

The AUTOBOX Advantage

David P. Reilly
Automatic Forecasting Systems Inc.
P.O. Box 563
Hatboro, Pa 19040
215-675-0652
<http://www.autobox.com>

Forecasting is...

The use of existing or historical (quantitative) information related to the causes, behavior and/or performance of processes to predict future values.

All forecasts are educated guesses and planning tools

All Models are wrong but some are useful (G.E.P. Box)

**Forecasting The Future Is More
Difficult Than Forecasting The Past !**



Hierarchical Structure

- Qualitative

 - Judgmental

 - Analogical

- Quantitative: Time Series Analysis

 - Causal Modeling

 - Smoothing or Memory Models

 - Trend Decomposition

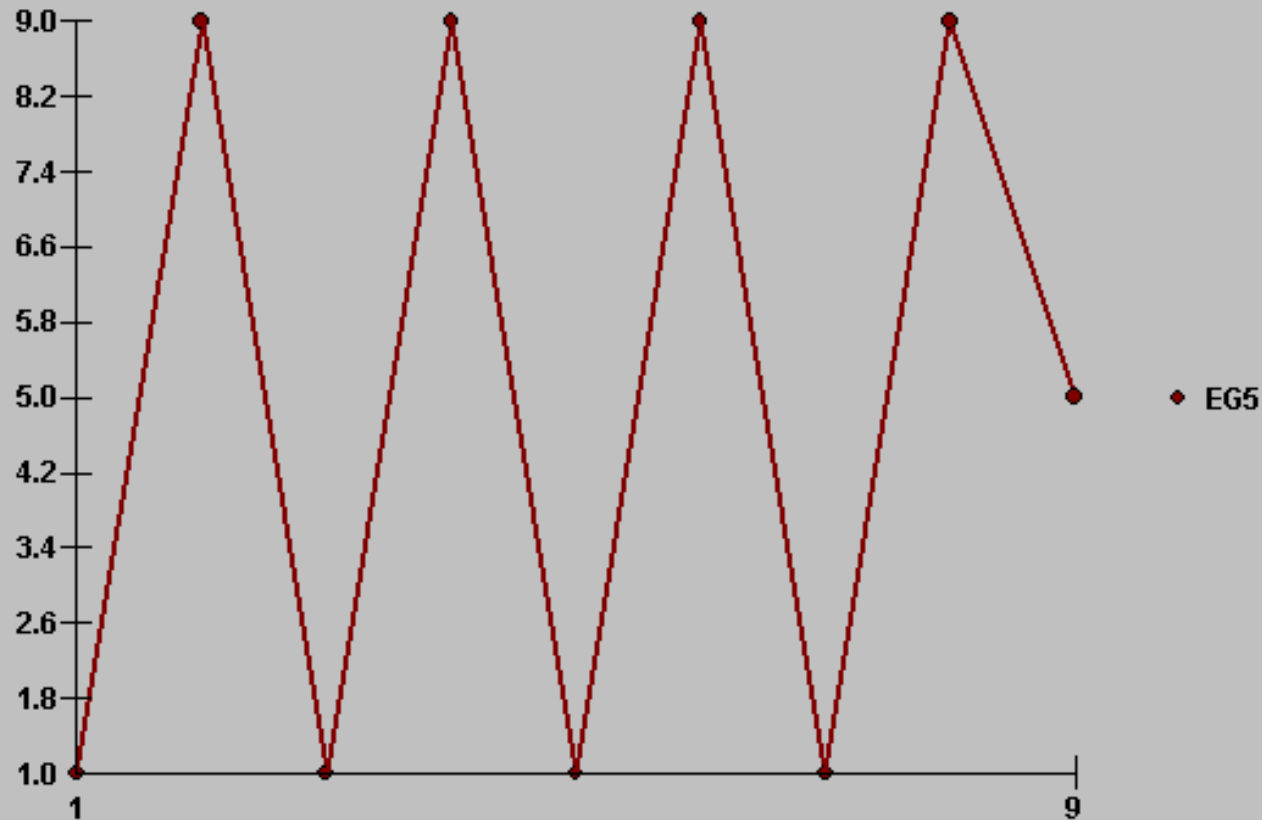
Forecasting patterns that have no historical precedent is not possible with Quantitative Methods. What is possible is to **detect that something unusual has occurred !**



How Would This Be Accomplished ?

By computing the **probability** of observing what was observed !

The mean can be unusual



**Periods From 1996/1 To 1996/9
(Seasonality:12)**

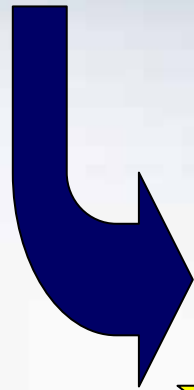
MaxVal 9.00000
MinVal 1.00000

Forecasting Models....

Model building based on data.

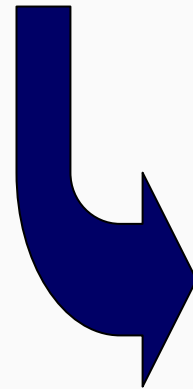


Data



Analysis

★ ★ **Forecasting** ★ ★



Information

Forecasting....



The purpose of forecasting is to summarize data so that information may be extracted from the data.

- **Data recorded sequentially through time is called “Time Series Data”.**
- **The analysis of time series data requires special mathematical techniques, called “Time Series Techniques”.**

Serious Disconnect between the Teaching and Practice of Statistics

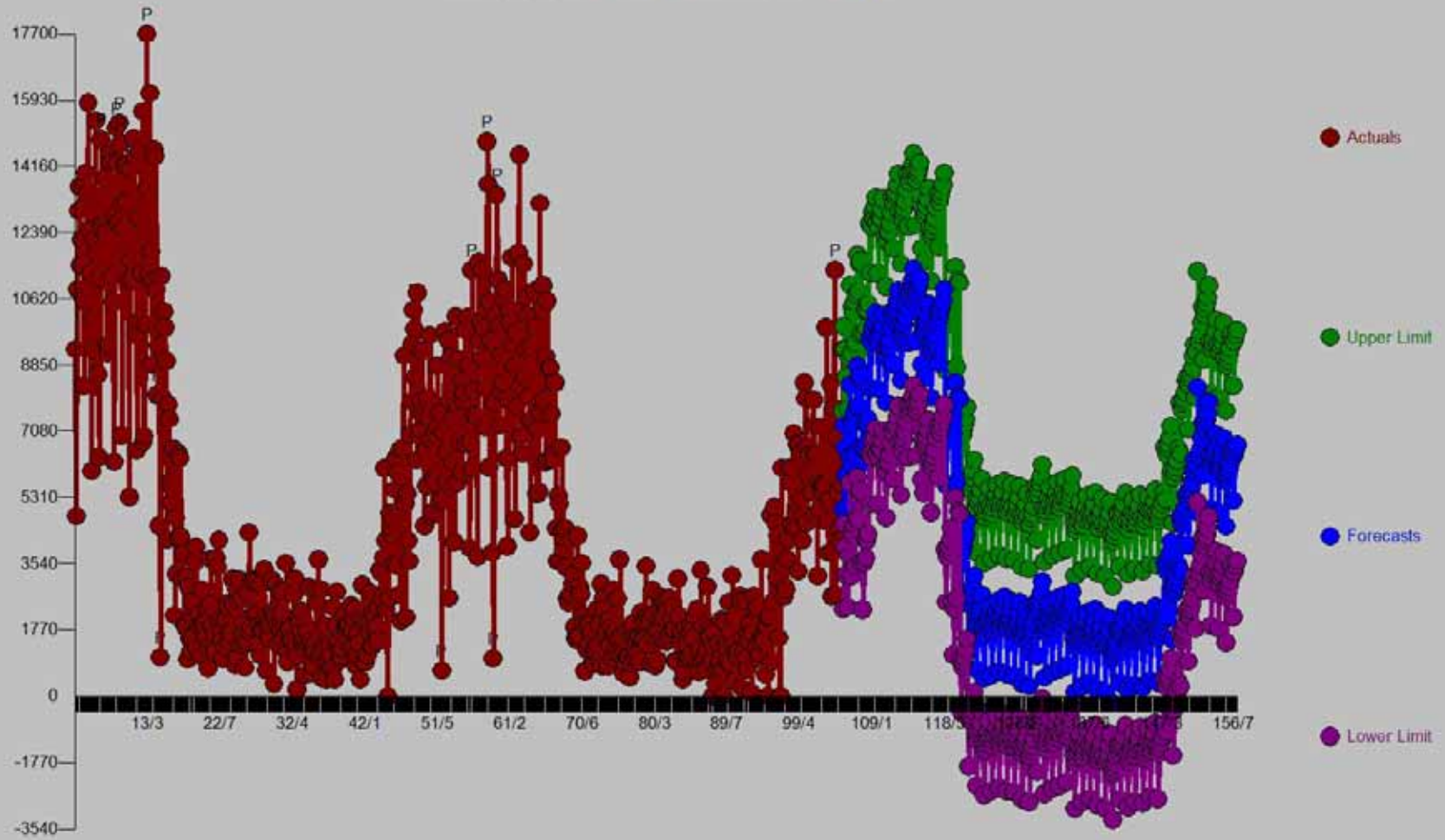
- 99.9% of all Academic presentation of statistical tools **REQUIRES** independent observations
- In time series data, this is clearly not the case

Statistical packages have enormous influence over analysis, especially over that of the **less sophisticated user**. There is a tendency for the user to do what is readily available in their software.

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 by determining the window of response around each user specified input.
- 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.

Actuals and Forecasts - dan



Periods 3/6 to 157/3(Seasonality of 7)

Two issues of Forecasting concern :

- **Early Warning Systems** detecting peculiar data (e.g. sudden onset of a seasonal structural change)
- **What if Analysis** (What is expected to happen if the weather forecast changes or if events are scheduled that have had historical impacts)

Quantitative: Time Series Analysis

Causal Modeling

Smoothing or Memory Models

Trend Decomposition

$Y_t = \text{Known Events} + \text{Previous Values of } Y + \text{Dummy}$

Autobox Seamlessly Integrates All Three Into One Equation

C A U S A L M O D E L (Simple Regression)

$$Y_t = V_0 + V_1 X_{1t} + A_t$$

M E M O R Y M O D E L (3 period moving average)

$$Y_t = 1/3 Y_{t-1} + 1/3 Y_{t-2} + 1/3 Y_{t-3} + A_t$$

D U M M Y M O D E L (where P is a Pulse a period t-2)

$$Y_t = V_1 L + V_2 P_{t-2} + A_t$$

Causal



- Using Known Events or User-Suggested auxiliary or helping series and future expectations of these (i.e. holidays, pay-days, 1st of the month, weather etc.)

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

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

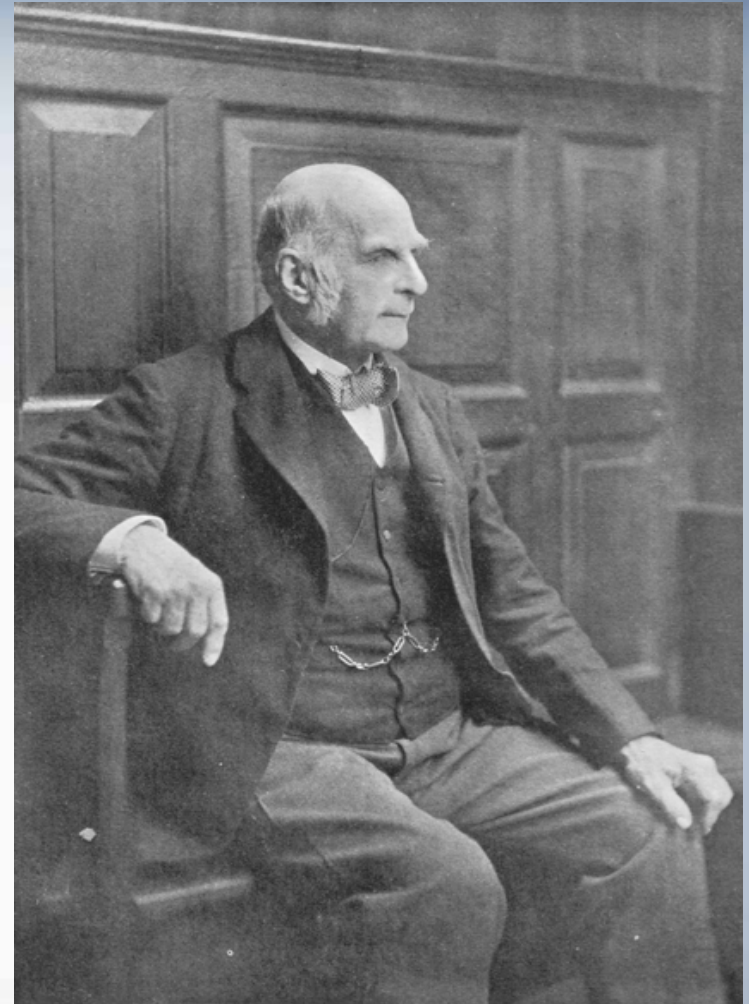


Historical Outline

- **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
- Eugenecist
- Statistician
- Anthropologist
- Criminologist
- Hereditarian
- Half-cousin of Charles Darwin
- Psychologist



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.

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)



Memory

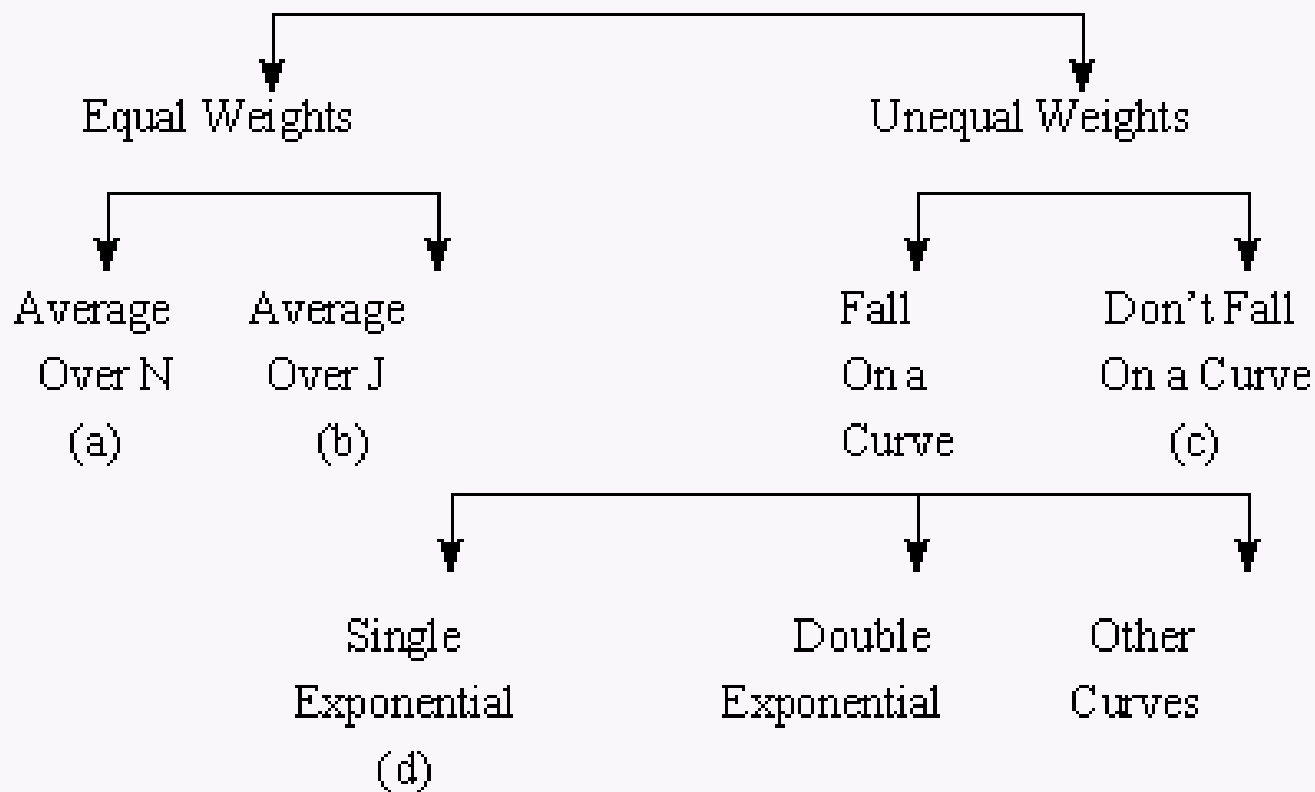


- Using historical values e.g. Cash Demand By Day for the last K periods. This raises the natural question of “How Much Data Should Be Used ” and “How Should It Be Used ?” .

