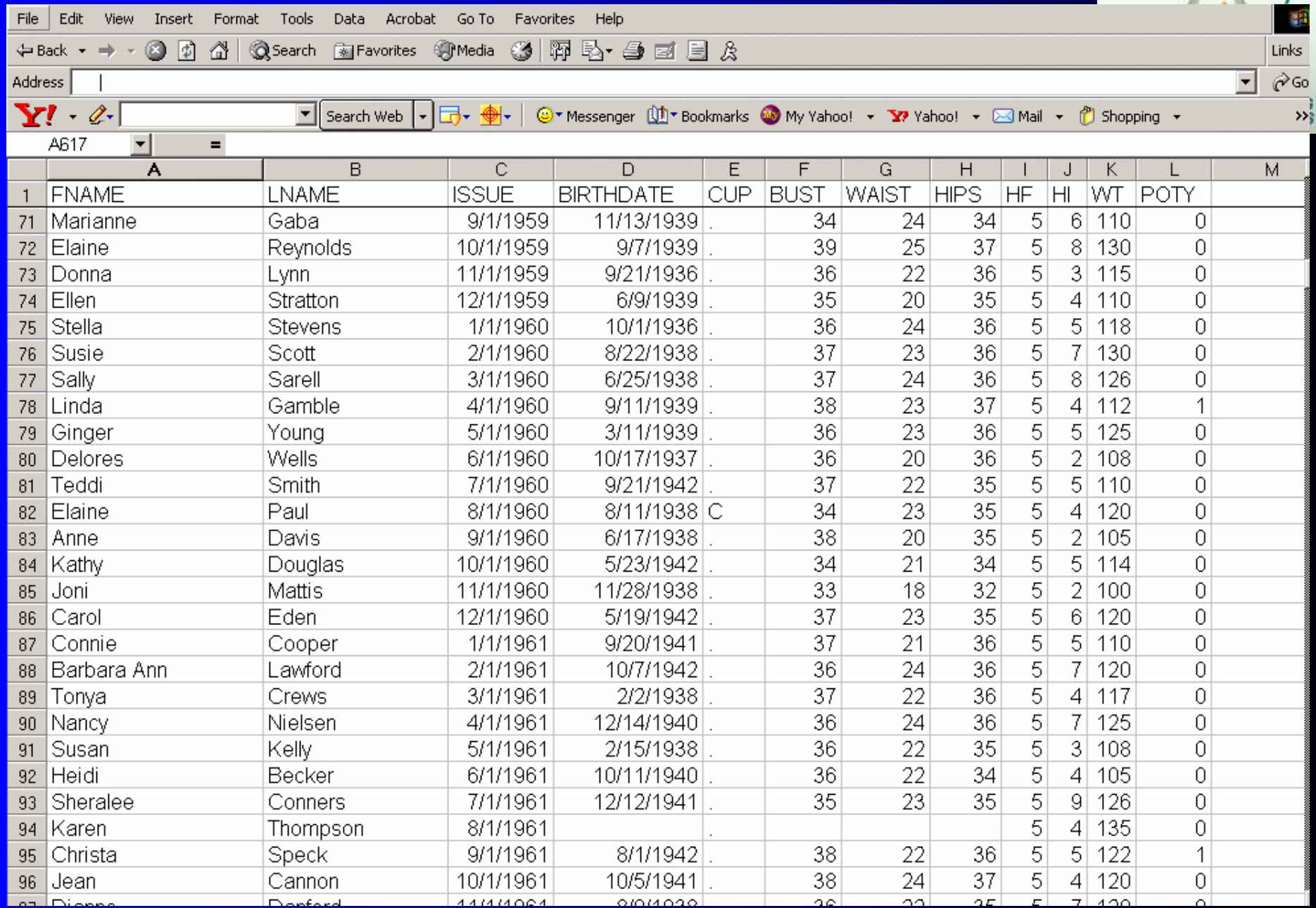
- We add the intervention variable into the model which then creates a new base case model. We can rerun the tournament and subsequent statistical testing to determine if a second intervention variable is needed. This process can be continued until no more variables are added to the base case model.

