Working Paper 384

January 1993

APPLICATION OF OUTLIER DETECTION AND MISSING VALUE ESTIMATION TECHNIQUES TO VARIOUS FORMS OF TRAFFIC COUNT DATA

S D CLARK, S WATSON, E REDFERN & M R TIGHT

This work was sponsored by the Science and Engineering Research Council

UNIVERSITY OF LEEDS
Institute for Transport Studies

ITS Working Paper 384

January 1993

APPLICATION OF OUTLIER DETECTION AND MISSING VALUE ESTIMATION TECHNIQUES TO VARIOUS FORMS OF TRAFFIC COUNT DATA

S D CLARK, S WATSON, E REDFERN & M R TIGHT

CONTENTS

1. INTRODUCTION 1

2. DATA FORMAT 1
   2.1 Department of Transport data 1
   2.2 Highways, Engineering and Technical Services data 2
   2.3 Leicester data 2

3. OUTLIER DETECTION AND ESTIMATION TECHNIQUES EMPLOYED 4
   3.1 Averaging technique 4
   3.2 Box-Jenkins ARIMA technique 5
   3.3 Influence technique 5
   3.4 By-eye technique 6

4. DATA ANALYSIS 6
   4.1 Department of Transport data 6
      4.1.1 Averaging technique 6
      4.1.2 ARIMA technique 7
      4.1.3 Influence technique 8
      4.1.4 Comparison of results 9
   4.2 Highways, Engineering and Technical Services data 18
      4.2.1 Comparison of results 18
   4.3 Leicester data 26
      4.3.1 Averaging technique 26
      4.3.2 ARIMA technique 26
      4.3.3 Influence technique 26
      4.3.4 Comparison of results 27

5. SUMMARY AND CONCLUSIONS 27

REFERENCES 35

Appendix 1
1a Averaging technique 36
1b ARIMA technique 36
1c Influence technique 38

APPLICATION OF OUTLIER DETECTION AND MISSING VALUE ESTIMATION TECHNIQUES TO VARIOUS FORMS OF TRAFFIC COUNT DATA.

1. INTRODUCTION

This paper reports on the application of suitable techniques for detecting outliers and suggesting estimates for missing values in various forms of traffic count data.

The data used in this study came from three sources. The first set was provided by the Department of Transport's (DOT) regional office in Leeds and consists of automatic hourly traffic counts at four sites. The second set was part of a larger database provided by West Yorkshire Highways, Engineering and Technical Services (HETS). This set consists of automatic half hourly traffic counts on a single site. The third and final set was provided by Nottingham University and consists of automatic five minute traffic counts at 40 locations, in close proximity to each other, from Leicester.

Three suitable techniques emerged from pilot studies of such series conducted by Watson et al (1992a) and Redfern et al (1992). The three techniques are:

a) Maintaining an average and variability measure over time;
b) ARIMA modelling with detection of large residuals;
c) A point's influence on the correlation structure of the series.

A fourth technique, by-eye detection and estimation, provides an intuitive comparison for the first three techniques.

2. DATA FORMAT

2.1 DEPARTMENT OF TRANSPORT DATA

The data provided by the DOT consists of automatic hourly traffic counts from two pairs of counters on the A650 trunk road between Keighley and Skipton. One counter of the pair measured the easterly flow whilst the other the westerly flow. The four detectors are coded W496, E497, W498 and E499. A schematic representation of the location of the vehicle detectors is given in Figure 1. Data is available covering the last four months of 1990 thereby providing four series, each containing 2928 observations. The quality of the series was good. Only one observation was missing, observation 900, in the W498 series. Two significant events occurred during the period of this data. The first was severe wintry weather on the weekend of 8th and 9th of December. The second is the Christmas period in the last week of the series. Both these events caused a significant change in the *typical* traffic flow profile.

For the purpose of analysis each series was divided into 24 individual series, each consisting of one hours data for the day. Thus the first series is the traffic flow from 0:00 to 01:00 in each day, the second is the flow from 01:00 to 02:00 each day, etc. The advantage of doing this is to exploit the natural seasonality of span seven (a week) which results in each of the 24 series. Further justification for this disaggregation is given in Section 3.

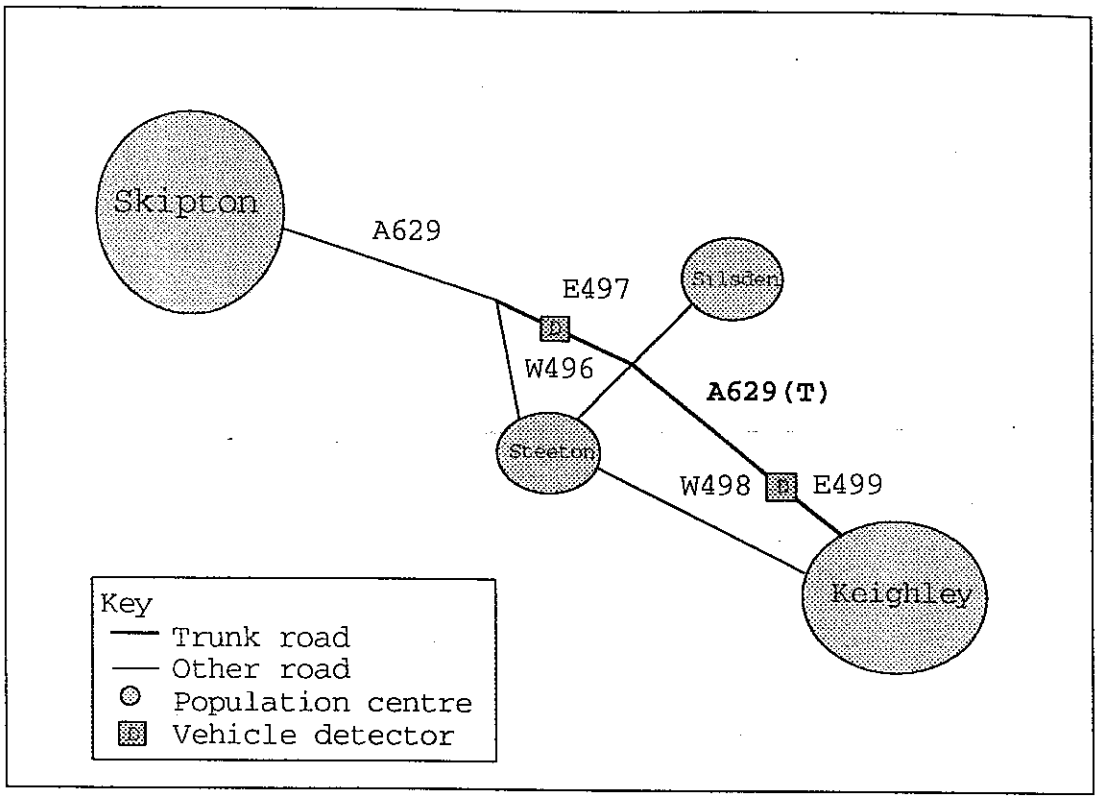

Figure 1 - Schematic diagram of position of DOT traffic detectors

2.2 HIGHWAYS, ENGINEERING AND TECHNICAL SERVICES DATA

The data from the Highways, Engineering and Technical Services division (HETS) was in a similar form to that of the DOT data. The half hourly flows are from an A road, the A653, just to the north of Dewsbury town centre. A full 12 months of data, in two directions, is available giving 17520 observations in each of the two series. The data is a lot more patchy than the DOT data, with a stretch of 1392 observations missing from both the north and south bound series. In addition the northbound traffic has a patch of 48 observations missing in an earlier part of the series. For the purposes of this study the traffic flows were summed to hourly flows, giving 8760 observations and missing patches of length 696 and 24. As in the case of the DOT data these series were divided into 24 individual hourly series.