The Memory Model Family Tree

Memory Model

BOX-JENKINS(ARIMA)



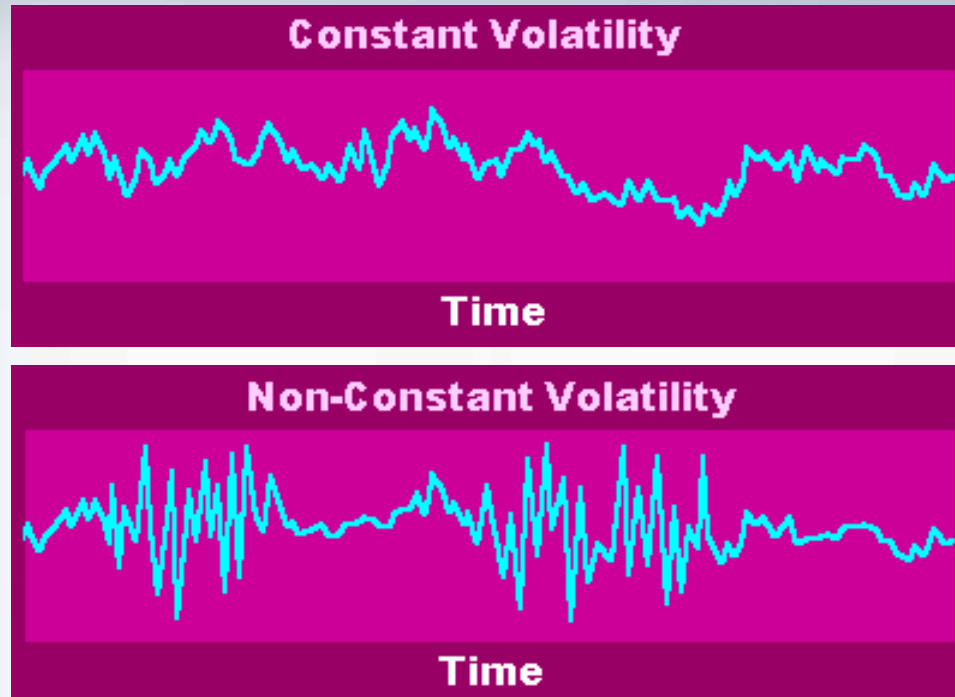
Three Questions



- 1. What is the optimal value of J (# of Lags)
- 2. What lag weights or coefficients should be applied to these J values and
- 3. Is there evidence of changing volatility that would suggest a transformation of the data be applied to history before assigning the lag weights

Homoscedasticity and Heteroscedasticity

Exhibit 1



The Family of Dummy Variables

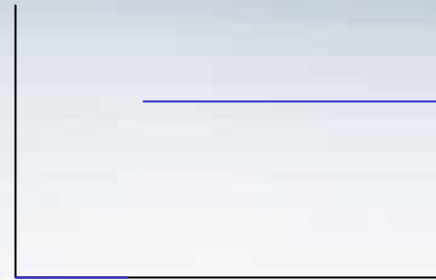
Pulse	$Z_t = 0,0,0,0,1,0,0,0$
Level Shift	$Z_t = 0,0,0,0,1,1,1,1,1, \dots$
Seasonal Pulse	$Z_t = 0,1,0,0,0,1,0,0,0,1, \dots$
Time Trend	$Z_t = 0,0,0,0,1,2,3,4,5, \dots$

The Family of Interventions

Intervention: Event leading to system response, characterized by type of response



Pulse



Level Shift



Seasonal Pulse



Time Trend

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:

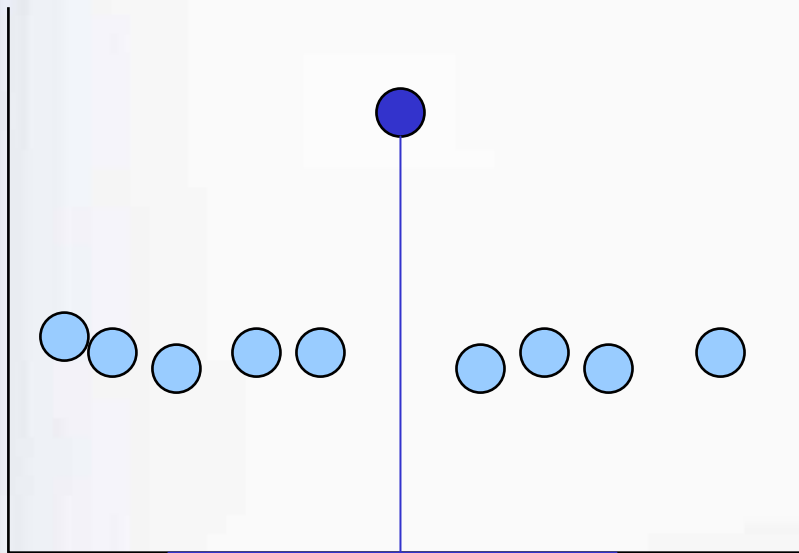
WHAT TO DO ABOUT THEM?

- 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.

Example of a Pulse Intervention

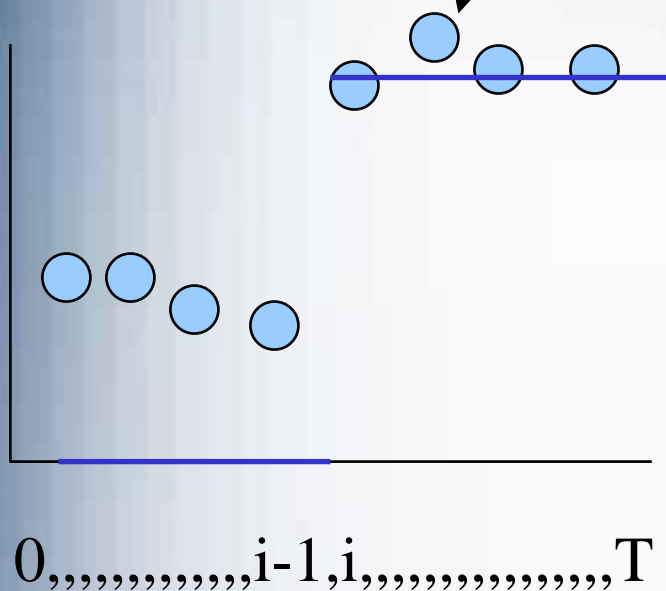
Z_t represents a pulse or a one-time intervention at time period 6.

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



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 $Y_t = B_0 + B_1 Z_t + U_t$



Assume the appropriate Z_t is
 $Z_t = 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, \dots, T$
 or $Z_t = 0 \quad t < i$
 $Z_t = 1 \quad t > i-1$ }

Modeling Interventions - Seasonal Pulses

There are other kinds of pulses that might need to be considered otherwise our model may be insufficient.

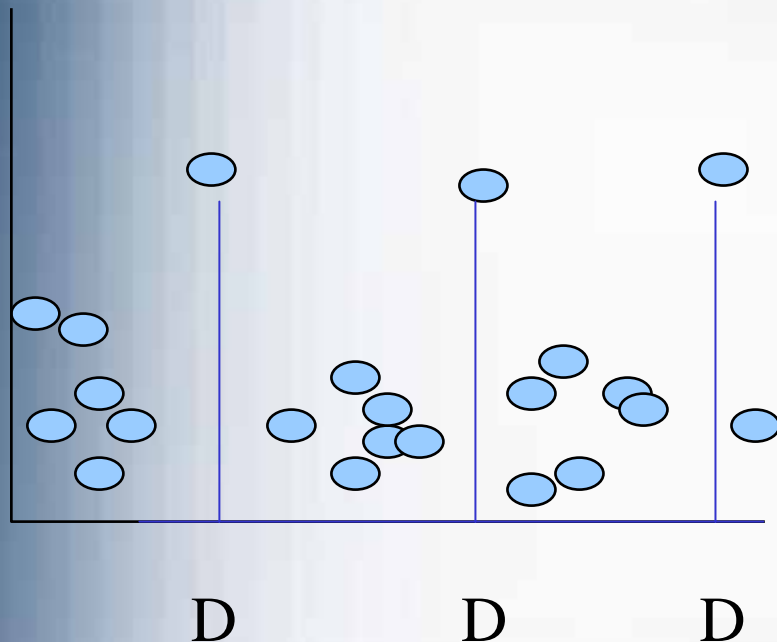
For example, December sales are high.

The data suggest this model

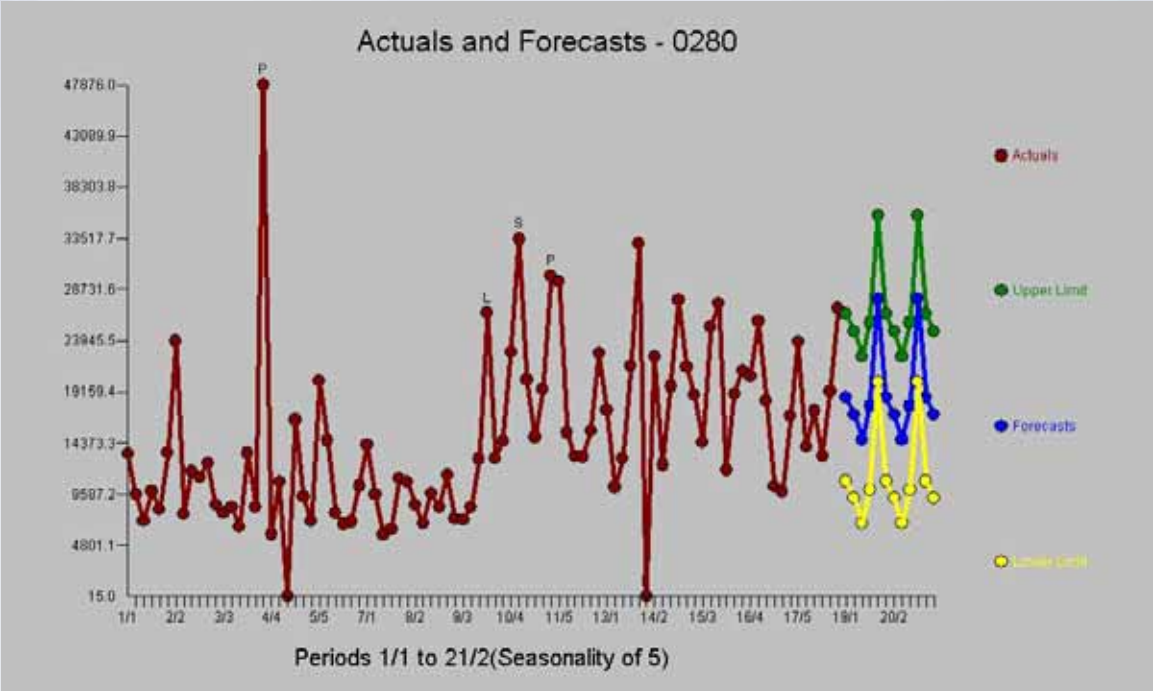
$$Y_t = B_0 + B_3 Z_t + U_t$$

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

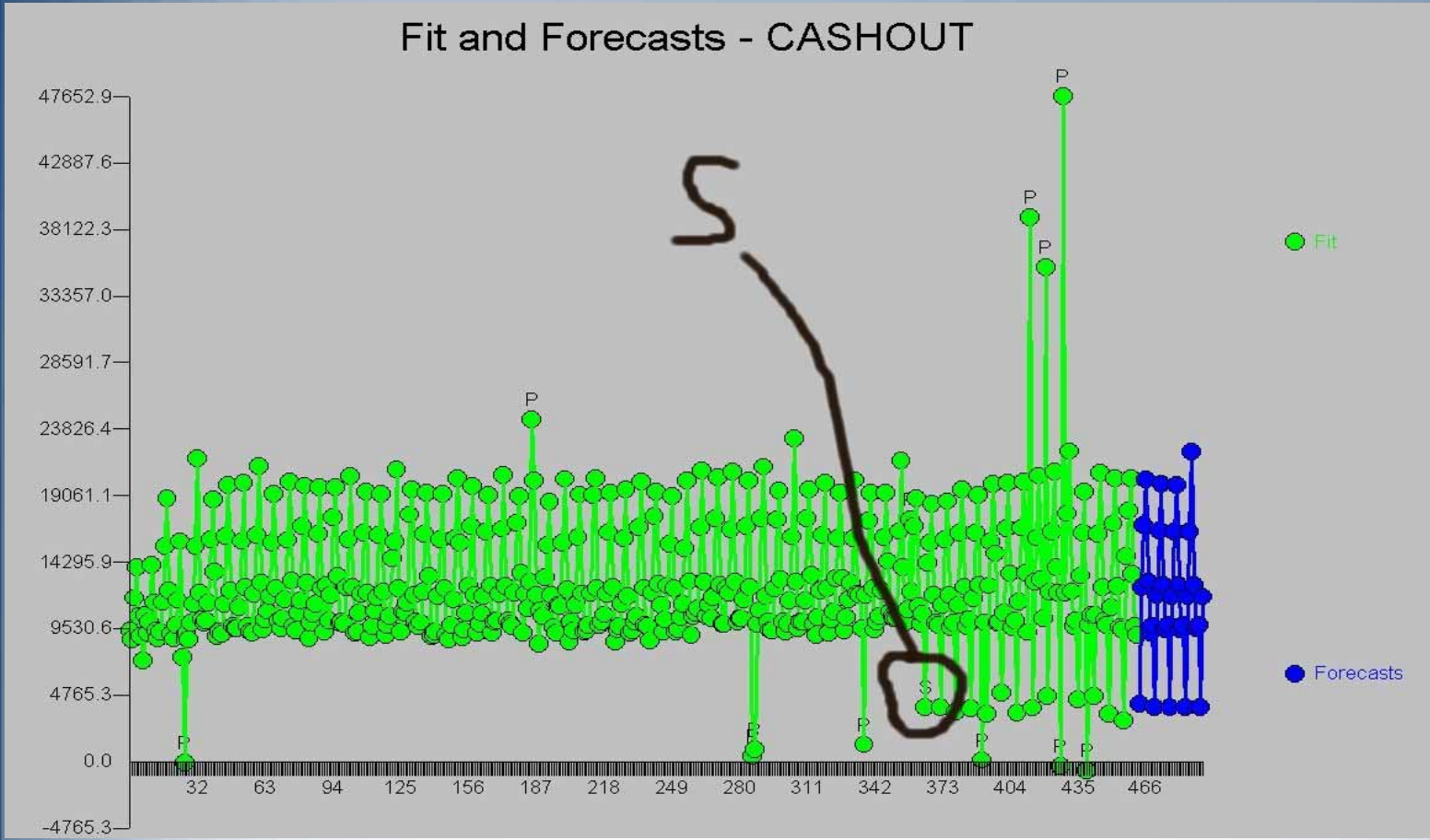
$$Z_t = 1 \quad i = 12, 24, 36, 48, 60$$



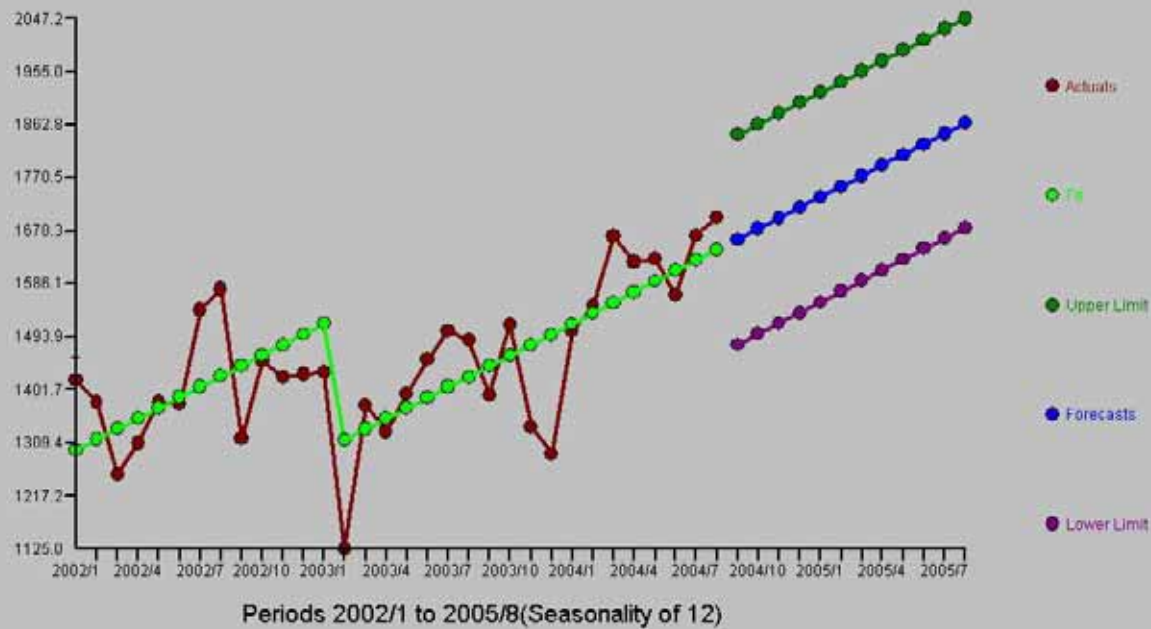
Detection of A Structural Change in the Daily Demand For Cash




Detection of A Structural Change in the Daily Demand For Cash



Actuals, Fit, Forecasts, Lower & Upper Limits - TOTAL INTAKES AUTOMATIC



$$Y_t = \text{Causals that you know about} + \text{Causals that you don't know about}$$


Future Value (at time t)
of Variable of Interest

$$Y_t = \text{Causals that you know about} \\ + \text{Memory} \\ + \text{Dummy Variables}$$

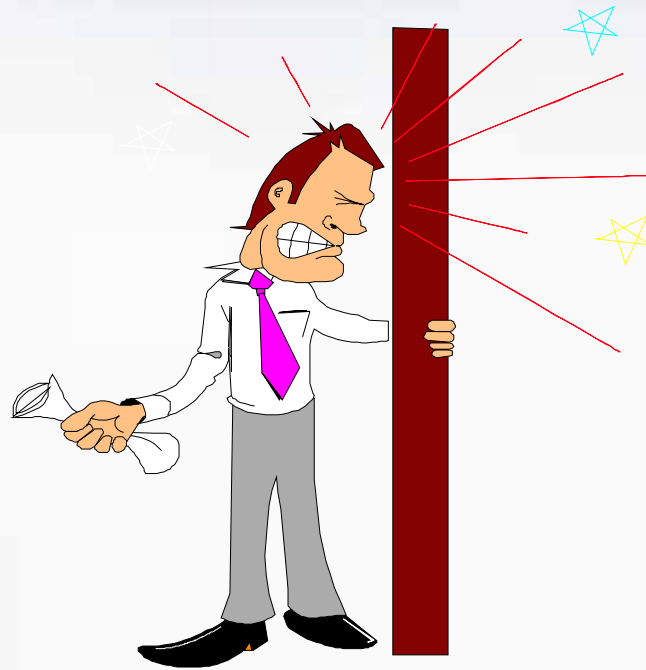
AUTOBOX Optimally Combines Three Kinds of Structures

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



Future Value (at time t) of
Variable of Interest

If only we had known sooner....