A Number of Leading Econometricians have been having serious discussions relating the Weights of Playboy Bunnies and the Economy …

So, I set out to get the historical weights for the Centerfold Bunnies and was able to locate the data on the web



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	FNAME	LNAME	ISSUE	BIRTHDATE	CUP	BUST	WAIST	HIPS	HF	HI	WT	POTY	
71	Marianne	Gaba	9/1/1959	11/13/1939	.	34	24	34	5	6	110	0	
72	Elaine	Reynolds	10/1/1959	9/7/1939	.	39	25	37	5	8	130	0	
73	Donna	Lynn	11/1/1959	9/21/1936	.	36	22	36	5	3	115	0	
74	Ellen	Stratton	12/1/1959	6/9/1939	.	35	20	35	5	4	110	0	
75	Stella	Stevens	1/1/1960	10/1/1936	.	36	24	36	5	5	118	0	
76	Susie	Scott	2/1/1960	8/22/1938	.	37	23	36	5	7	130	0	
77	Sally	Sarell	3/1/1960	6/25/1938	.	37	24	36	5	8	126	0	
78	Linda	Gamble	4/1/1960	9/11/1939	.	38	23	37	5	4	112	1	
79	Ginger	Young	5/1/1960	3/11/1939	.	36	23	36	5	5	125	0	
80	Delores	Wells	6/1/1960	10/17/1937	.	36	20	36	5	2	108	0	
81	Teddi	Smith	7/1/1960	9/21/1942	.	37	22	35	5	5	110	0	
82	Elaine	Paul	8/1/1960	8/11/1938	C	34	23	35	5	4	120	0	
83	Anne	Davis	9/1/1960	6/17/1938	.	38	20	35	5	2	105	0	
84	Kathy	Douglas	10/1/1960	5/23/1942	.	34	21	34	5	5	114	0	
85	Joni	Mattis	11/1/1960	11/28/1938	.	33	18	32	5	2	100	0	
86	Carol	Eden	12/1/1960	5/19/1942	.	37	23	35	5	6	120	0	
87	Connie	Cooper	1/1/1961	9/20/1941	.	37	21	36	5	5	110	0	
88	Barbara Ann	Lawford	2/1/1961	10/7/1942	.	36	24	36	5	7	120	0	
89	Tonya	Crews	3/1/1961	2/2/1938	.	37	22	36	5	4	117	0	
90	Nancy	Nielsen	4/1/1961	12/14/1940	.	36	24	36	5	7	125	0	
91	Susan	Kelly	5/1/1961	2/15/1938	.	36	22	35	5	3	108	0	
92	Heidi	Becker	6/1/1961	10/11/1940	.	36	22	34	5	4	105	0	
93	Sheralee	Conners	7/1/1961	12/12/1941	.	35	23	35	5	9	126	0	
94	Karen	Thompson	8/1/1961		.				5	4	135	0	
95	Christa	Speck	9/1/1961	8/1/1942	.	38	22	36	5	5	122	1	
96	Jean	Cannon	10/1/1961	10/5/1941	.	38	24	37	5	4	120	0	
97	Dianne	Danford	11/1/1961	8/9/1938		36	22	35	5	7	120	0	