2.3 LEICESTER DATA

The data from Leicester, part of the instrumented city project, was of a different form to any previously considered in this project. The counts were five minute traffic flows collected by the SCOOT traffic control system along a large number of links. These detectors are in close geographical proximity to each other, typically between 50m and 250m apart. Figure 2 gives a schematic diagram of region W of the network whilst Figure 3 gives a schematic representation of region B. The series vary in start time and duration. Typically the collection of data for the operation of SCOOT starts at 07:30 and goes on until 19:45, giving 147 observations in each series, although some stop mid morning giving in the region of 30 observations. The quality of the series vary. In most only the occasional observation is missing, whilst in one set there are frequent patches of missing observations.

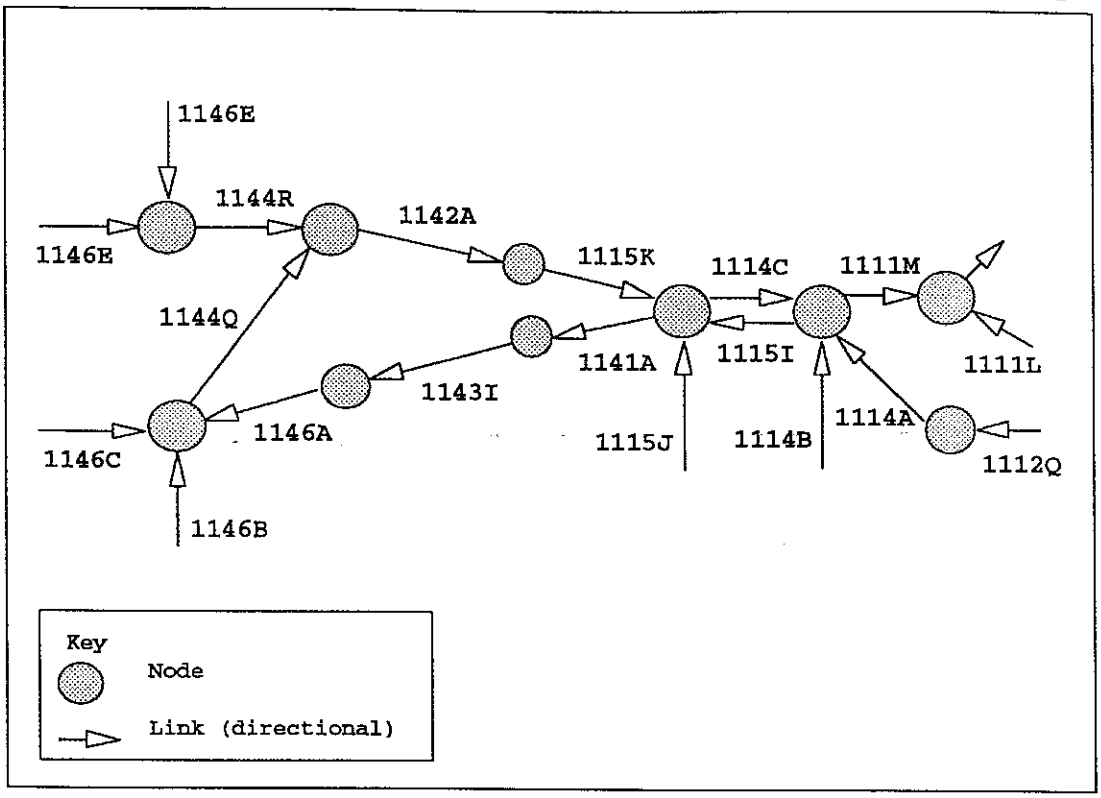

Figure 2 - Region W of Leicester SCOOT network

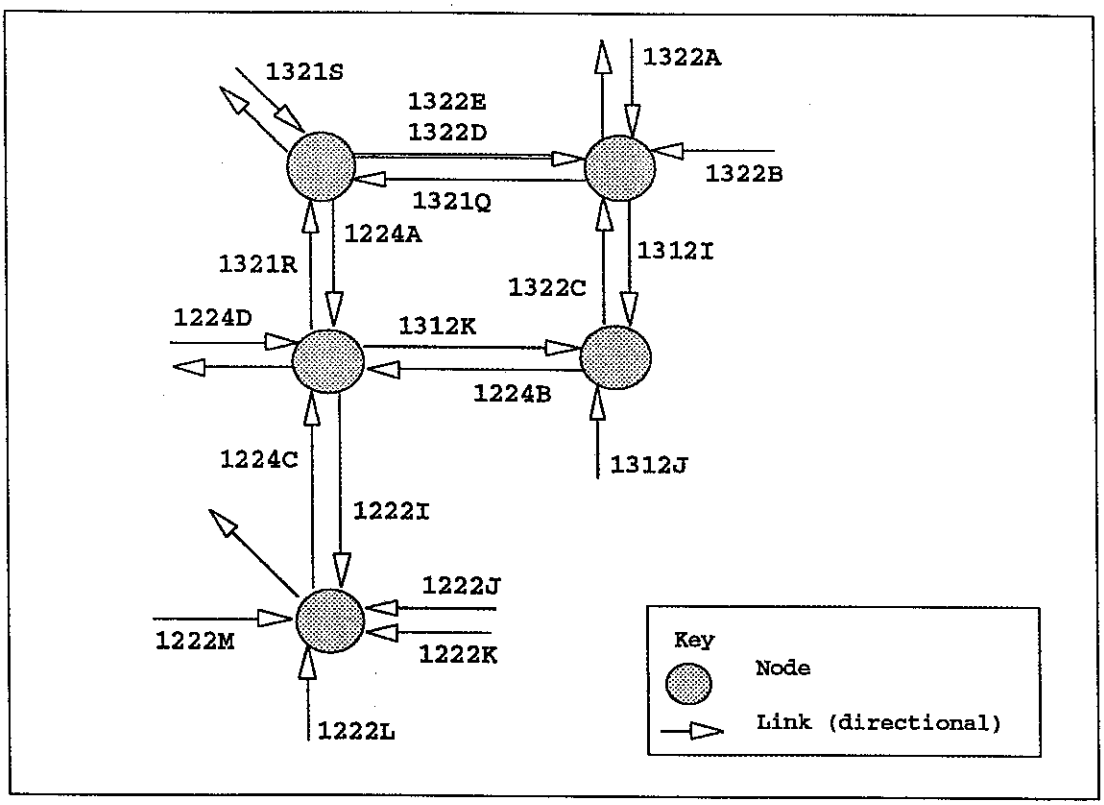

Figure 3 - Region B of Leicester SCOOT network

Table 1 lists the days for which data was analysed and the number of observations in each series.

Date	Day	Observations
06/04/91	Sunday	145
12/04/91	Saturday	146
24/04/91	Wednesday	146
26/04/91	Friday	31
01/05/91	Wednesday	35
03/05/91	Friday	52
06/05/91	Monday	146
08/05/91	Wednesday	145

Table 1 - Leicester data summary

3. OUTLIER DETECTION AND REPLACEMENT TECHNIQUES EMPLOYED

All the three techniques considered in this study are able to detect outliers, provide replacement values for these outliers and estimate missing values in a time series.

3.1 AVERAGING TECHNIQUE

This is a development of the Department of Transport's method for validating and patching traffic flow census data. An average and variability measure is maintained for each season under study, usually one for each hour of the day, since the traffic flows are hourly. Each new observation is tested to see if it falls in the interval given by seasonal average ± 4 seasonal standard deviations. If the new observation does fall in this interval then it is accepted and used to update the period average and variance. If, however, it does not fall in this range then it is flagged as an outlier. A replacement value is then suggested using an exponentially smoothed weighted average of previous observations at this season. Estimates for missing values can also be obtained using this smoothing method. Appendix 1a contains a mathematical description of this technique.