How Would This Be Accomplished ?

By computing the **probability** of observing what was observed !

Early Warning Systems

Find out why

Early warning systems should not simply detect high and low values, but should detect unusual activity inconsistent with expectations.

Historical development of Memory

Consider an “N Period” Equally Weighted Model

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

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

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.

$$Y(\text{new})=(1-a)*Y(\text{old})+a*\text{error}$$

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(α)

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

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 modelthus 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).



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.

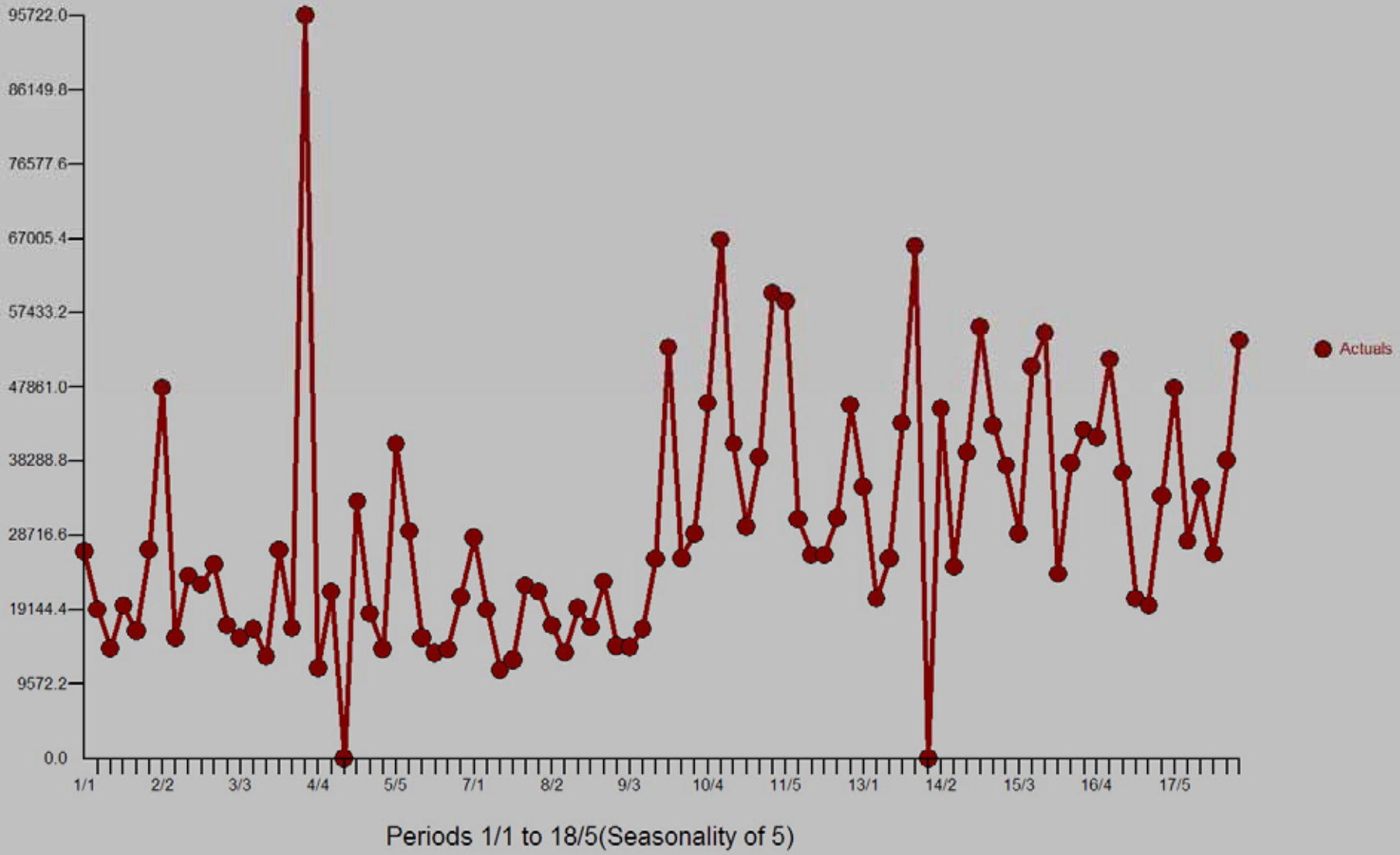
Reilly, D.P. (1980). "Experiences with an Automatic Box-Jenkins Modeling Algorithm," in *Time Series Analysis*, ed. O.D. Anderson. (Amsterdam: North-Holland), pp. 493-508.

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.

Actuals - CARR0280

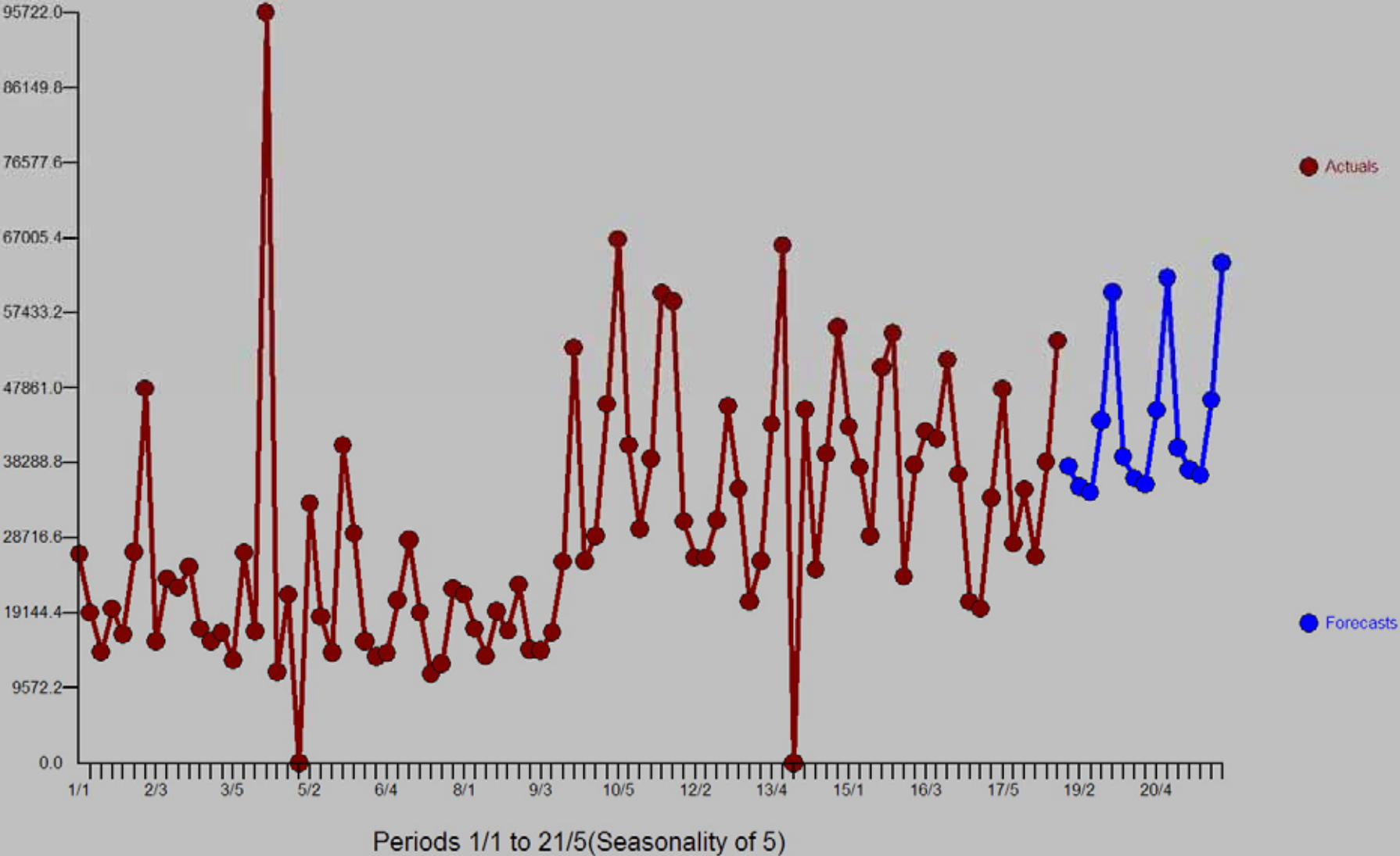


Previous Forecasting Tools Employed By Carreker (BMF*)