File Edit View Insert Format Tools Data Acrobat Go To Favorites Help

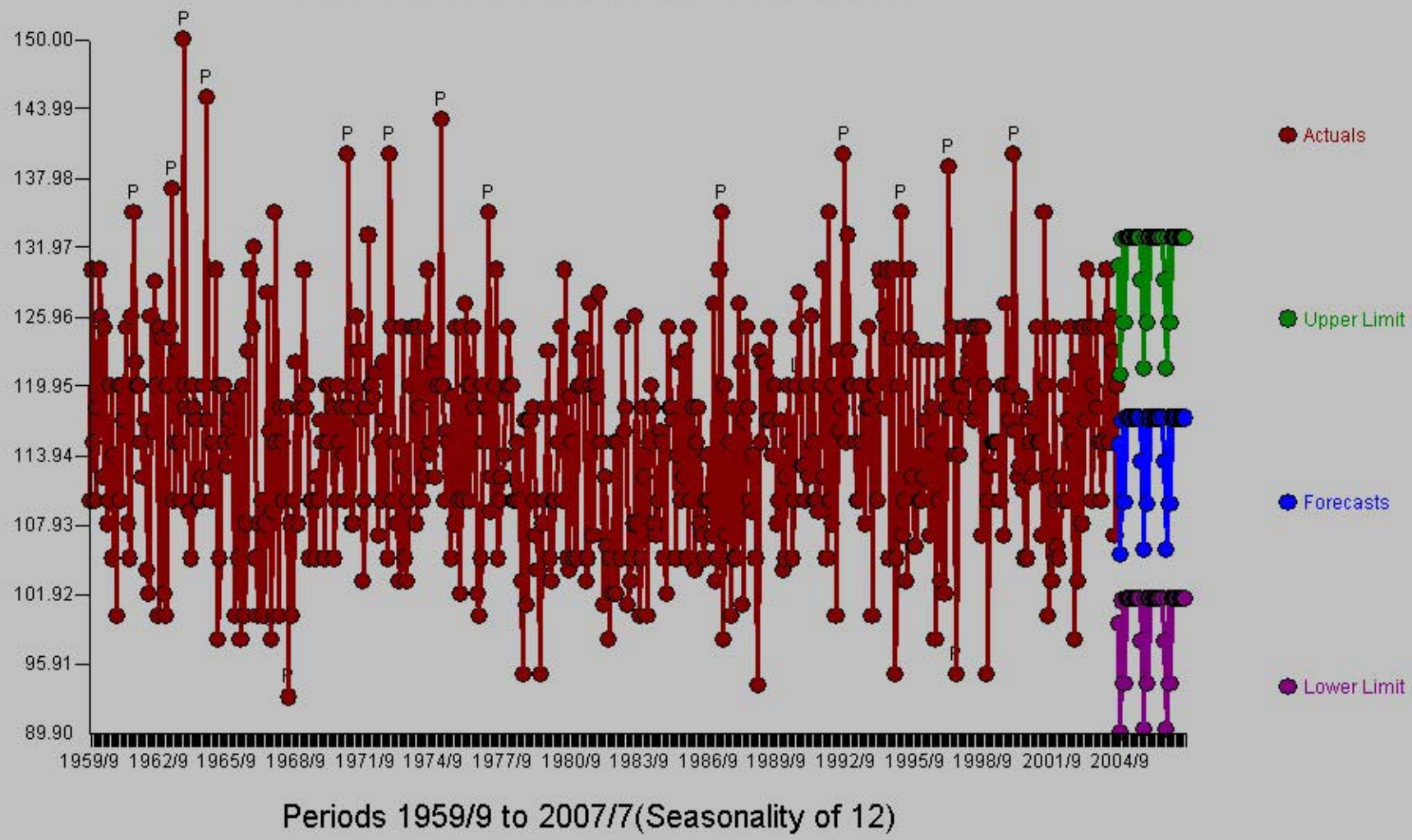
Back Search Favorites Media Links