The technique has the advantages that it is intuitive and requires little outside intervention. A disadvantage is that since the season averages and variances need to be primed using observations from the first three seasons no outliers can be detected during this period. For the DOT and HETS data sets each of the series consist of the same hours flow in each day of the week. So for the purposes of this study the season in the DOT and HETS data is of length 7, necessitating the maintenance of 7 season averages and variances and no outliers being detected during the first 21 observations of each series. Since there are 24 such series the number of observations used for priming the averages is 21*24=504. Section 4 contains a more detailed discussion about the effect of using different priming periods for this technique with the DOT, and hence also the HETS, data sets. With the Leicester data this priming is performed with only three observations. The behaviour of the series during this priming period needs to be typical of the remaining part of the series otherwise spurious outliers are detected.

3.2 BOX-JENKINS (1976) ARIMA TECHNIQUE

This type of technique relies upon an underlying period-to-period relationship in the data. Thus the observation at the current time period is related to previous observations, either the immediate predecessor or a predecessor from a number of observations back. This relationship can be a straight forward linear relationship (referred to as Autoregressive, AR) or can be related to previous errors in the model (referred to as Moving average, MA). Indeed it is possible to have both these relationships in the same model and such models are termed ARMA models. Before this technique can be applied it is necessary that the series under consideration should be roughly constant in its mean level and its variance. This property is referred to as stationarity. Weak stationarity in the mean can usually be achieved by differencing the data so that any time trend or level shift is reduced to a constant term. In effect rather than modelling the series itself we are modelling the change in the series from observation to observation. This is commonly referred to as integration (I). Stationarity in the variance can also be achieved by applying a transform to the data to dampen any change in variability, usually a logarithmic transform is applied. The three elements of the models are usually brought together and given the term ARIMA models (Appendix 1b gives a more mathematical description of the Box-Jenkins approach).

The DOT and HETS data is highly seasonal with a span of seven days. A suitable model form will include a component from the previous time period and perhaps the time period 7 observations back. Here the justification for the division of the original series into 24 same hour series becomes apparent. If the series was not so divided then the same hours flow in the previous week would be 7*24=168 observations back in the series which is too long a span for meaningful model identification, fitting and diagnostic procedures.

An visual inspection of the Leicester data suggest that a differencing is required in the series to reduce the time trends to a stationary level. A strong relationship may also exist between an observation and the one for the previous period.

Once a suitable model has been identified 0/1 intervention variables can be introduced to account for known missing values. Outliers may then be identified as those observations whose residual (the difference between the observation and what the model predicts for the observation) is large. For our purposes large is defined as being beyond three times the standard error for the fitted model. Additional 0/1 intervention variables are then introduced to account for these outliers and the model refitted. When no further outliers have been found the identification process finishes.

The advantages of this technique are that it provides a large degree of descriptive information for time relationships in the series and has a proven track record in application. The major disadvantage is that a degree of specialist knowledge is required to identify, fit and interpret the model. This can be a time-consuming task. Through the process of differencing, the technique will lose observations at the start of a series. In our case a differencing of seven will lose seven observations. Hence outliers cannot be detected in that period.

3.3 INFLUENCE TECHNIQUE

These methods were developed in a time series context by Chernick, Downing and Pike (1982) and later extended by Watson (1987). The technique attempts to measure the influence of each observation on the underlying correlation structure in the series. Outlying observations will tend

to affect this relationship and can also produce spurious correlations. Observations which are calculated to have a large influence are designated as outliers. Replacement values for these outliers and any missing observations can also be obtained from this technique. Appendix 1c gives a more mathematical description of how an individual observation's influence is calculated and interpreted.

This technique has the advantage that, unlike the Box-Jenkins approach and to a lesser extent the averaging technique, no hypothesised model is required for the data. All that is required is that the time span of the correlation structure is adequate enough to encompass all the associations in the series. This requirement justifies the division of the data into 24 same hour series. If the series was treated as one long series then a strong correlation would be expected at lag 24 and lag 168 which is too great a span for practical purposes.

The disadvantage of this technique is that observations during the final season of the series can not have an influence value calculated and can not therefore be detected as outliers. This can be a serious drawback since it is the recent history of a series which is usually of the most interest. A method for overcoming this problem is to reverse the series so that influence values can no longer be calculated for the first season in the series. This relies upon the series being time reversible. This property can be tested using the method described in Tsay (1992). Tsay's method is, however, more stringent than is required in this study. Here a series is said to be reversible if the autocorrelation structure calculated in a forward direction ($r(k) = f(t, t+k)$) is the same as that in a backward direction ($r(k) = f(t, t-k)$). This is sufficient since only the correlation structure of the series is at the heart of the detection and replacement technique.

3.4 BY-EYE TECHNIQUE

This simply involves inspecting the series for any unusual observations. Estimates for missing observations can be guessed at using neighbouring observations. The best results can be obtained by graphing the series over time and inspecting the resultant graph. This method also provides a good intuitive yardstick against which to measure the performance of the other methods. The disadvantage of this method is that an element of subjective judgement is involved.

4 DATA ANALYSIS

4.1 DEPARTMENT OF TRANSPORT DATA

To asses the effectiveness of each of the techniques when applied to this data set the existence of the bad weather and Christmas periods of low flow will be used. It is to be hoped that most techniques would detect this situation and suggest reasonable replacements for the atypical observations during this period. It should be noted that the original observed flows are the correct flows and that the replacements are only a suggestion of what might have happened if the cause of the event did not occur.

4.1.1 Averaging technique

A number of approaches to the analysis of this data set are available. The first is to only maintain one average and variability measure throughout the whole series. This approach suggests that the flow throughout the day is fairly uniform in its level and variability. This is unlikely to be the

case since peak hour flows can change the traffic flow dramatically. The second is to maintain 24 average and variability measures, one for each hour of the day. This relies upon an assumption that, day to day, the flow will be uniform. Whilst this may be true for the weekdays it will not usually be so for the weekends. The flow on a weekday in any hour will usually be much greater than that on a Sunday or even a Saturday. The third and final approach is to maintain an average and variability measure for each hour of the week, 168 in total. This approach suggests that the flow on any particular day for any particular hour will be uniform from week to week. This was the approach adopted. To justify this decision a number of preliminary exercises were conducted. The first was to use approach one and maintain only one pair of average and variability measures. Different lengths for the priming period were adopted to try and improve the performance of the technique. The second and third exercises implement approaches two and three. Figure 4 shows how approach one performs at the start of a series. The thick solid line is the original series, the solid line is the suggested replacement for rejected observations and the dashed lines are the lower and upper bounds. Clearly the approach is unable to react to the large increase in flow in the hour to 07:00. This approach goes on to reject subsequent flows until 23:00 when the original flow falls back within the limits. Table 2 gives the number and percentage of observations detected as outliers in each exercise. Within Table 2 the 'Priming' row indicates the number of observations required to prime the series averages and variability measures. The 'Replacement' row is the suggested replacement for the missing observation.

Approach	ONE			TWO	THREE
Priming	3	72	504	72	504
W496	1632 (56%)	1591 (54%)	1338 (46%)	277 (9%)	148 (5%)
E497	1673 (57%)	1633 (56%)	1377 (47%)	298 (10%)	135 (5%)
W498	1595 (54%)	1567 (54%)	1237 (42%)	297 (10%)	164 (6%)
E499	1726 (58%)	1673 (57%)	1403 (48%)	312 (11%)	156 (5%)
Replacement	238	238	238	1253	933

Table 2 - Number and percentage of observations detected as outliers

Clearly approach one is unacceptable since frequently, more than half the observations are detected as being outliers. Increasing the priming period does not significantly enhance the performance of this approach. Use of approach two does improve the situation but still the percentage of outliers is large. Clearly approach three is the most suitable, with the percentage number of outliers detected being near the 5% mark. A further guide to the quality of the results is the replacement value suggested for observation 900 in the W498 series. For approach one it is very much on the low side. For approach two it is on the high side whilst the value for approach three is reasonable. Figure 5 presents this replacement result, and others, in context.