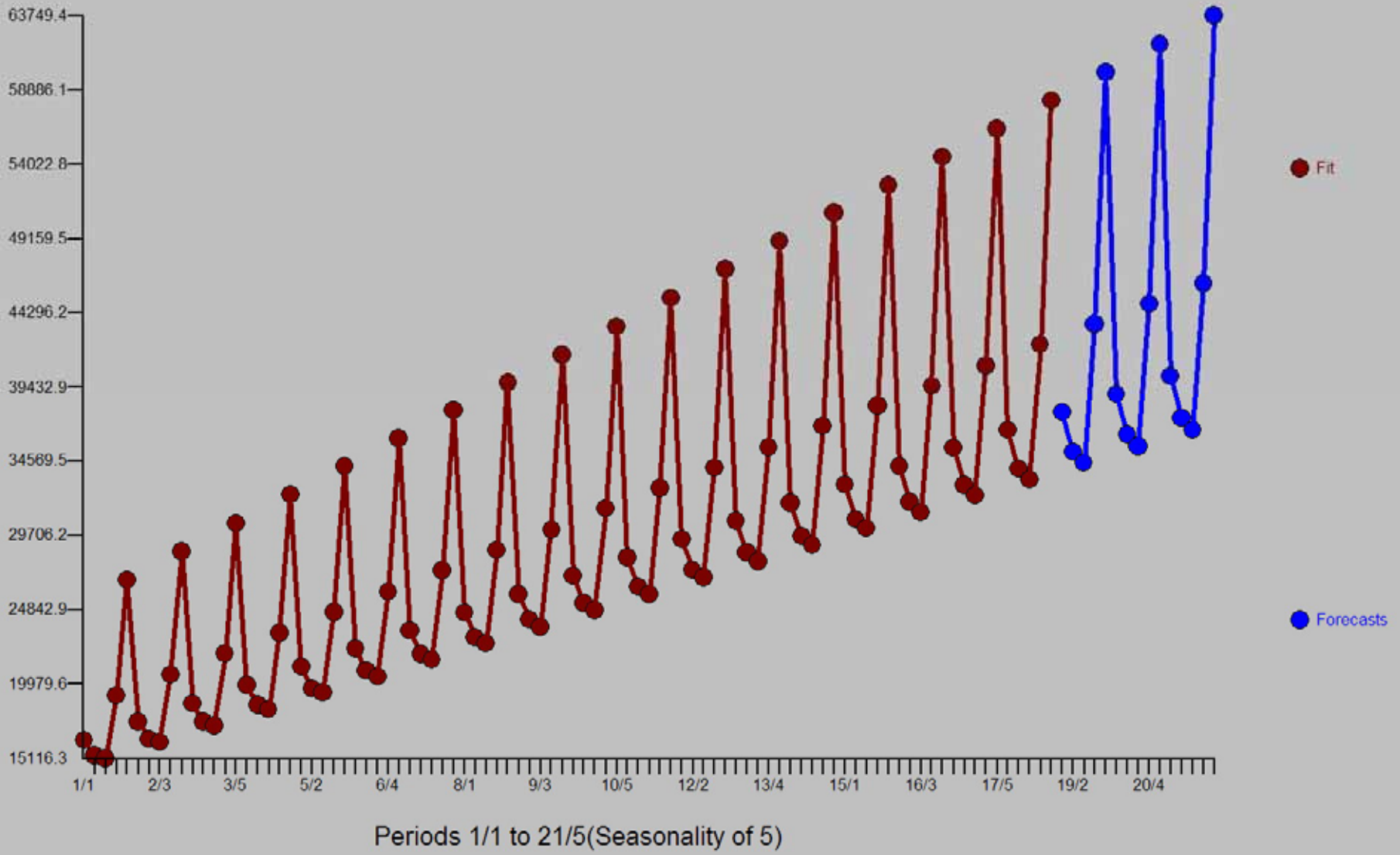
Also Known as The SAS Regime

***Before Mark Frost**

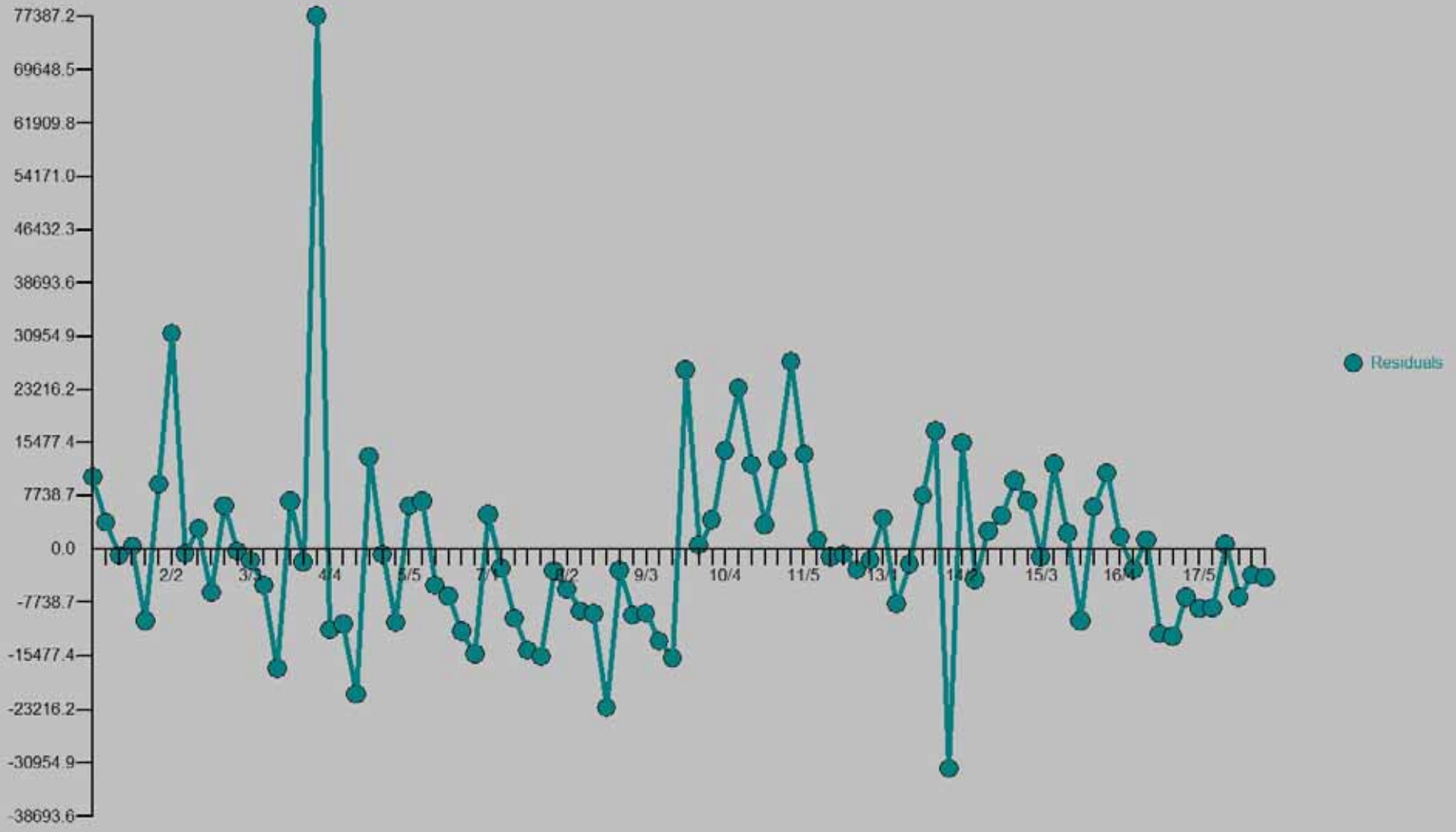
Actuals and Forecasts - CARR0280



Fit and Forecasts - CARR0280

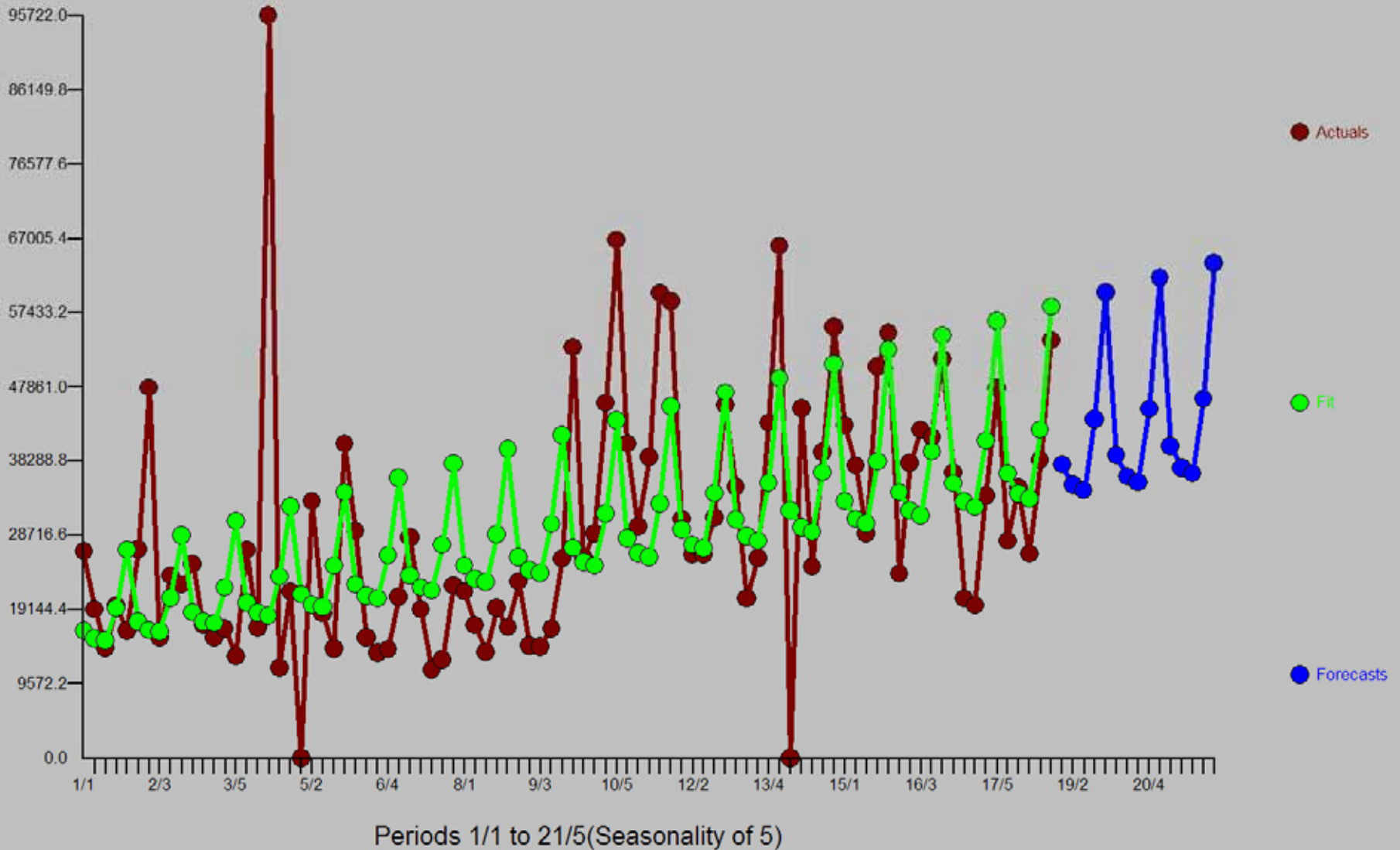


Residuals - CARR0280



Periods 1/1 to 18/5(Seasonality of 5)

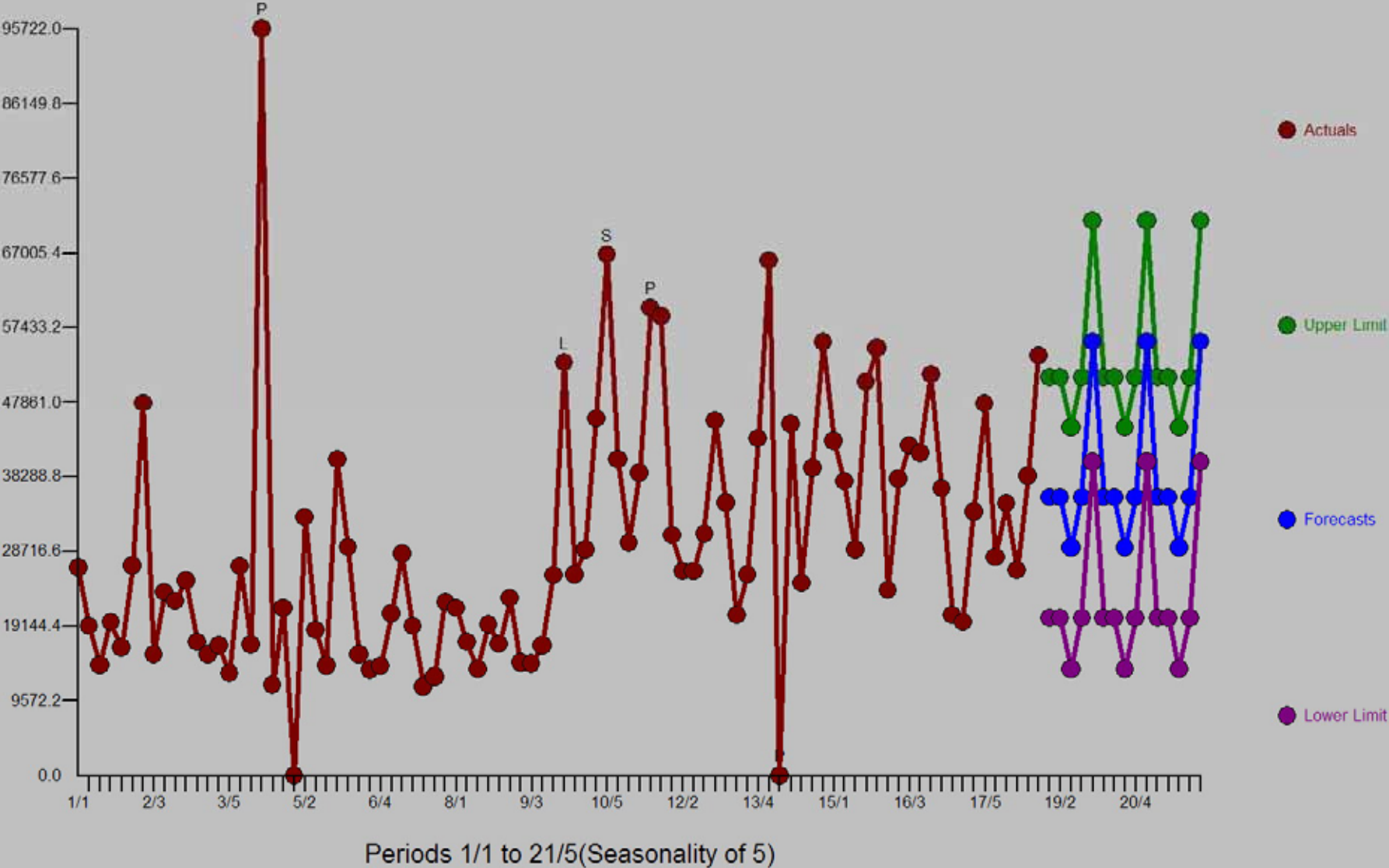
Actuals, Fit and Forecasts - CARR0280



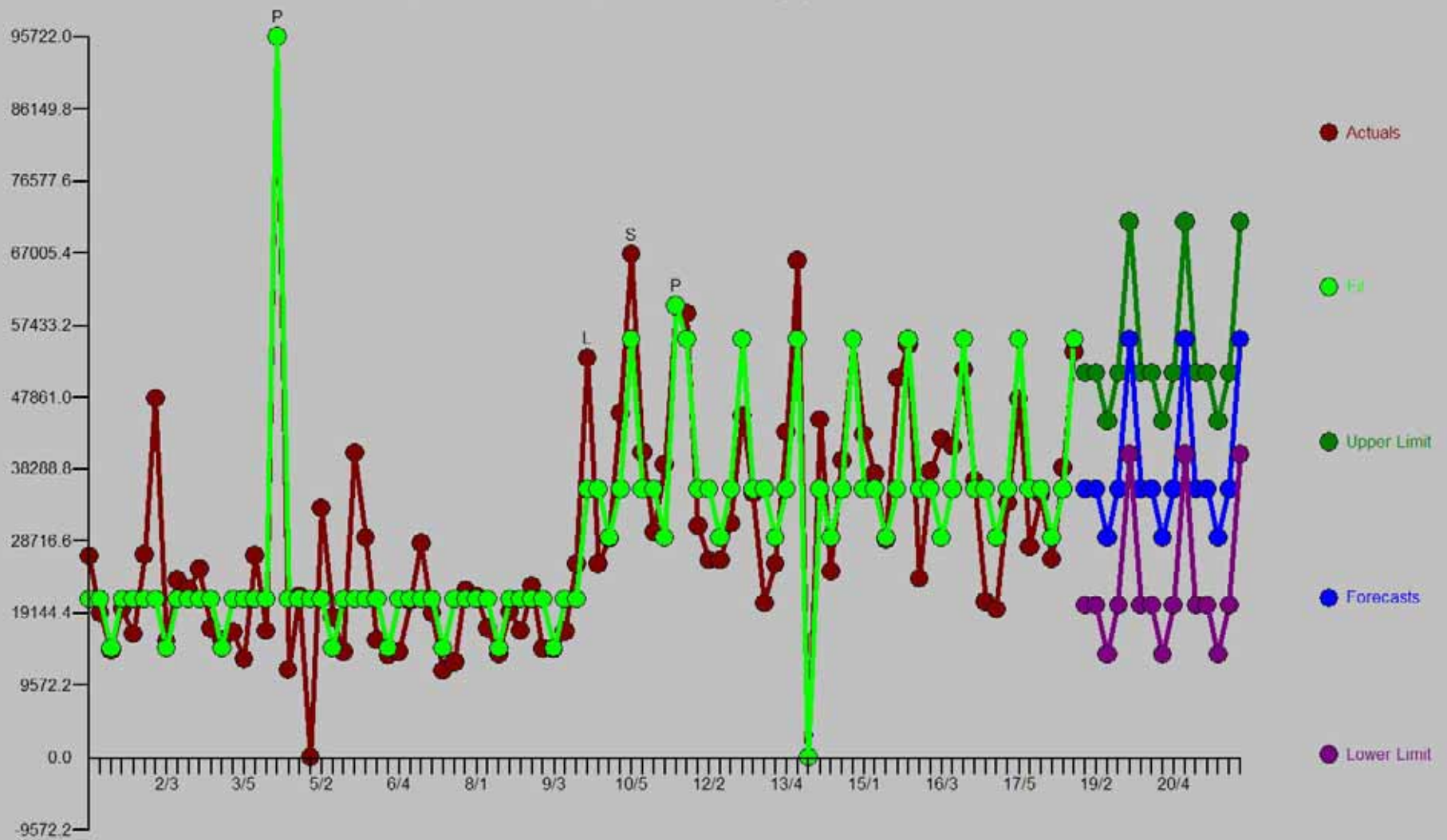
Current Forecasting Tools Employed By Carreker (BEM)