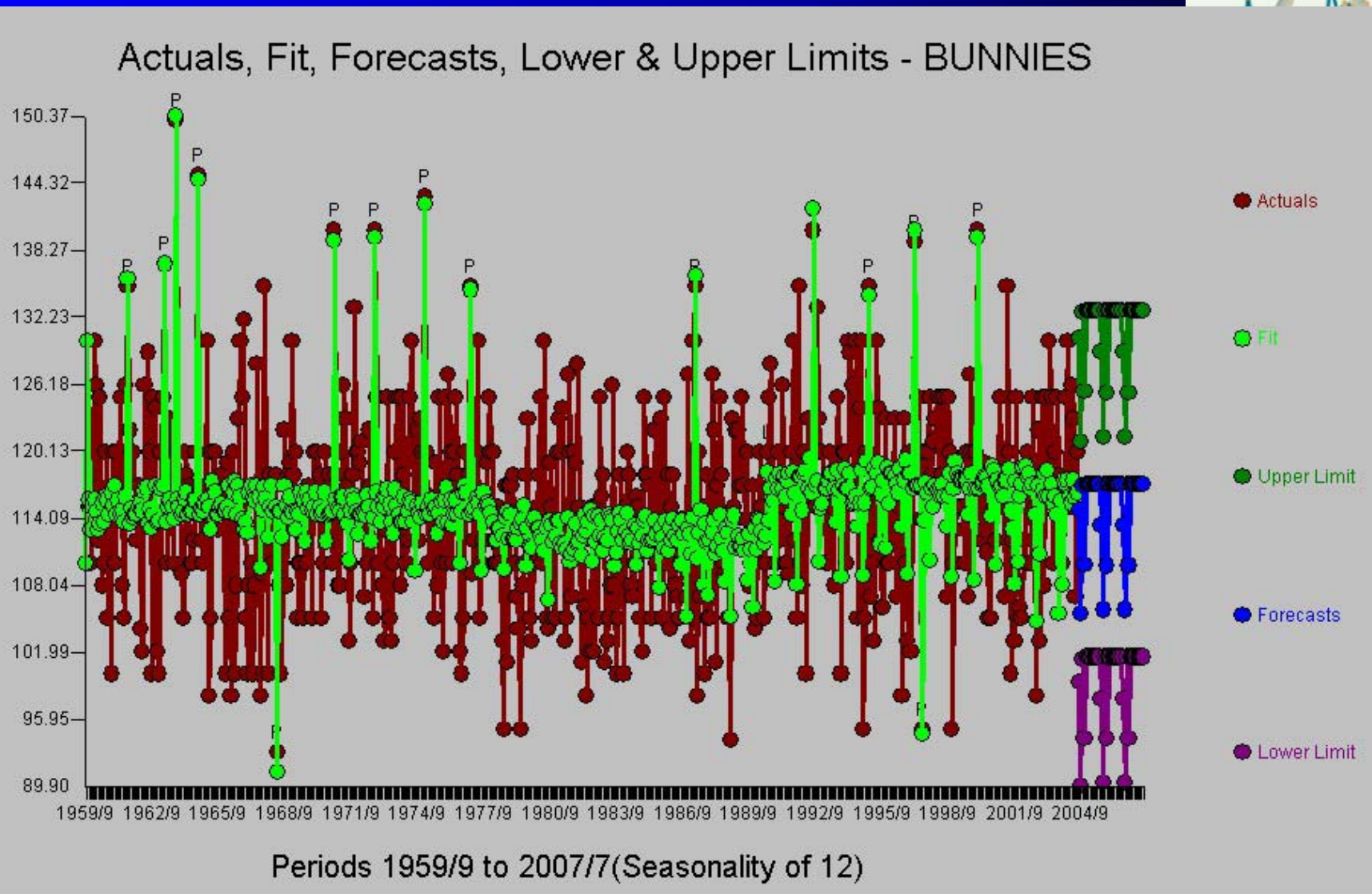
Address Go

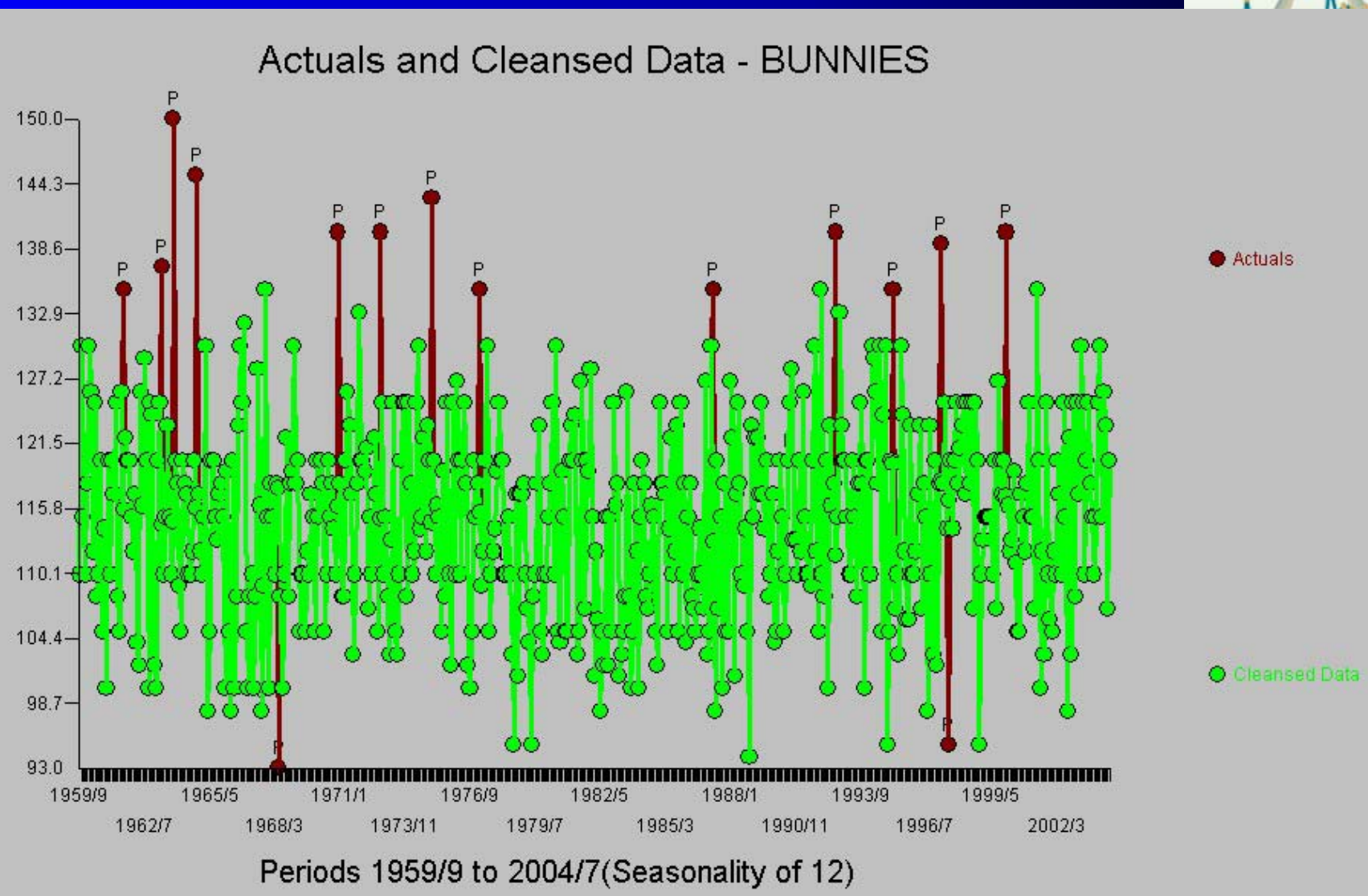
Search Web Messenger Bookmarks My Yahoo! Yahoo! Mail Shopping