4.1.2 ARIMA technique

From earlier studies (Watson et al, 1992a and Redfern et al, 1992) and other explorative work a suitable form of ARIMA model for this type of data is an ARIMA$(1,0,0)(0,1,1)_7$ model. This model form was applied universally to all the time series and proved to be suitable and descriptive. In those cases (13%) where this form was not appropriate as indicated by poor

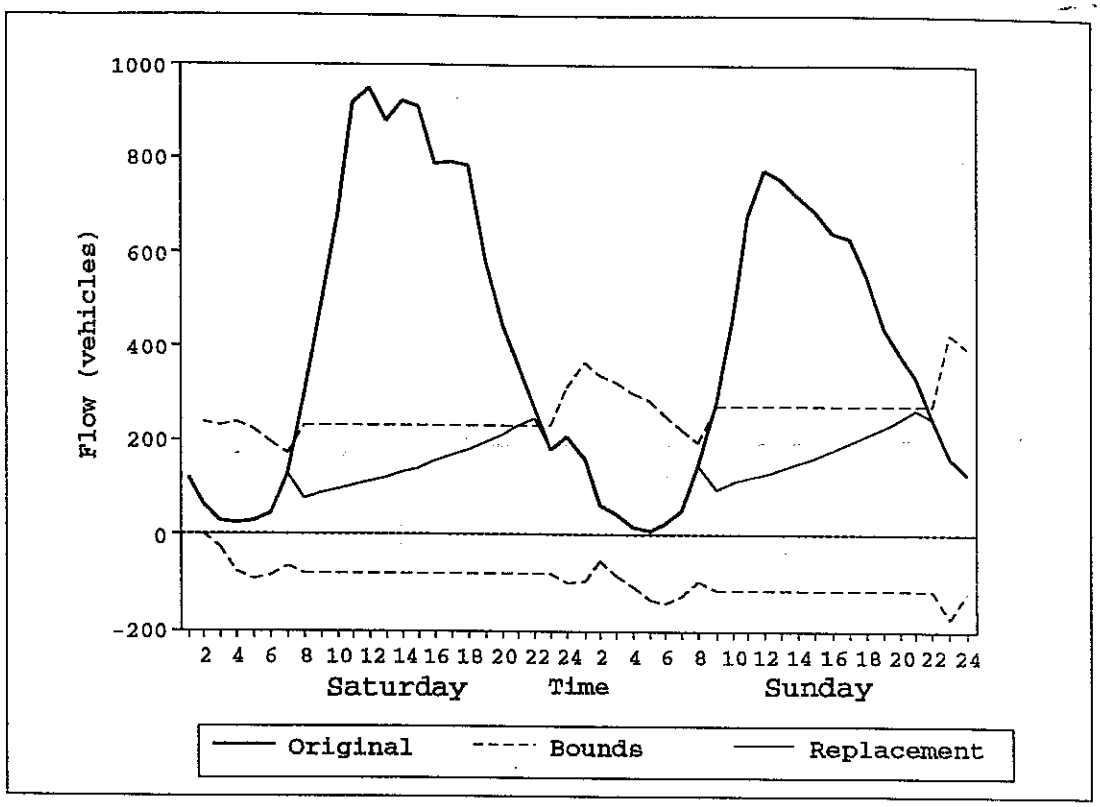

Figure 4 - Performance of approach one

diagnostics or due to a lack of convergence alternative forms were adopted. These alternatives were (in decreasing order of use):

a) ARIMA(p,0,0)(0,1,1)$_7$ where p=2 or 3. The number of non-seasonal autoregressive parameters was increased to account for any remaining autocorrelation in the model's residuals. Such a model possessed more descriptive power than the universal model;

b) ARIMA(1,1,0)(0,1,1)$_7$. A non-seasonal trend term was required to account for a day to day change in traffic flows. This model was most often required in the peak hours;

c) ARIMA(p,0,q)(P,1,Q)$_7$ where p=0 or 1, q=0 or 1, P=1-p and Q=1-q. A reversal of the types of non-seasonal and seasonal parameters was required so that convergence of the estimation criterion procedure could take place. These models were not as descriptive as the universal model.

The suggested replacement value for the missing observation in the W498 series is 882.

4.1.3 Influence technique

For a 95% critical value point approximately 180 observations (6%) were detected as outliers in each series. This is a reasonable number to expect given the critical value chosen and the presence of a large number of atypical observations due to bad weather and Christmas. At the 99% level around 40 observations (1.4%) were detected as outliers. This again is a reasonable result. A 99% level was adopted for the remainder of this study. The Christmas period fell in the last 7 days of

these series so ordinarily the influence technique would be unable to detect outliers and suggest replacements for this event. If, however, the series is reversed then the period of non-function for this technique applies to the first 8 days (since autocorrelations up to lag 8 are being used) of the series and not the last 8. The detection of outliers and suggested replacements now become available for the Christmas event. The other outliers detected (ie non-Christmas) were, reassuringly, the same irrespective whether the series was reversed or not.

The suggested replacement value for the missing observation in the W498 series is 894.

4.1.4 Comparison of results

Figure 5 shows the replacement value suggested by each DOT approach and the other two techniques.

A reasonable by-eye estimate for the missing observation is 820. The averaging technique primed with 504 observations is within 13% of this value whilst the ARIMA and Influence techniques are both within 8%.

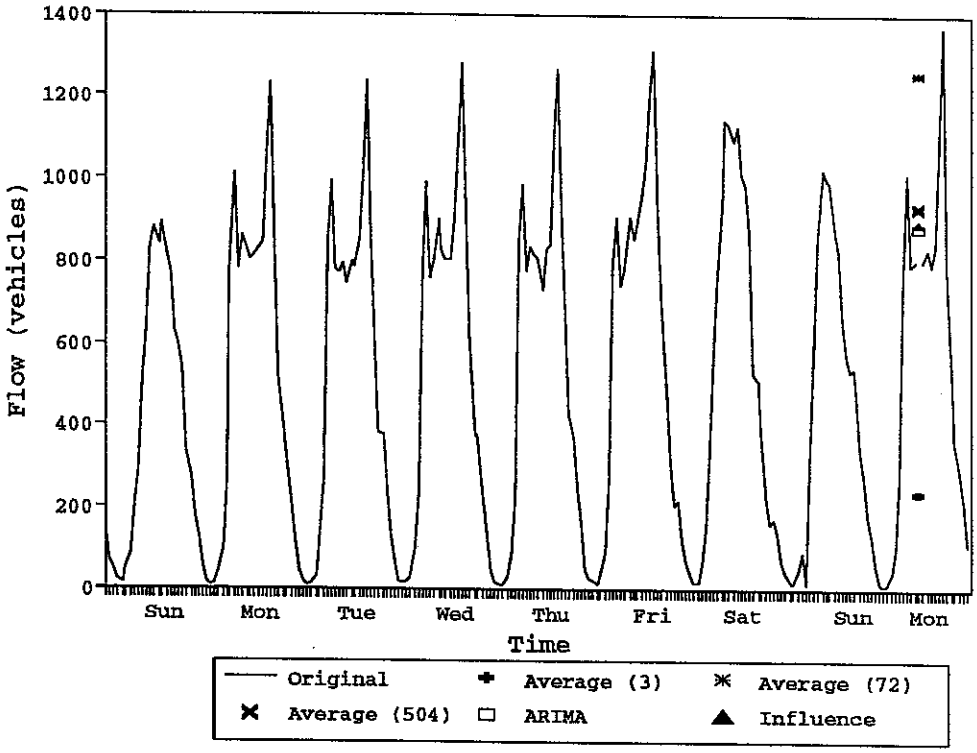

Figure 5 - Suggested replacements for W498 missing observation

Figures 6-9 show the outliers detected and the replacement values for the period of bad weather in each series. All the techniques detect the unusually low flow throughout the Saturday and most of the Sunday. The average technique produces a replacement estimate which is consistently on the large size and somewhat unbelievable. The ARIMA technique tends to produce the middle estimate for the replacement whilst the influence technique produces the lower estimate. Only the ARIMA and influence replacements are credible.