*Before Event Modelling

Actuals and Forecasts - CARR0280

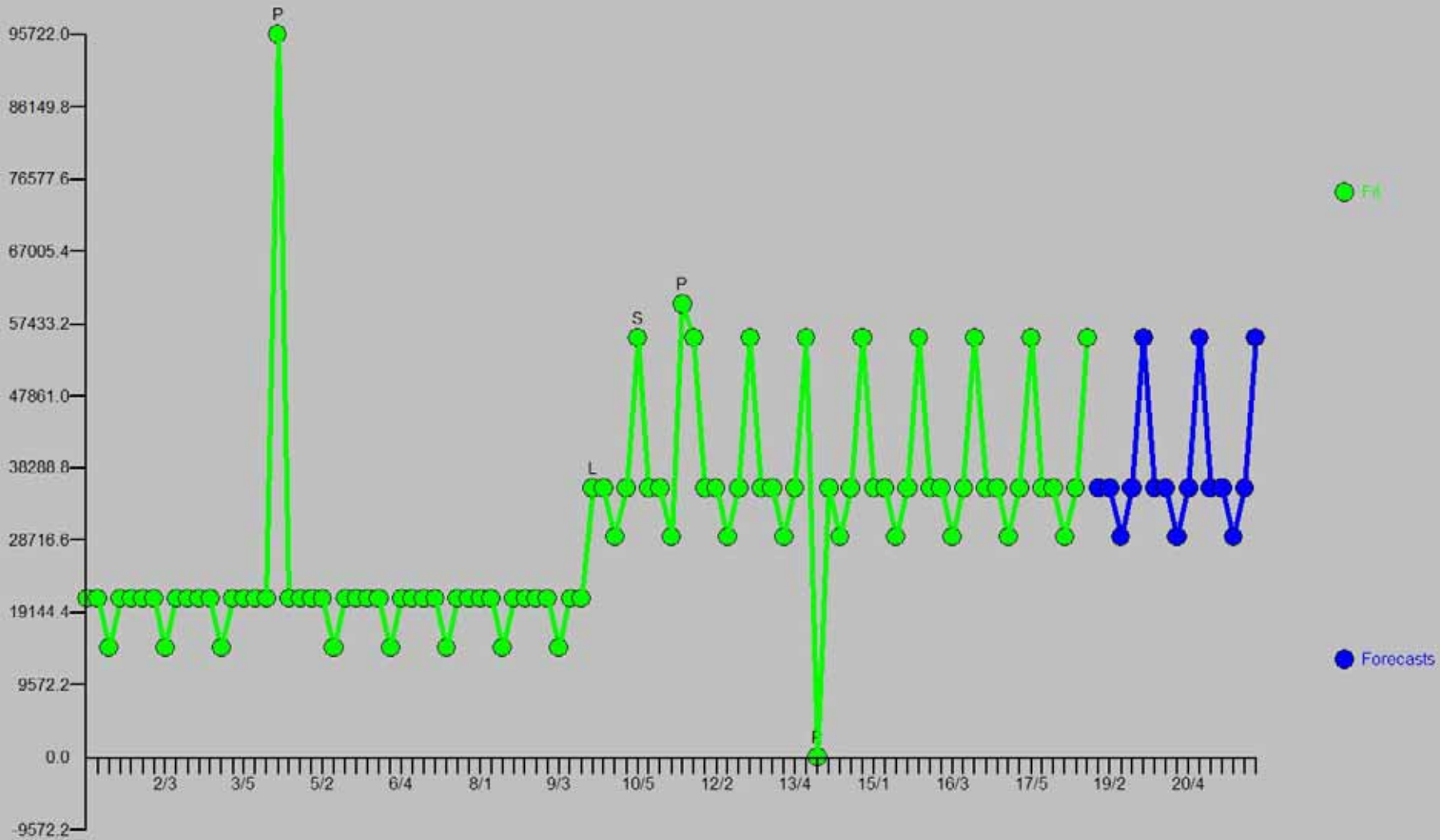


Actuals, Fit, Forecasts, Lower & Upper Limits - CARR0280



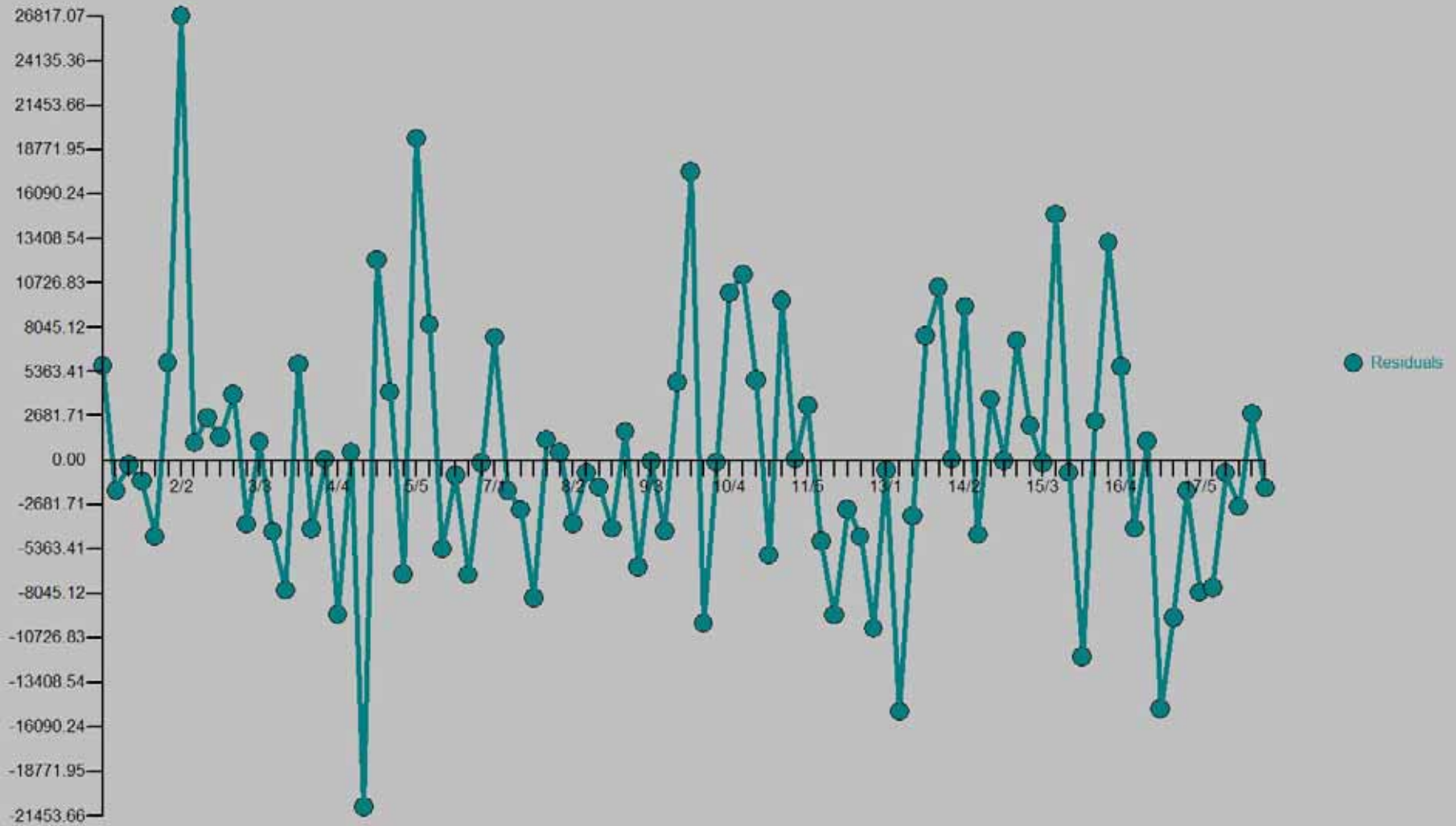
Periods 1/1 to 21/5(Seasonality of 5)

Fit and Forecasts - CARR0280



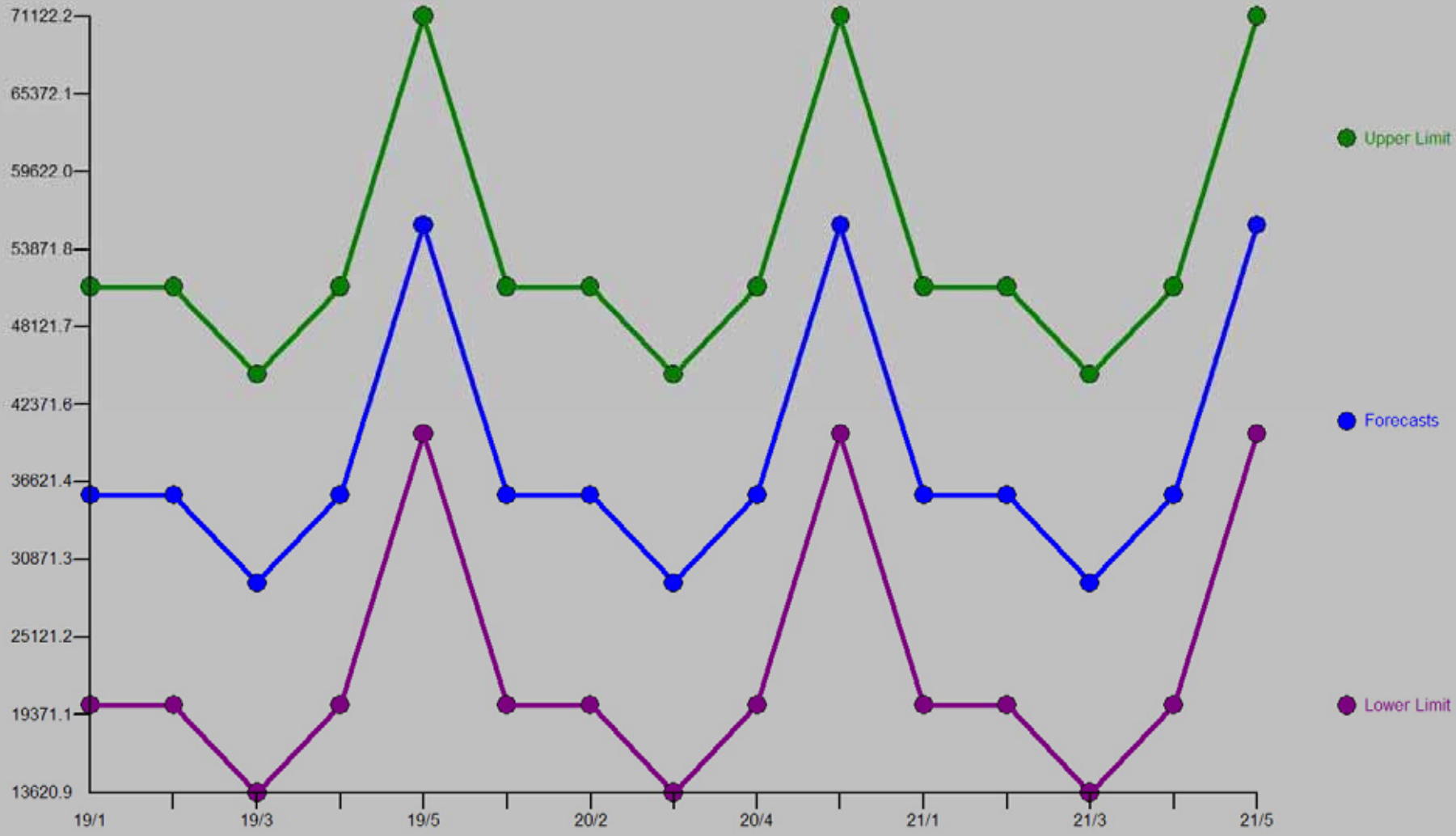
Periods 1/1 to 21/5(Seasonality of 5)

Residuals - CARR0280



Periods 1/1 to 18/5(Seasonality of 5)

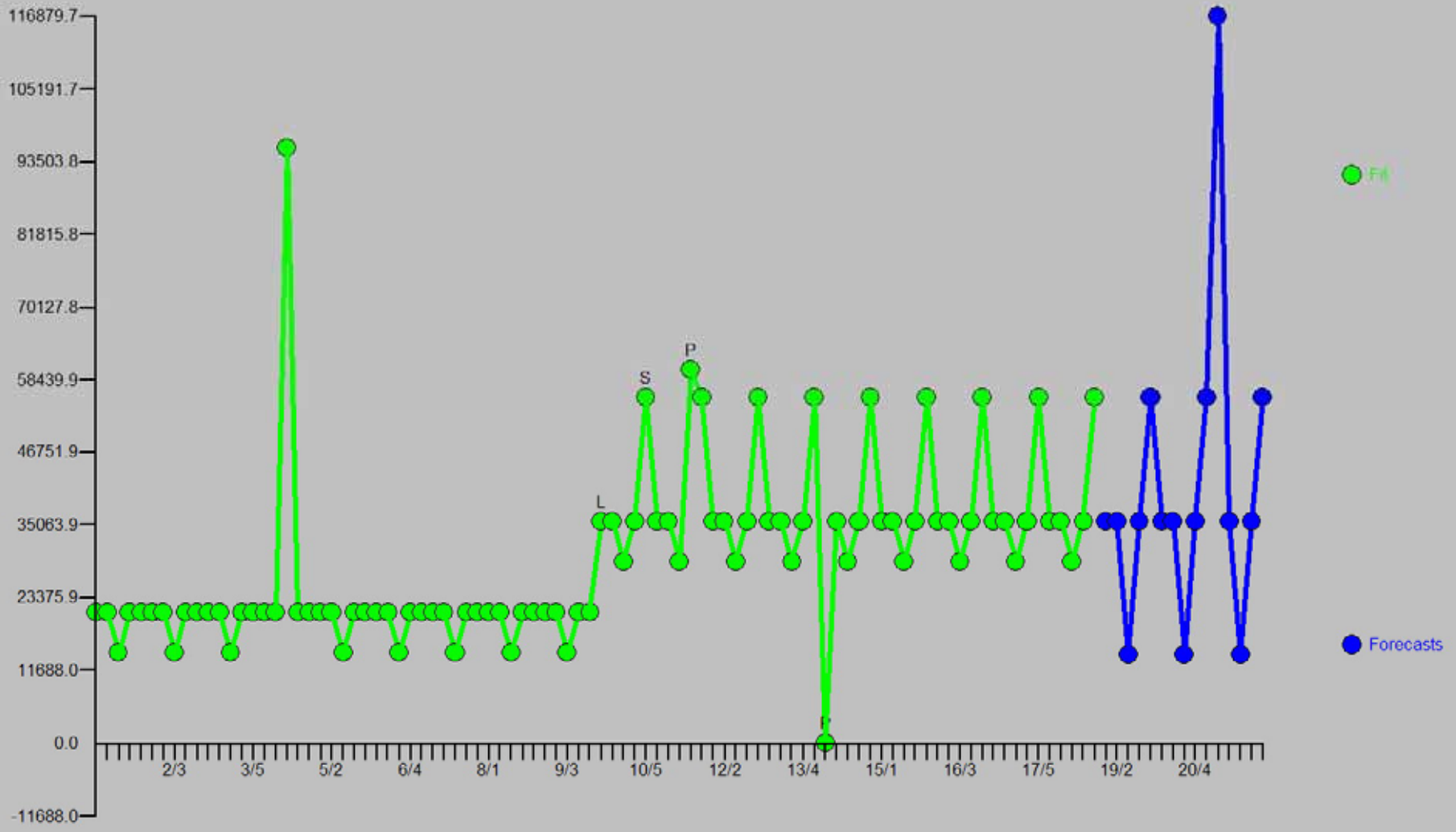
Forecasts, Lower and Upper Limits - CARR0280



Periods 19/1 to 21/5(Seasonality of 5)

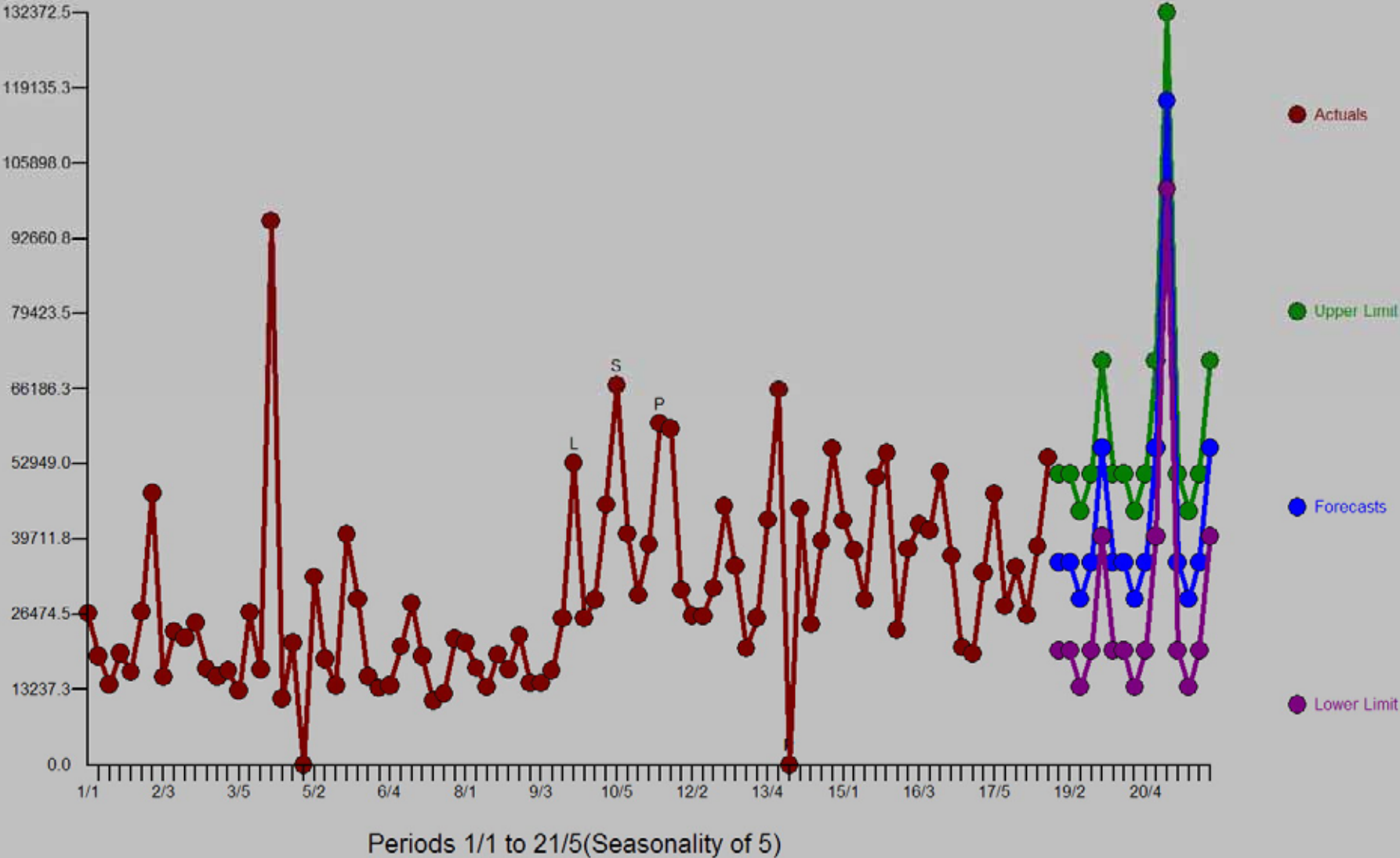
**Next Generation Of Forecasting
Tools Employed By Carreker
Incorporating Event Modelling**