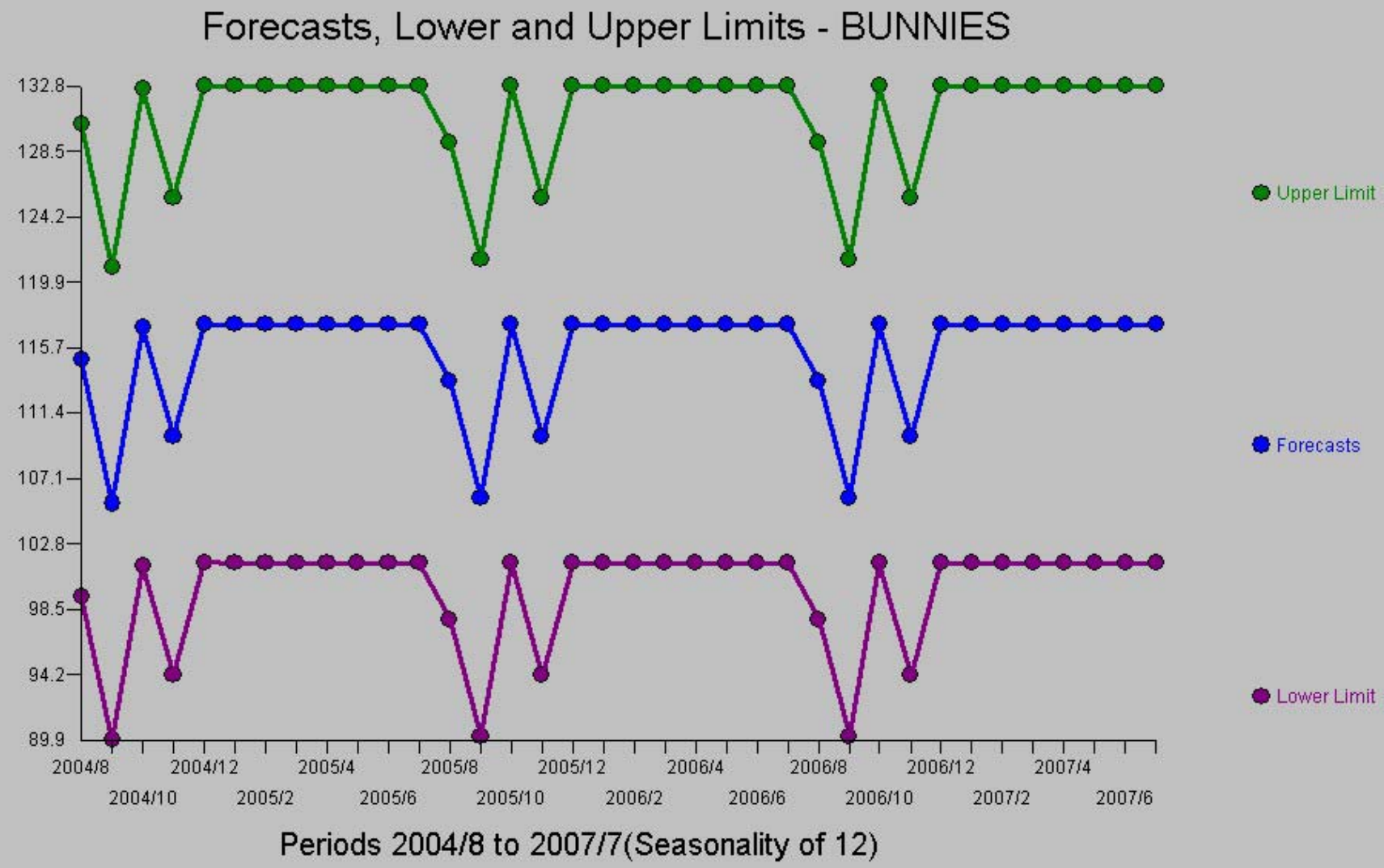
A617 =

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	FNAME	LNAME	ISSUE	BIRTHDATE	CUP	BUST	WAIST	HIPS	HF	HI	WT	POTY	
71	Marianne	Gaba	9/1/1959	11/13/1939	.	34	24	34	5	6	110	0	
72	Elaine	Reynolds	10/1/1959	9/7/1939	.	39	25	37	5	8	130	0	
73	Donna	Lynn	11/1/1959	9/21/1936	.	36	22	36	5	3	115	0	
74	Ellen	Stratton	12/1/1959	6/9/1939	.	35	20	35	5	4	110	0	
75	Stella	Stevens	1/1/1960	10/1/1936	.	36	24	36	5	5	118	0	
76	Susie	Scott	2/1/1960	8/22/1938	.	37	23	36	5	7	130	0	
77	Sally	Sarell	3/1/1960	6/25/1938	.	37	24	36	5	8	126	0	
78	Linda	Gamble	4/1/1960	9/11/1939	.	38	23	37	5	4	112	1	
79	Ginger	Young	5/1/1960	3/11/1939	.	36	23	36	5	5	125	0	
80	Delores	Wells	6/1/1960	10/17/1937	.	36	20	36	5	2	108	0	
81	Teddi	Smith	7/1/1960	9/21/1942	.	37	22	35	5	5	110	0	
82	Elaine	Paul	8/1/1960	8/11/1938	C	34	23	35	5	4	120	0	
83	Anne	Davis	9/1/1960	6/17/1938	.	38	20	35	5	2	105	0	
84	Kathy	Douglas	10/1/1960	5/23/1942	.	34	21	34	5	5	114	0	
85	Joni	Mattis	11/1/1960	11/28/1938	.	33	18	32	5	2	100	0	
86	Carol	Eden	12/1/1960	5/19/1942	.	37	23	35	5	6	120	0	
87	Connie	Cooper	1/1/1961	9/20/1941	.	37	21	36	5	5	110	0	
88	Barbara Ann	Lawford	2/1/1961	10/7/1942	.	36	24	36	5	7	120	0	
89	Tonya	Crews	3/1/1961	2/2/1938	.	37	22	36	5	4	117	0	