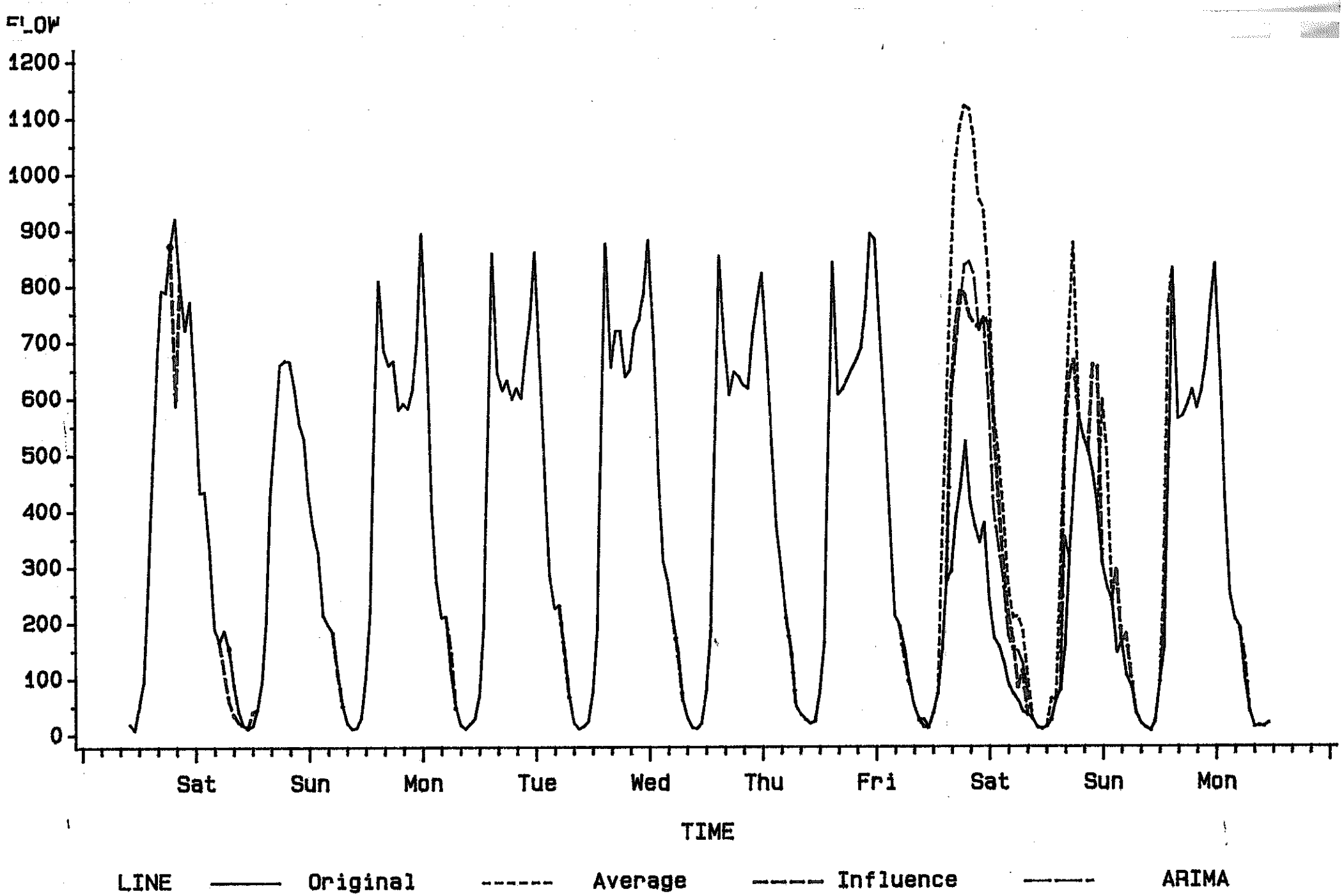

Figure 6 - Outliers detected and replaced for W496

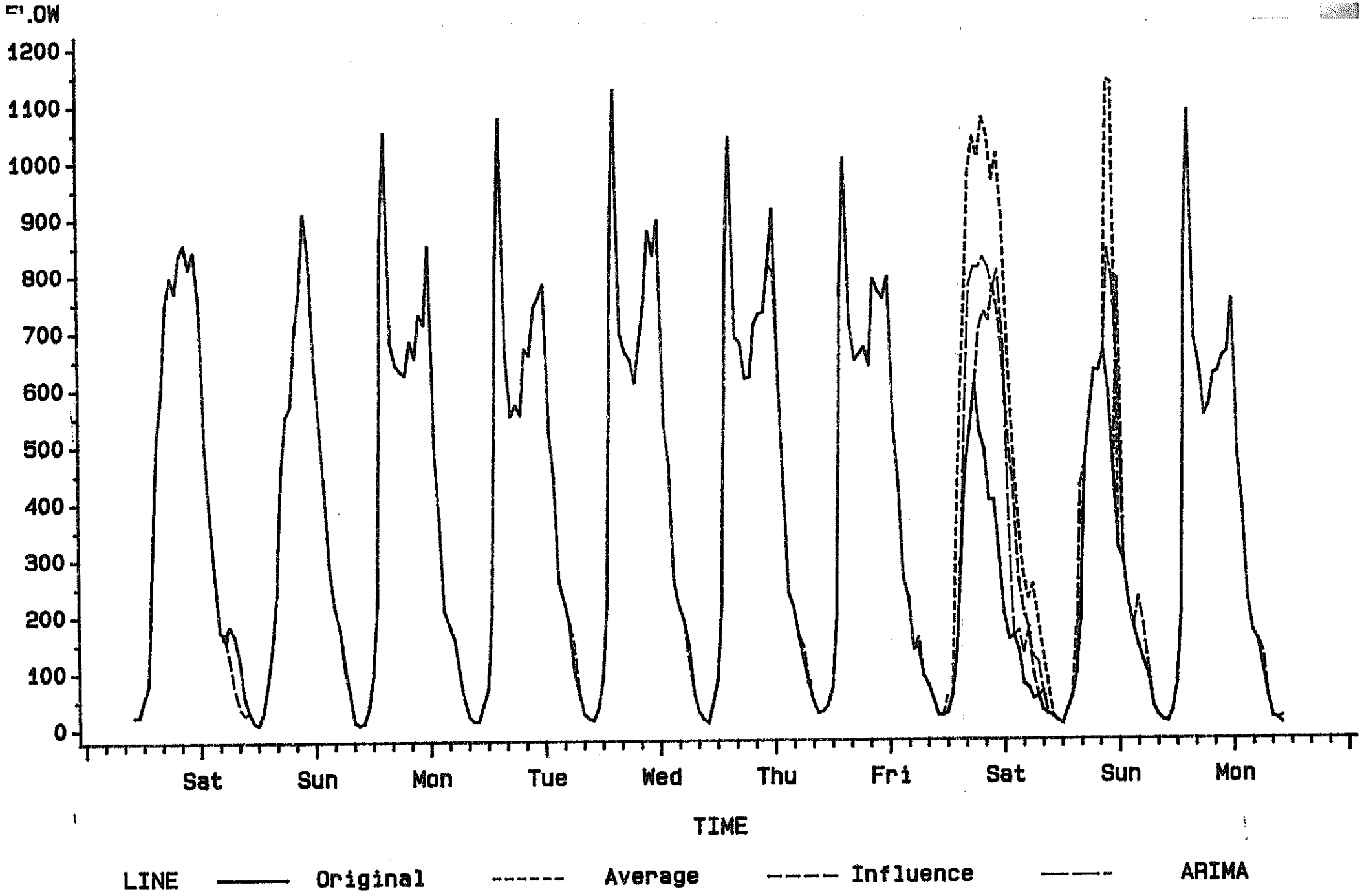

Figure 7 - Outliers detected and replaced for E497

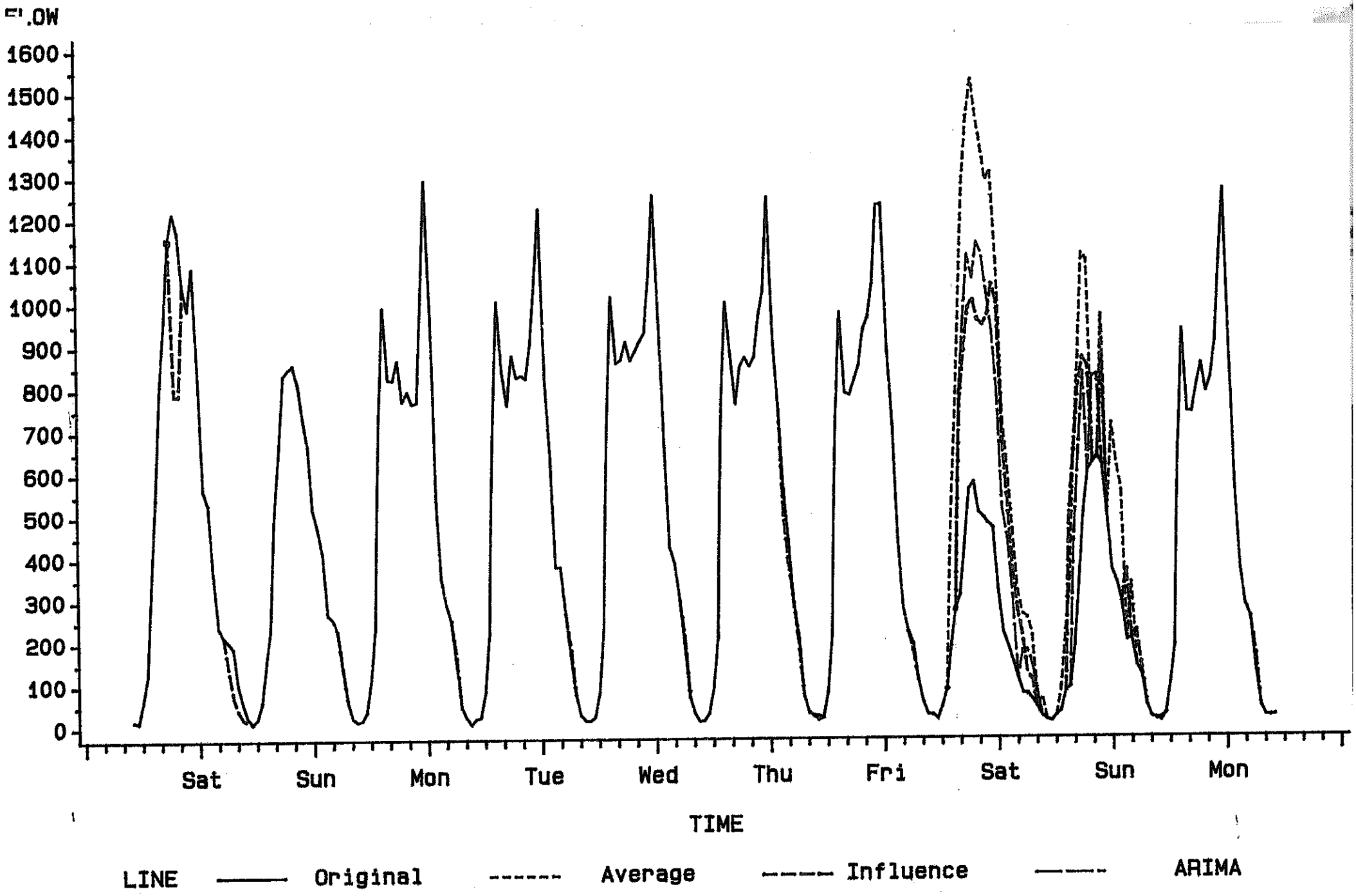