Fit and Forecasts - CARR0280

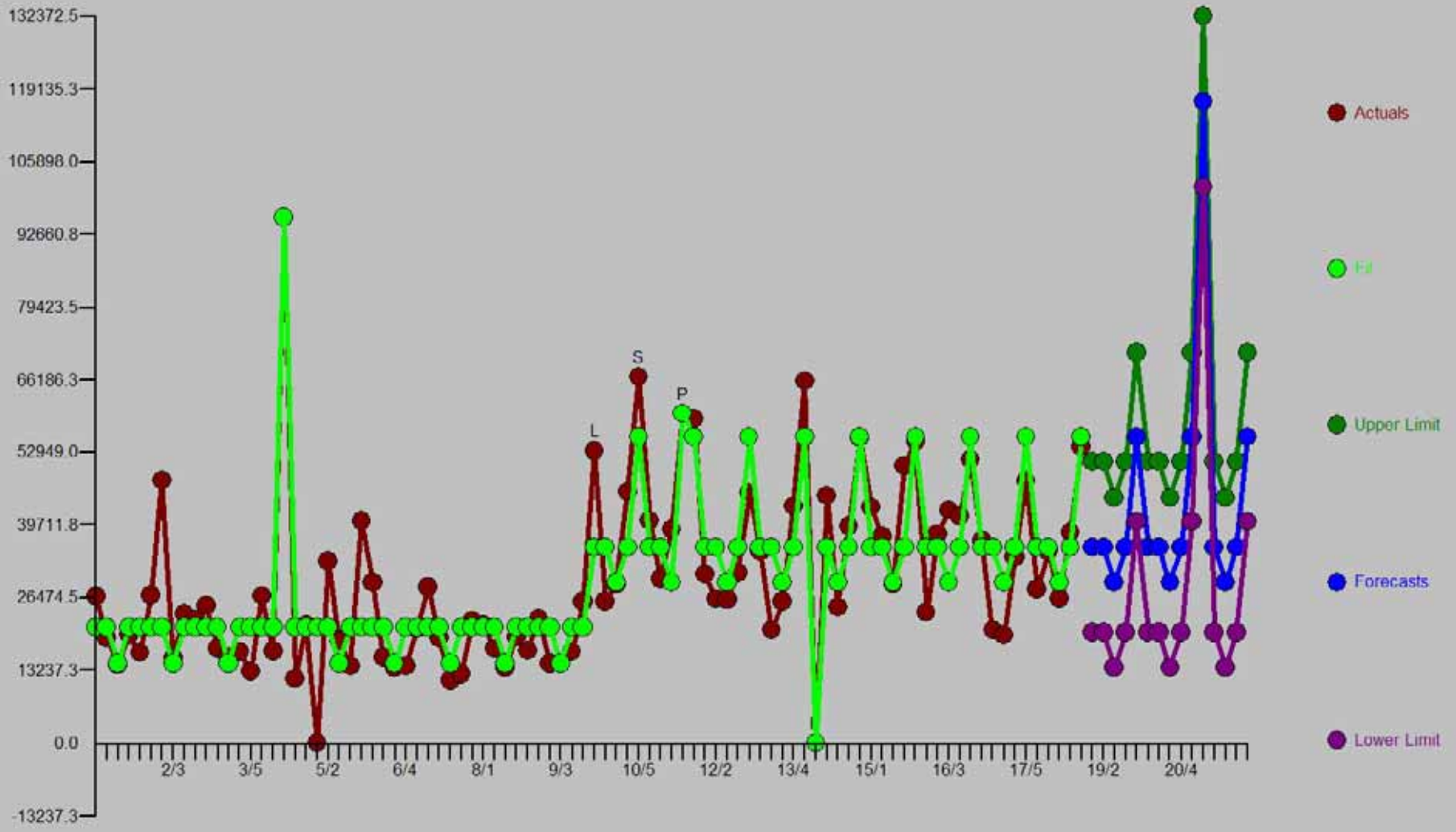


Periods 1/1 to 21/5(Seasonality of 5)

Actuals and Forecasts - CARR0280

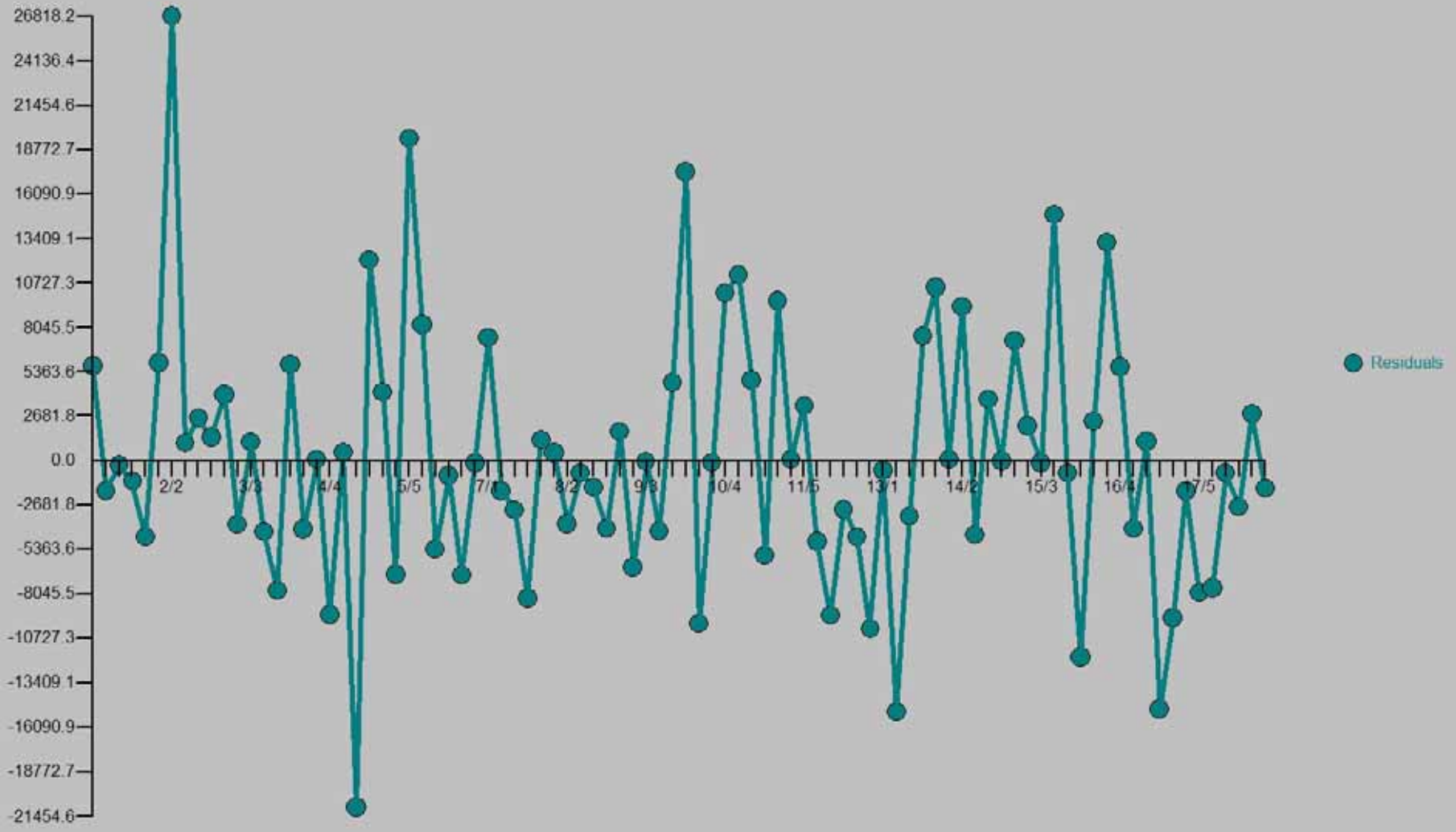


Actuals, Fit, Forecasts, Lower & Upper Limits - CARR0280



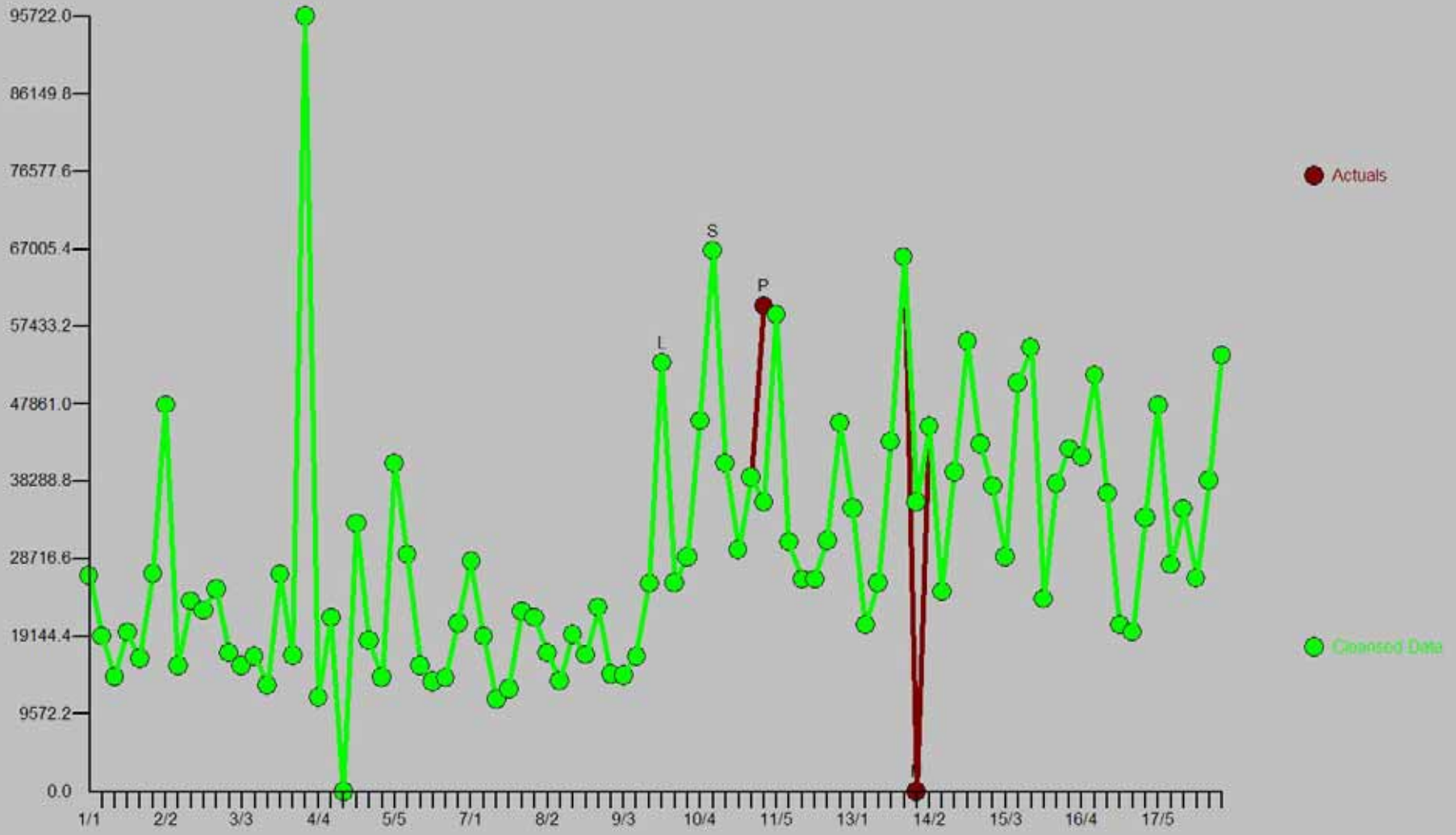
Periods 1/1 to 21/5(Seasonality of 5)

Residuals - CARR0280



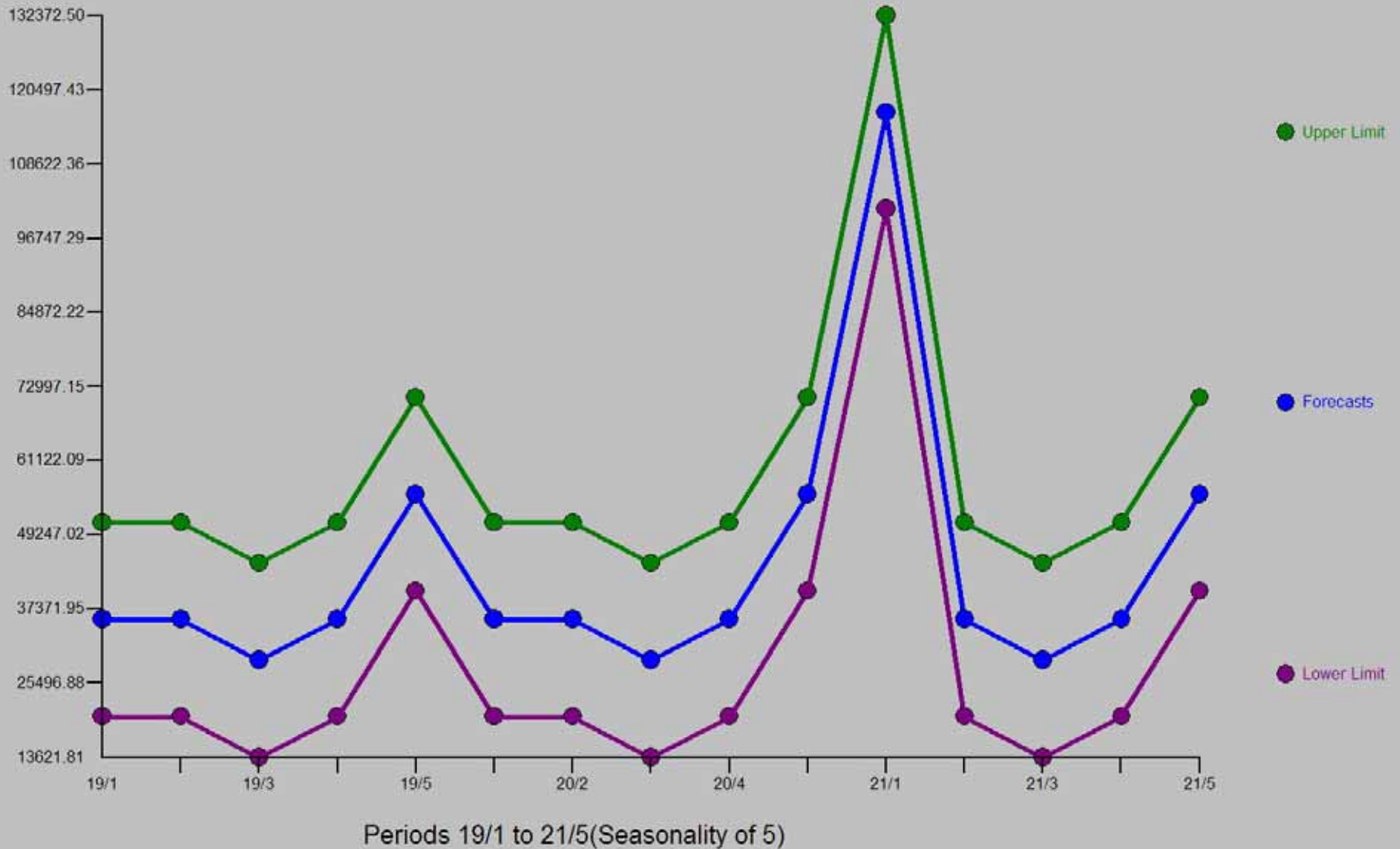
Periods 1/1 to 18/5(Seasonality of 5)

Actuals and Cleansed Data - CARR0280



Periods 1/1 to 18/5(Seasonality of 5)