90	Nancy	Nielsen	4/1/1961	12/14/1940	.	36	24	36	5	7	125	0	
91	Susan	Kelly	5/1/1961	2/15/1938	.	36	22	35	5	3	108	0	
92	Heidi	Becker	6/1/1961	10/11/1940	.	36	22	34	5	4	105	0	
93	Sheralee	Conners	7/1/1961	12/12/1941	.	35	23	35	5	9	126	0	
94	Karen	Thompson	8/1/1961		.				5	4	135	0	
95	Christa	Speck	9/1/1961	8/1/1942	.	38	22	36	5	5	122	1	
96	Jean	Cannon	10/1/1961	10/5/1941	.	38	24	37	5	4	120	0	


Actuals and Forecasts - BUNNIES
150.00
143.99
137.98
131.97
125.96
119.95
113.94
107.93
101.92
95.91
89.90
1959/9
1962/9
1965/9
1968/9
1971/9
1974/9
1977/9
1980/9
1983/9
1986/9
1989/9
1992/9
1995/9
1998/9
2001/9
2004/9
Actuals
Upper Limit
Forecasts
Lower Limit
Periods 1959/9 to 2007/7(Seasonality of 12)


Actuals, Fit, Forecasts, Lower & Upper Limits - BUNNIES
150.37
144.32
138.27
132.23
126.18
120.13
114.09
108.04
101.99
95.95
89.90
1959/9
1962/9
1965/9
1968/9
1971/9
1974/9
1977/9
1980/9
1983/9
1986/9
1989/9
1992/9
1995/9
1998/9
2001/9
2004/9
Actuals
Fit
Upper Limit
Forecasts
Lower Limit
Periods 1959/9 to 2007/7(Seasonality of 12)


Actuals and Cleansed Data - BUNNIES
150.0
144.3
138.6
132.9
127.2
121.5
115.8
110.1
104.4
98.7
93.0
1959/9
1962/7
1965/5
1968/3
1971/1
1973/11
1976/9
1979/7
1982/5
1985/3
1988/1
1990/11
1993/9
1996/7
1999/5
2002/3
Actuals
Cleansed Data
Periods 1959/9 to 2004/7(Seasonality of 12)