Figure 8 - Outliers detected and replaced for W498

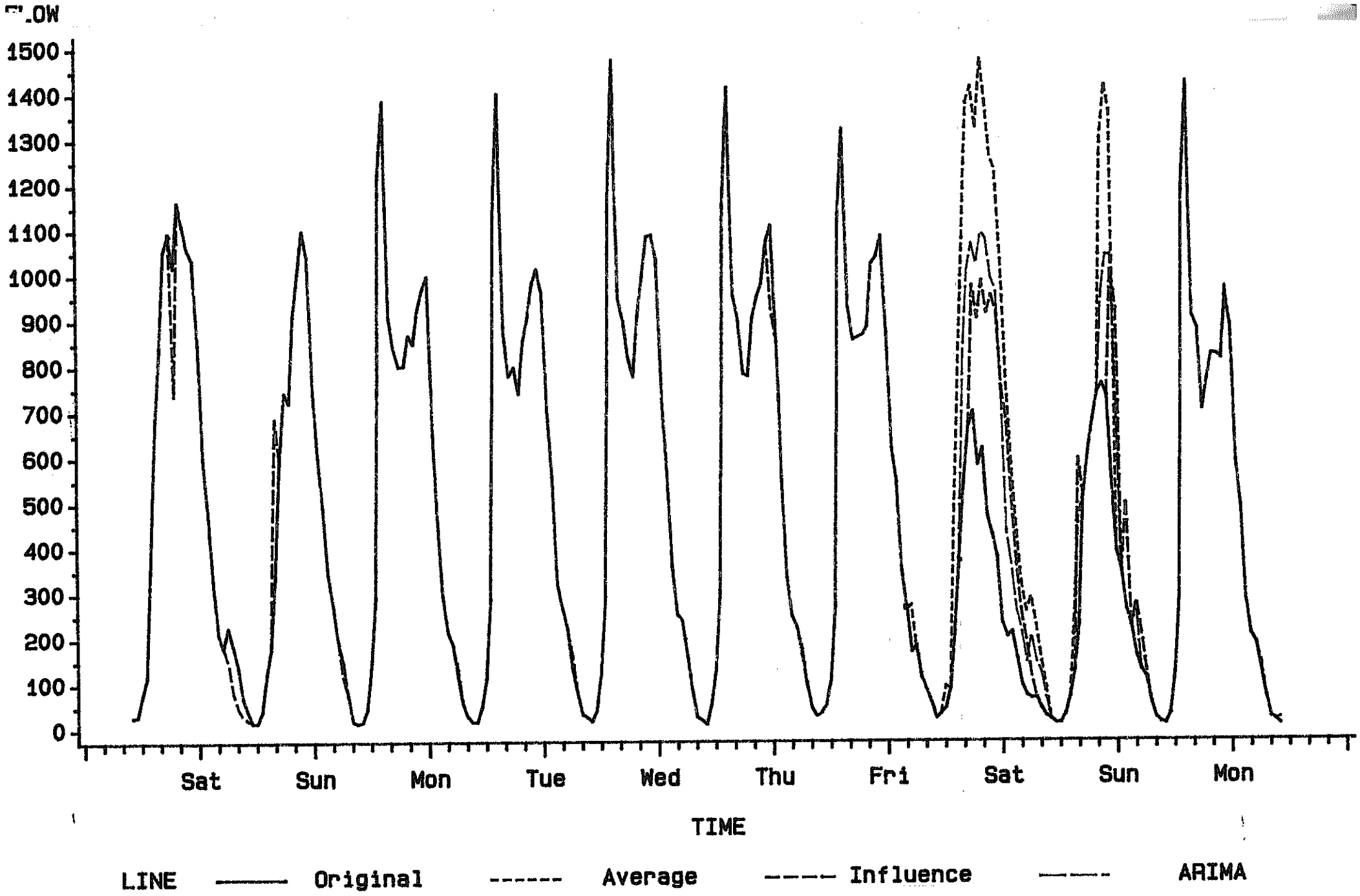

Figure 9 - Outliers detected and replaced for E499

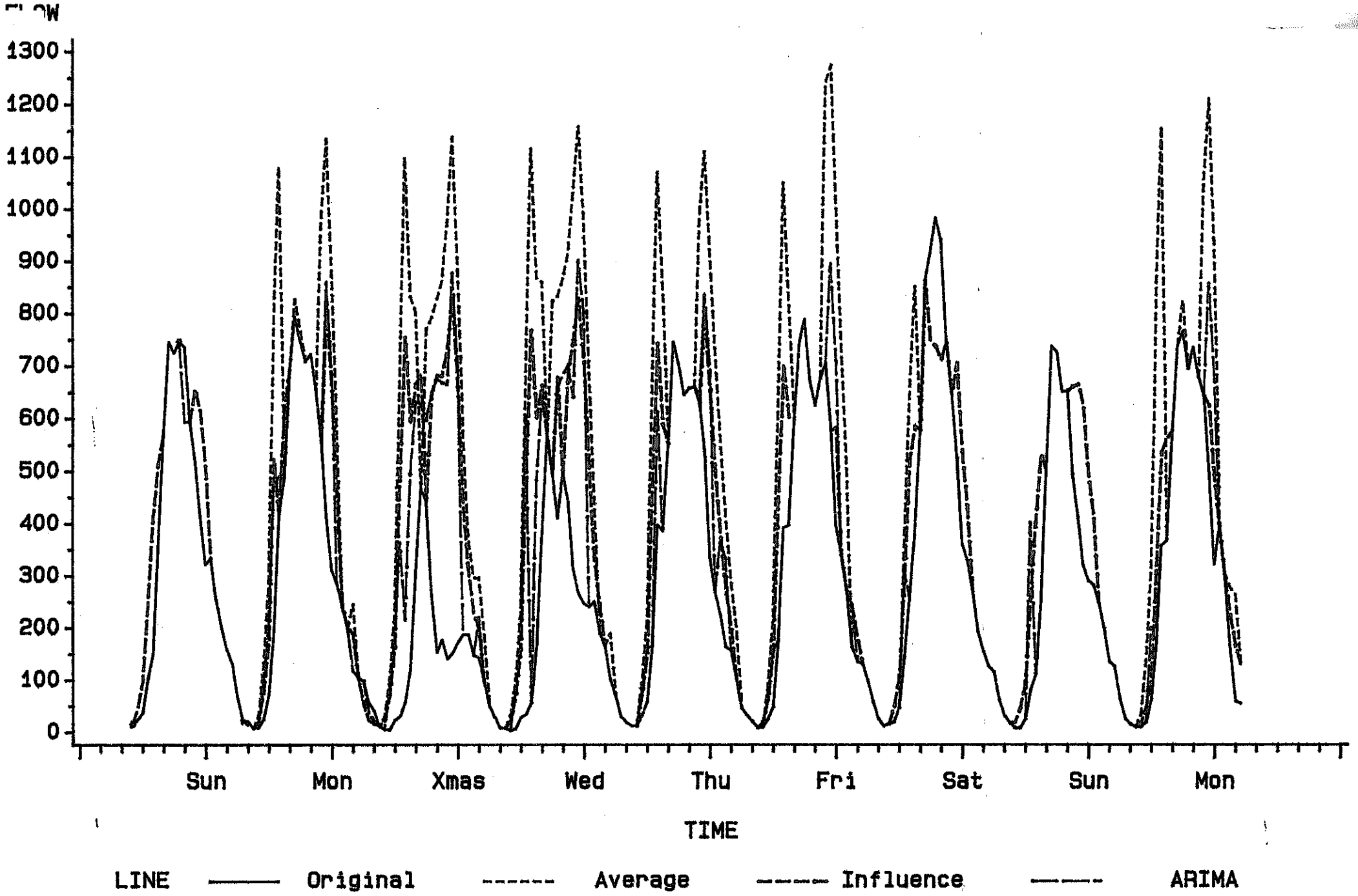

Figure 10 - Outliers detected and replacements for Christmas on W496