Forecasts, Lower and Upper Limits - CARR0280

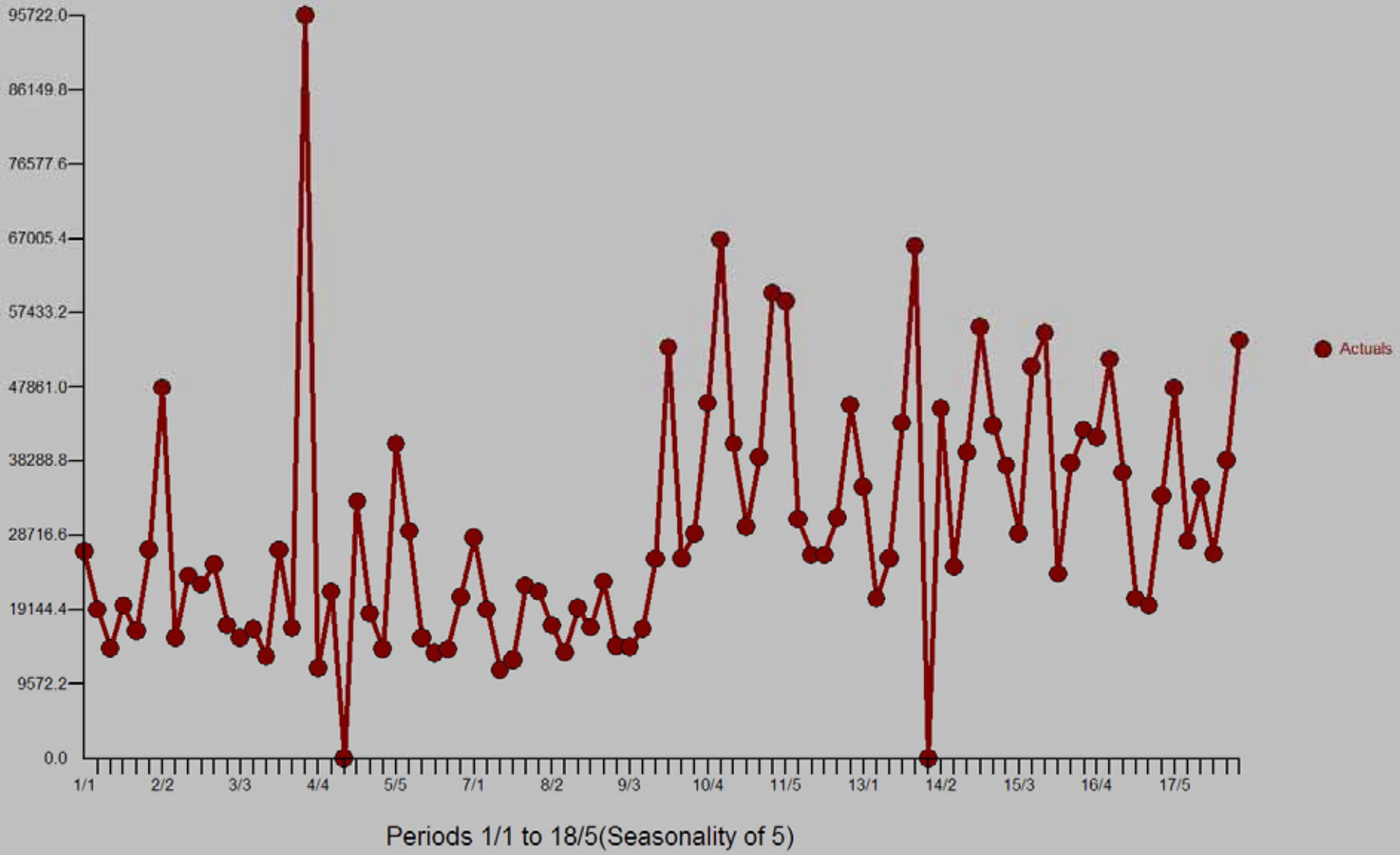


Future Developments

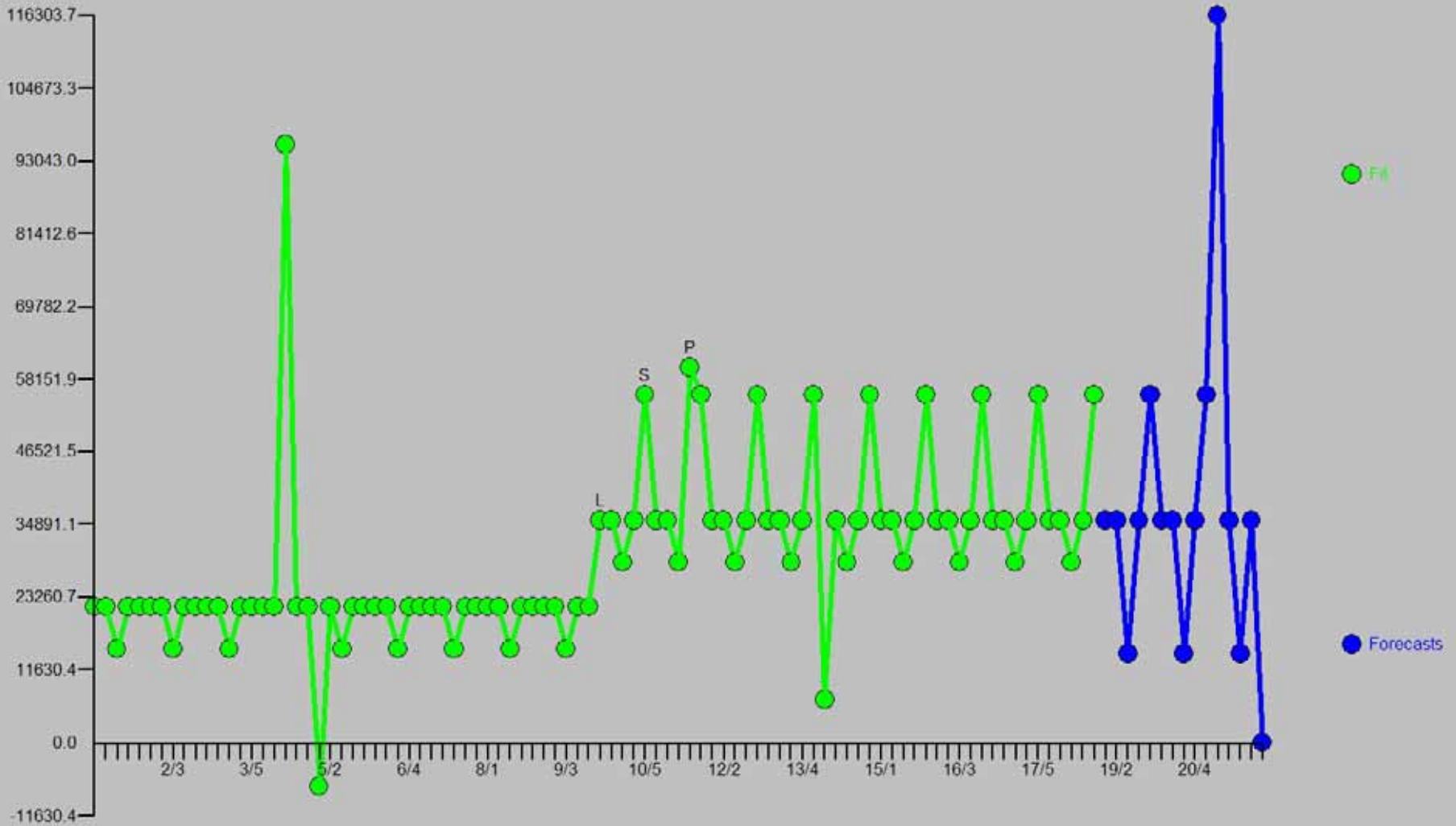
**Let Us Assume That We Knew
“WHY” We Observed **Zero** Values**

The ATM was physically unavailable those two days and will be unavailable one day in the future (the 15th day).

Actuals - CARR0280

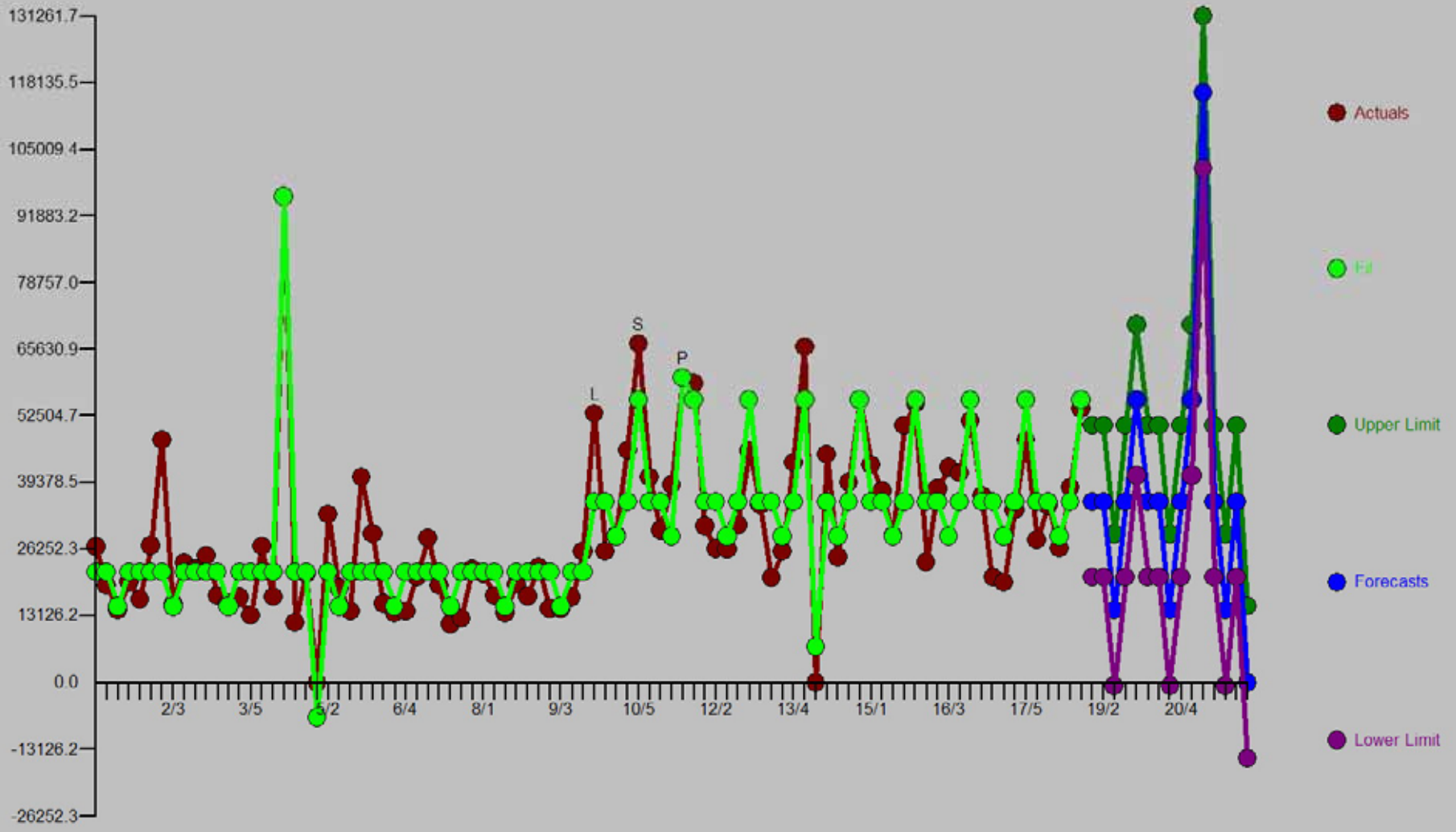


Fit and Forecasts - CARR0280



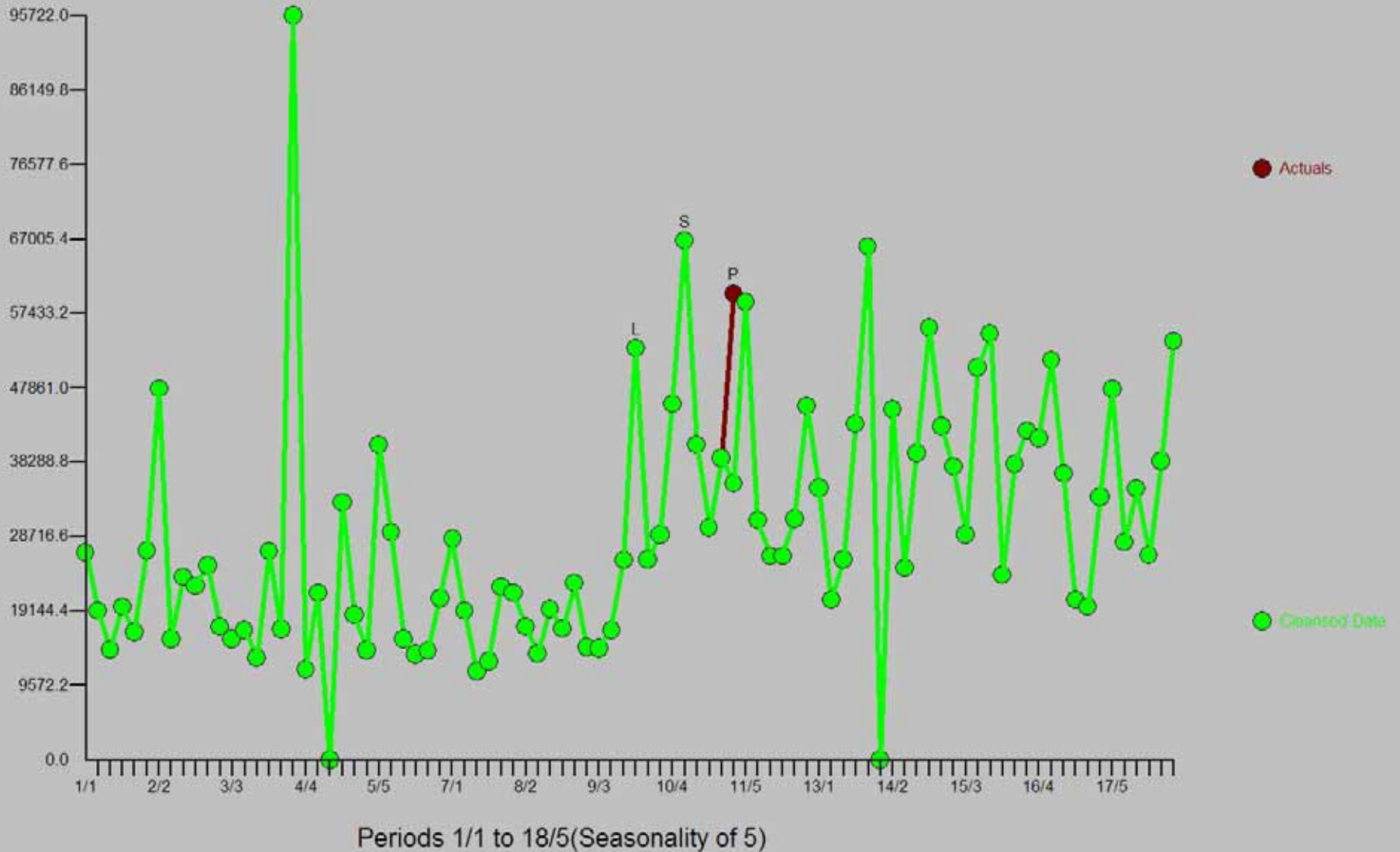
Periods 1/1 to 21/5 (Seasonality of 5)

Actuals, Fit, Forecasts, Lower & Upper Limits - CARR0280

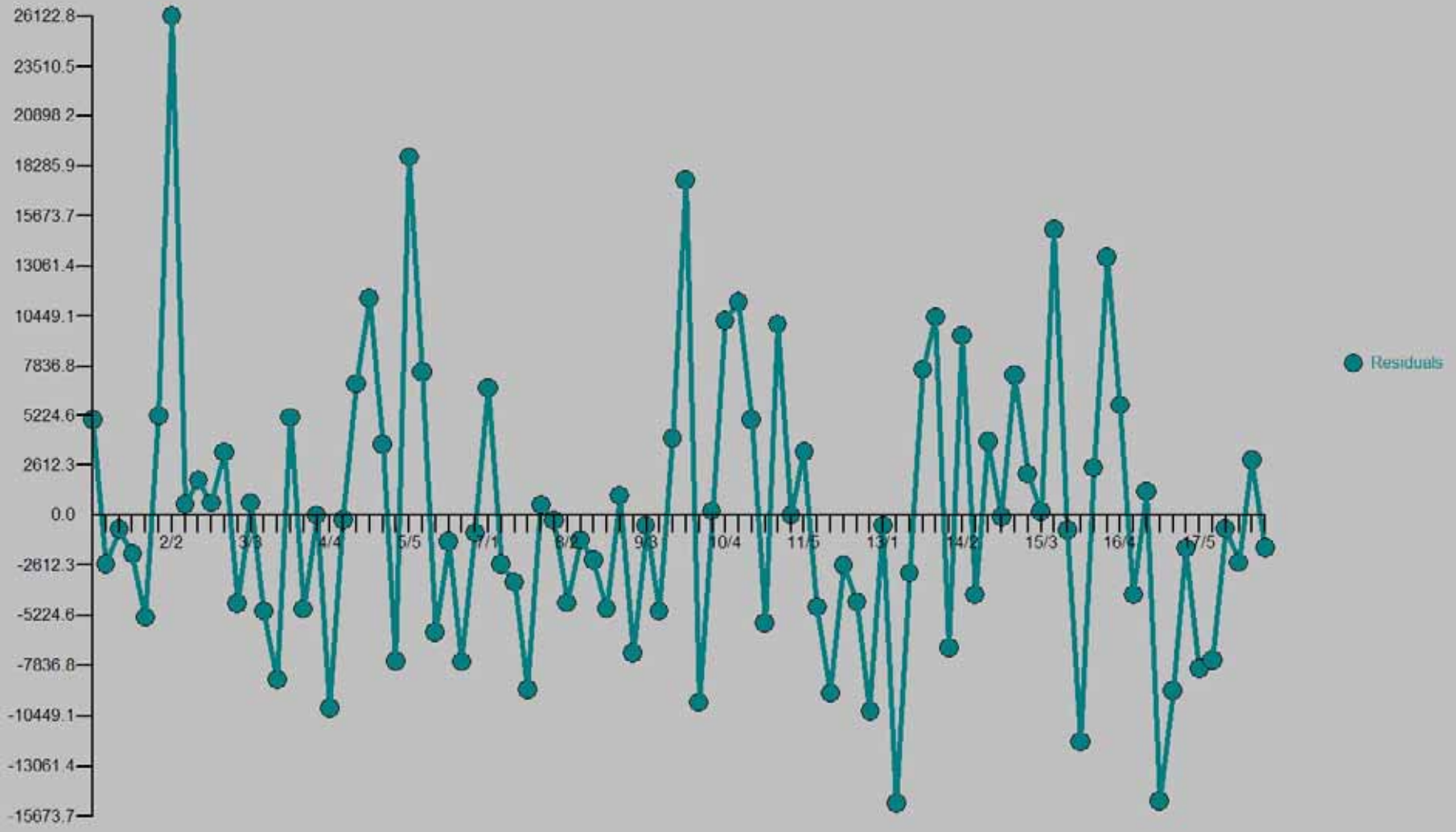


Periods 1/1 to 21/5(Seasonality of 5)