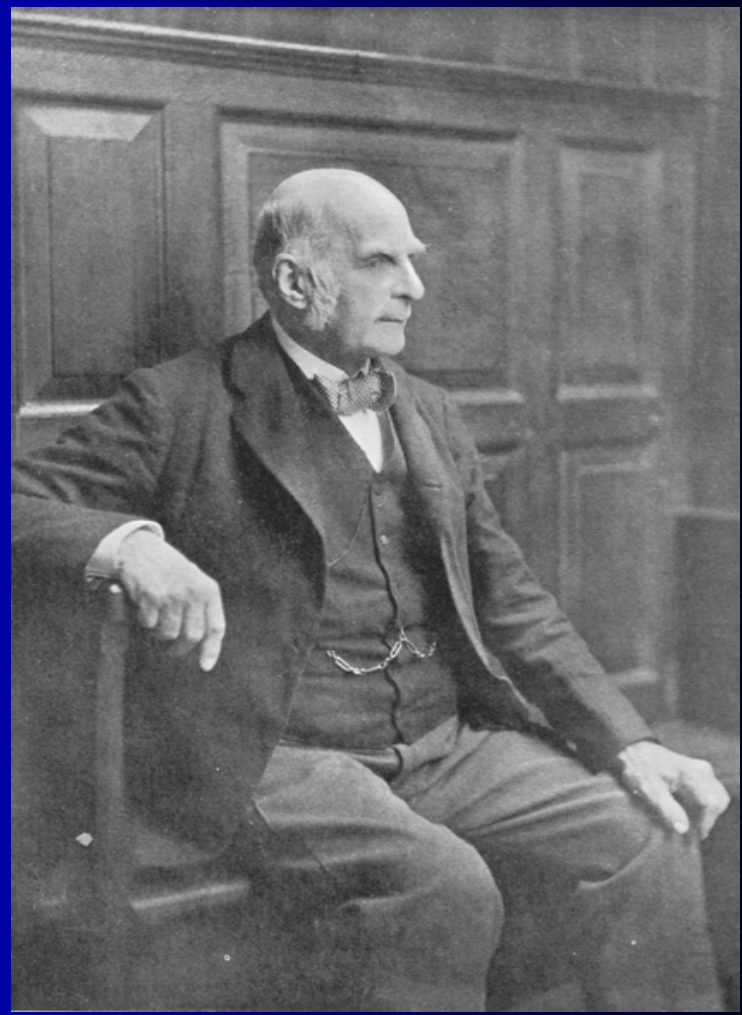

Forecasts, Lower and Upper Limits - BUNNIES
132.8
128.5
124.2
119.9
115.7
111.4
107.1
102.8
98.5
94.2
89.9
2004/8
2004/10
2004/12
2005/2
2005/4
2005/6
2005/8
2005/10
2005/12
2006/2
2006/4
2006/6
2006/8
2006/10
2006/12
2007/2
2007/4
2007/6
Upper Limit
Forecasts
Lower Limit
Periods 2004/8 to 2007/7(Seasonality of 12)

Contact Information

- Automatic Forecasting Systems, Inc. (AFS)
  P.O. Box 563
  Hatboro, PA 19040
  Phone: 215-675-0652
  Fax: 215-672-2534
  email: sales@Autobox.com
  Web Site: www.Autobox.com