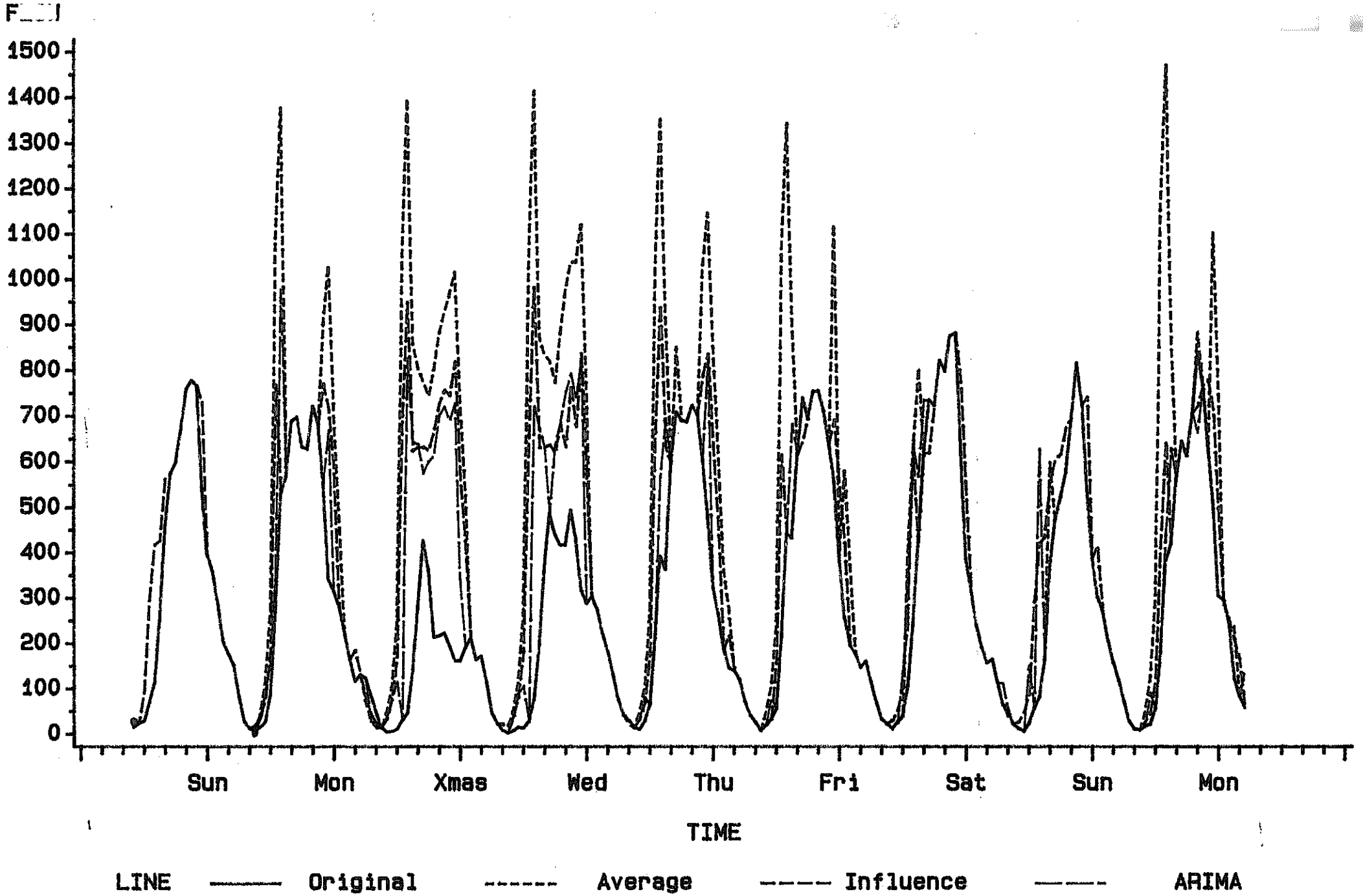

Figure 11 - Outliers detected and replacements for Christmas on E497

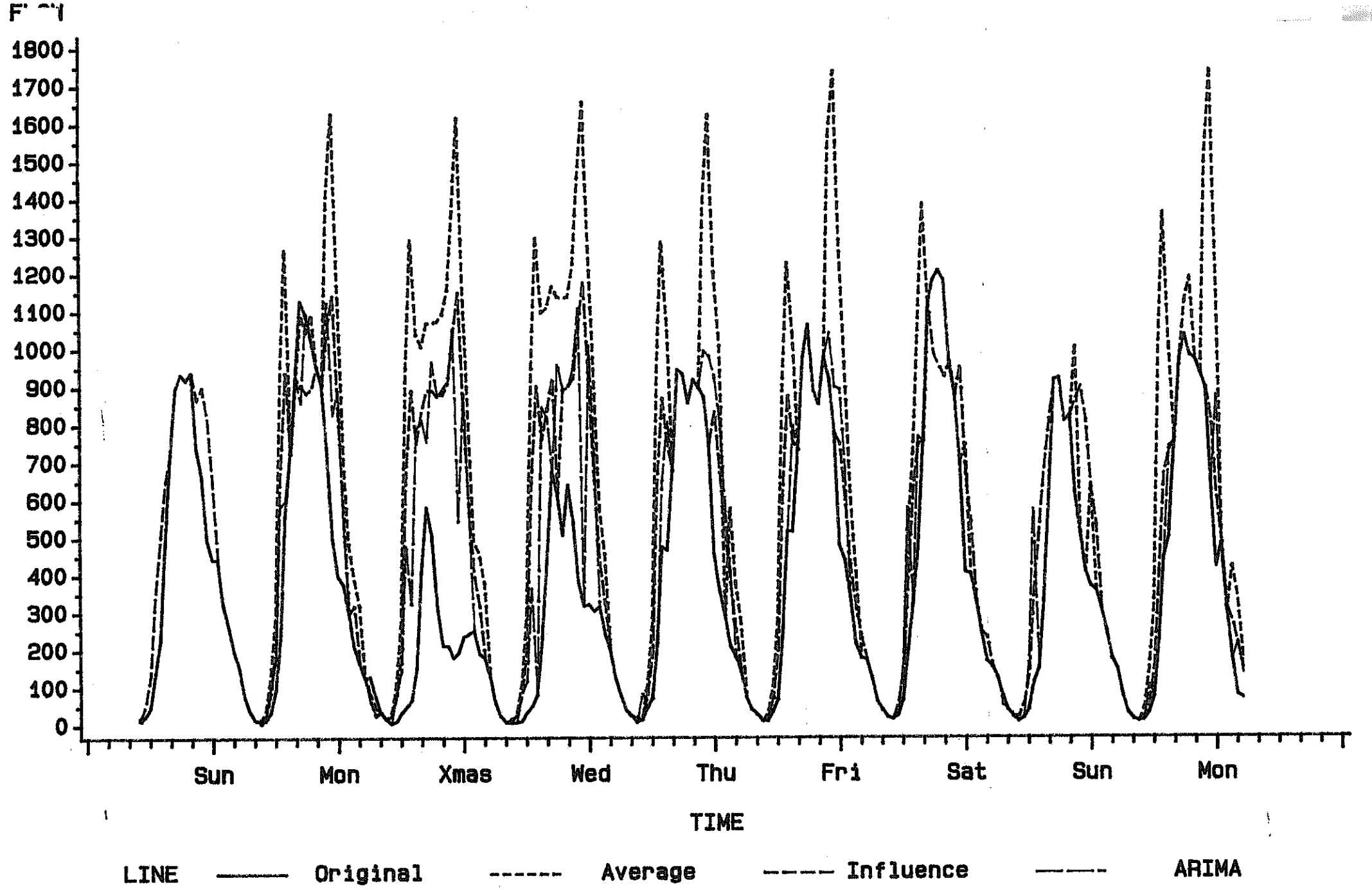

Figure 12 - Outliers detected and replacements for Christmas on W498