Actuals and Cleansed Data - CARR0280

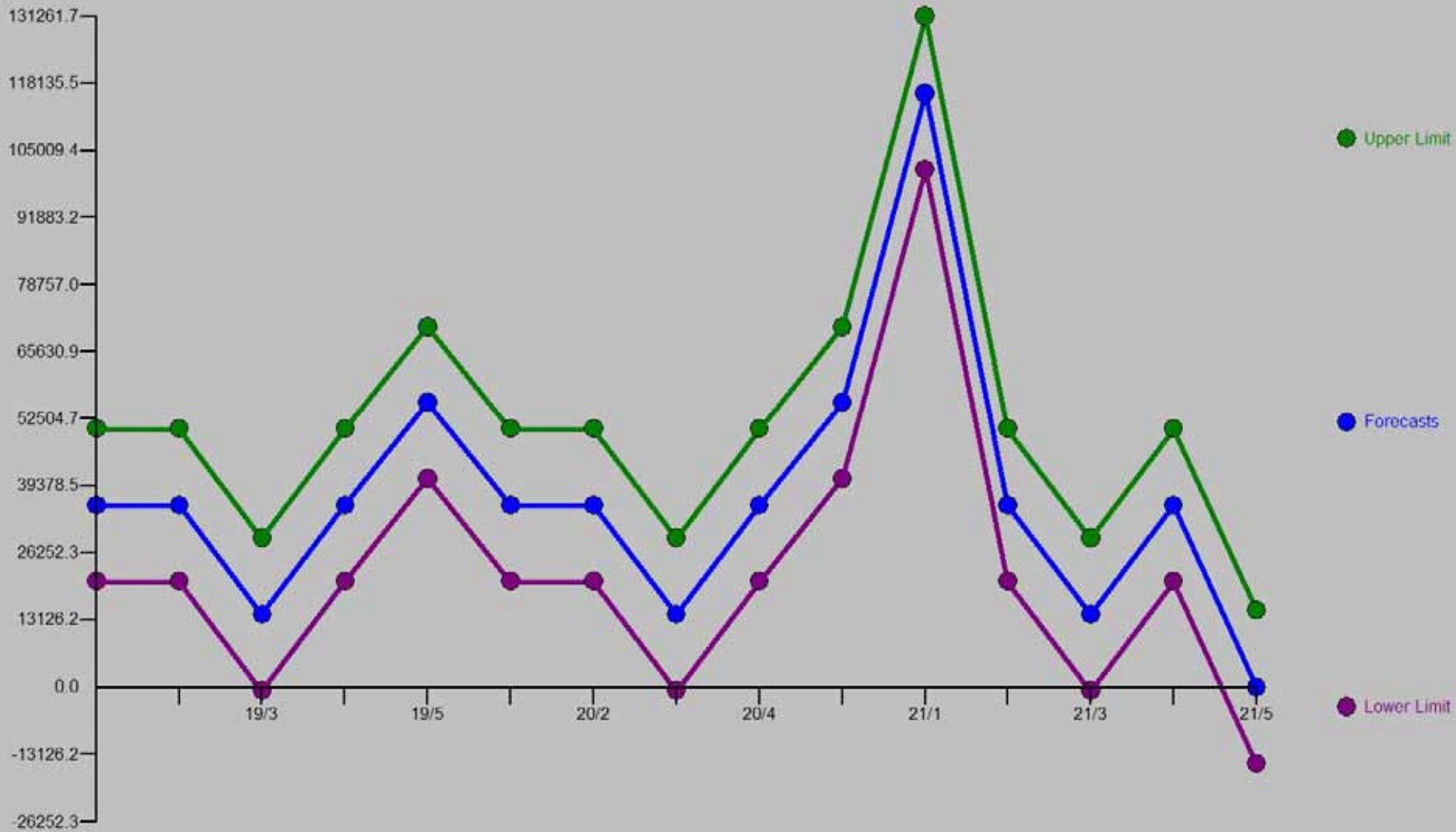


Residuals - CARR0280



Periods 1/1 to 18/5(Seasonality of 5)

Forecasts, Lower and Upper Limits - CARR0280



Periods 19/1 to 21/5(Seasonality of 5)

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

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	Connors	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 Forward Stop Home Search Favorites Media Print Mail Print Preview Print Setup

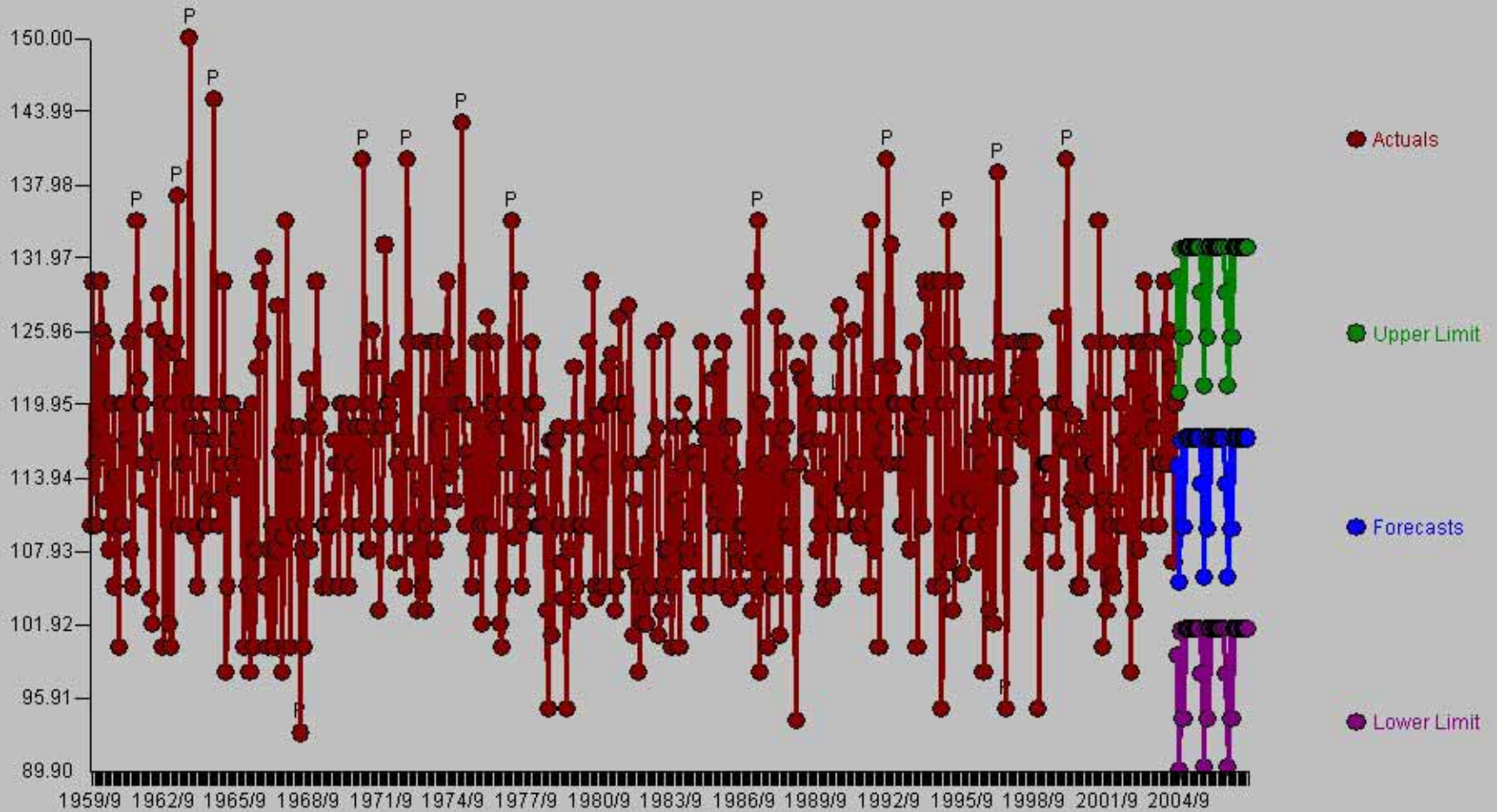
Address

Y! Search Web Messenger Bookmarks My Yahoo! Yahoo! Mail Shopping

A617 =

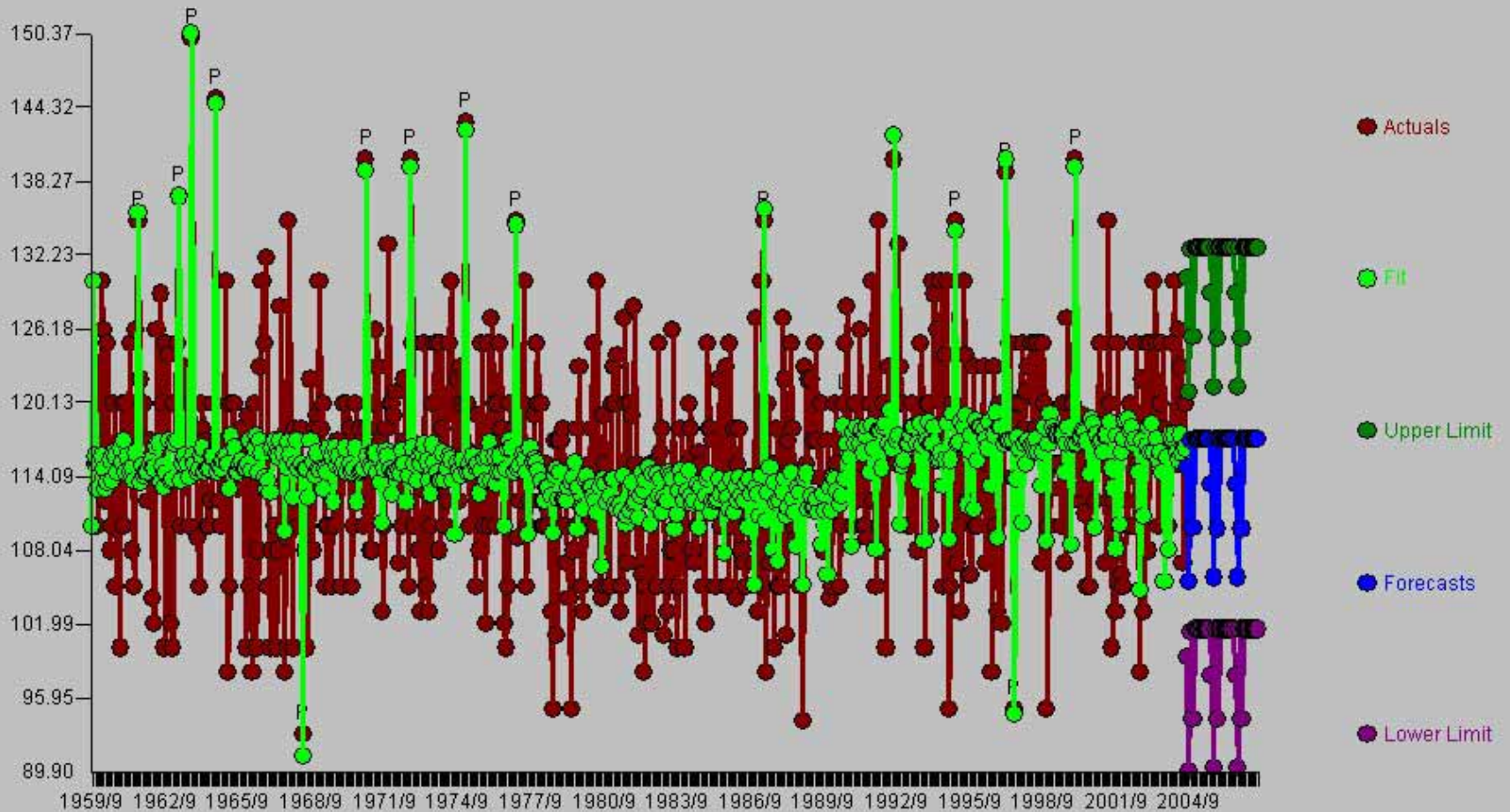
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	FNAME	LNAME	ISSUE	BIRTHDATE	CUP	BUST	WAIST	HIPS	HF	HI	WT	POTY	
71	Marianne	Gaba	9/1/1959	11/13/1939	.	34	24	34	5	6	110	0	
72	Elaine	Reynolds	10/1/1959	9/7/1939	.	39	25	37	5	8	130	0	
73	Donna	Lynn	11/1/1959	9/21/1936	.	36	22	36	5	3	115	0	
74	Ellen	Stratton	12/1/1959	6/9/1939	.	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	Connors	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	Dianna	Darford	11/1/1961	8/10/1938	.	36	22	35	5	7	120	0	

Actuals and Forecasts - BUNNIES



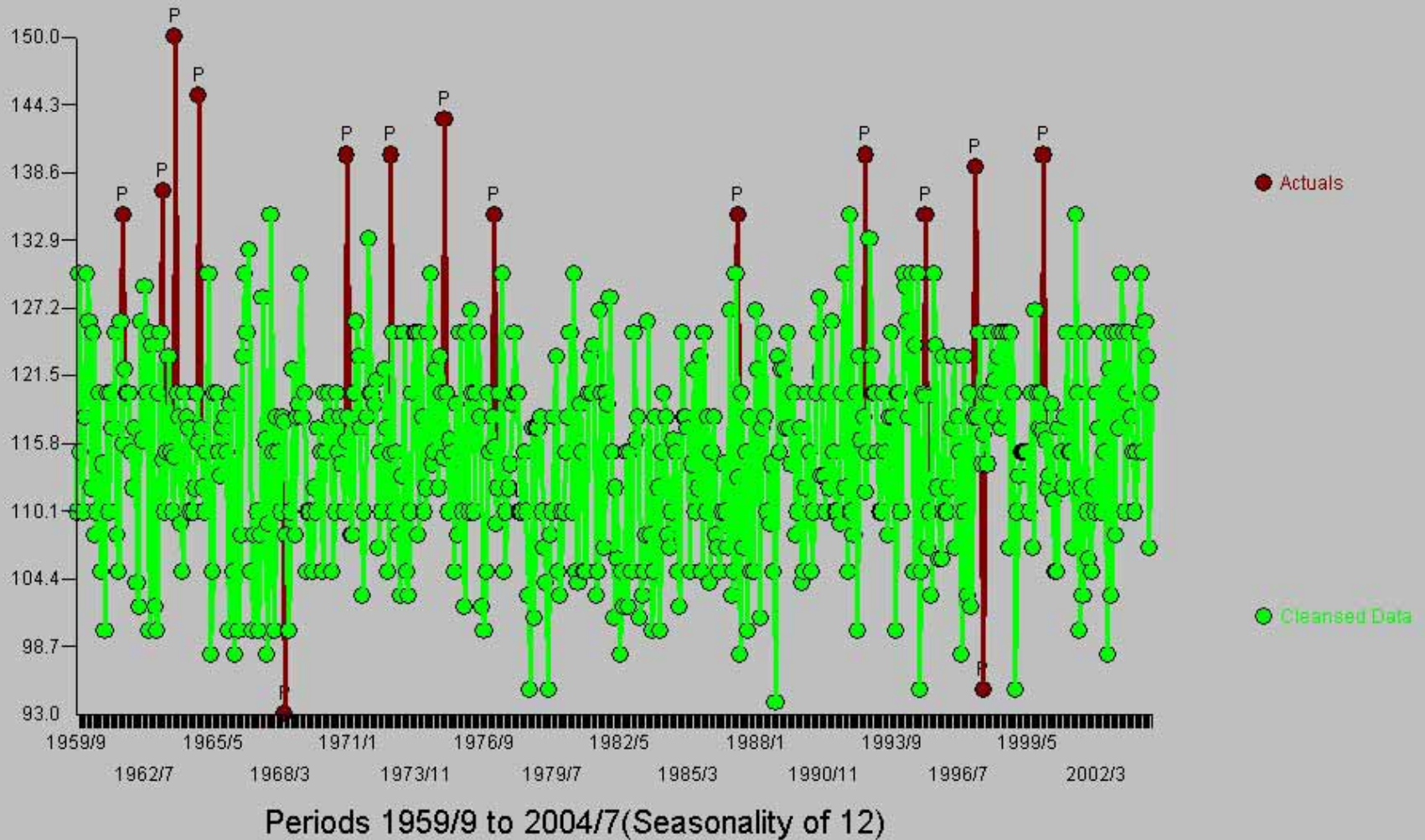
Periods 1959/9 to 2007/7(Seasonality of 12)

Actuals, Fit, Forecasts, Lower & Upper Limits - BUNNIES



Periods 1959/9 to 2007/7(Seasonality of 12)

Actuals and Cleansed Data - BUNNIES



Forecasts, Lower and Upper Limits - BUNNIES