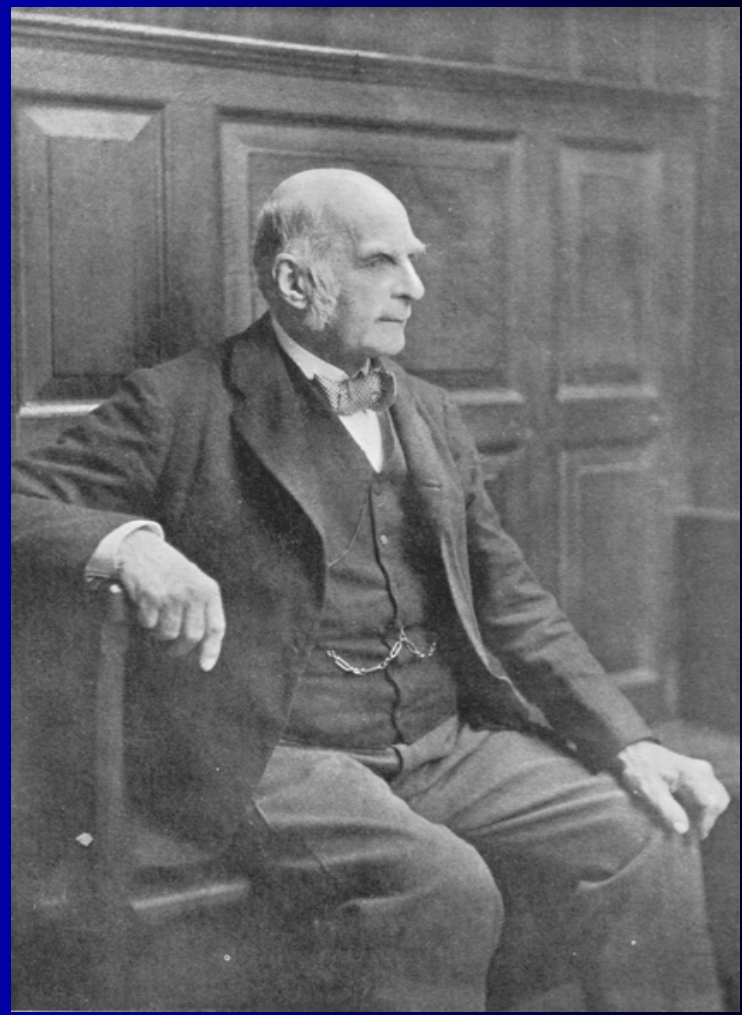
Sir Francis Galton

- *Webster's Third International Dictionary* defines *racism* as "the assumption that psycho cultural traits and capacities are determined by biological race and that races differ decisively from one another which is usually coupled with a belief in the inherent superiority of a particular race and its right to domination over others."



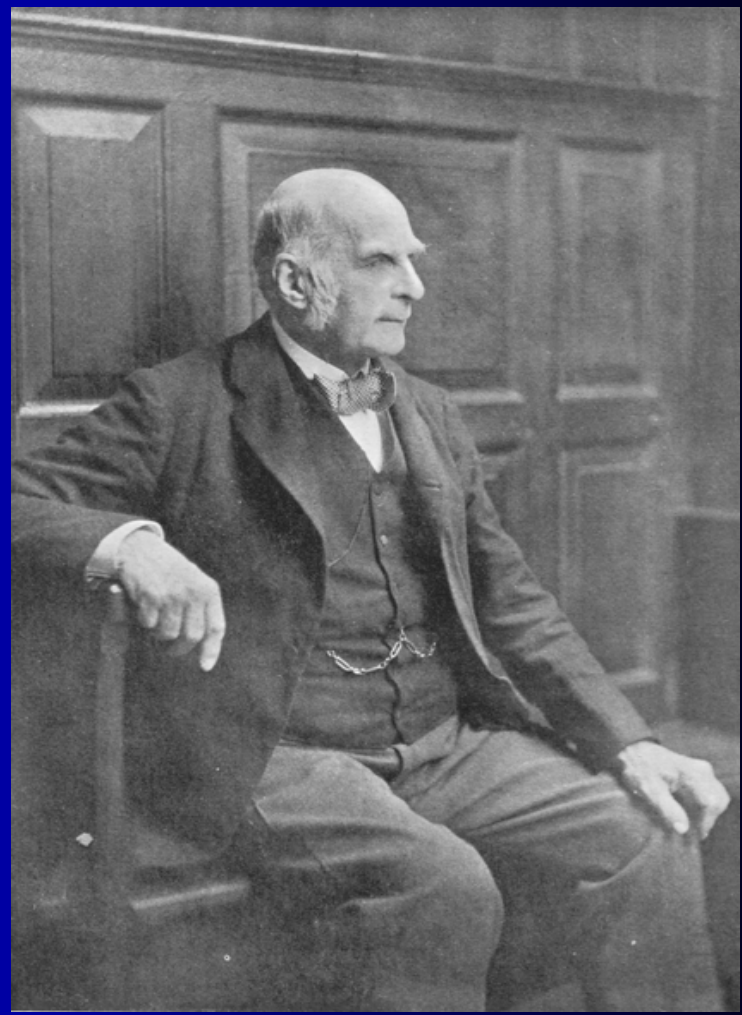
Sir Francis Galton

- From its inception the eugenics movement has embodied this definition of racism. Francis Galton was the father of the eugenics movement and widely known as The Father of Regression .



Sir Francis Galton

- He believed, without citing any evidence, that there was "a difference of not less than two grades between the black and white races..." (Galton, 1869/1962, p.394).



References

Hald, Anders. *A history of Mathematical Statistics from 1750 to 1930*. John Wiley & Sons, NY, 1998.

Stigler, Steven. *The history of Statistics: The measurement of Uncertainty before 1900*. Harvard University Press. 1986.

Fisher, Ronald. *Statistical Methods for Research Workers. 14th edition*. Oliver and Boyd, Edinburgh, 1970.

Stanton, Jeffrey. *Galton, Pearson, and the Peas: A brief History of Linear regression for statistics instructors.* http://www.amstat.org/publications/jse/v9n3/stanton.html