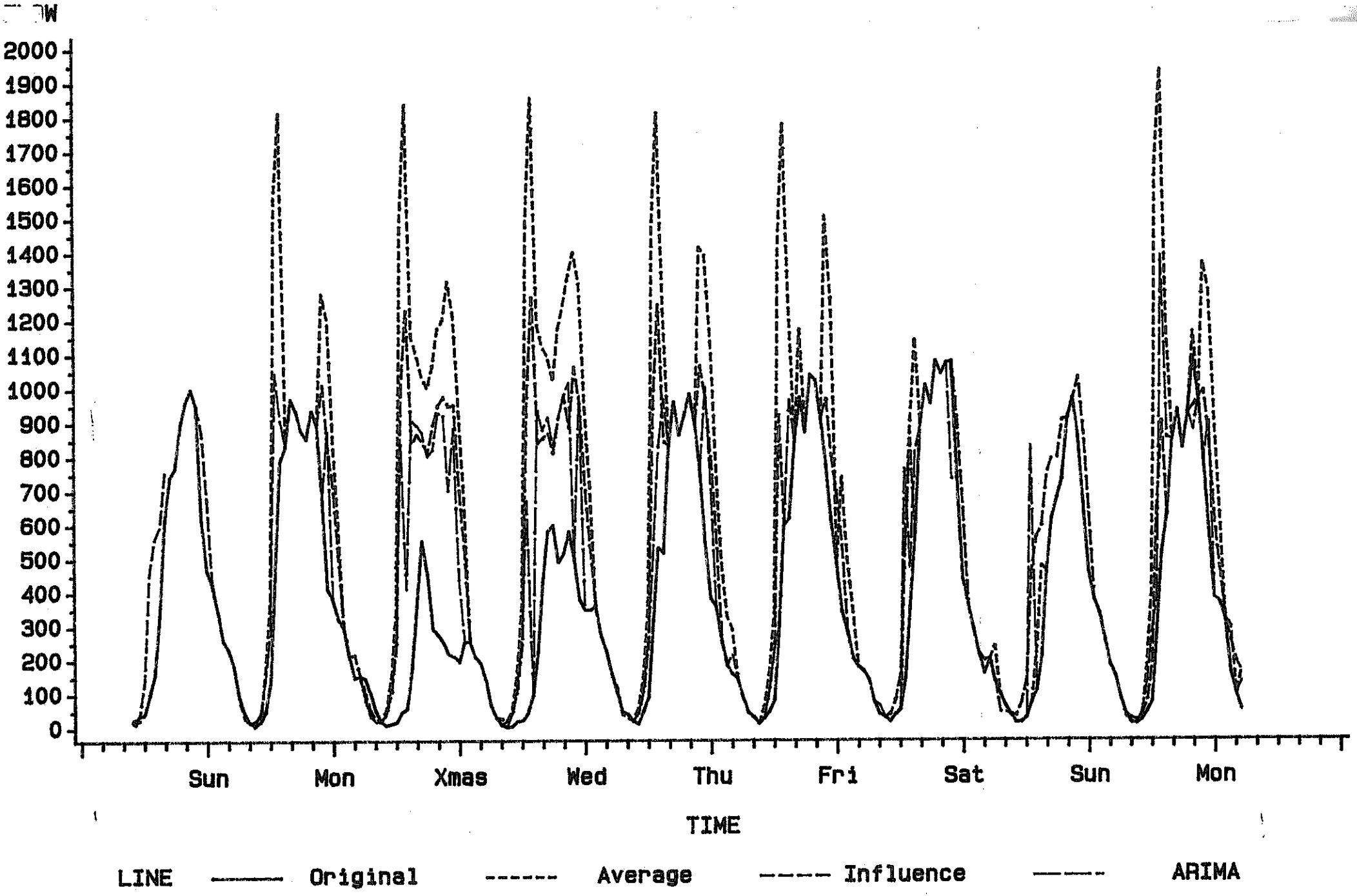

Figure 13 - Outliers detected and replacements for Christmas on E499

Figures 10-12 show the outliers detected and the replacement values for the Christmas period. Once again all the techniques have highlighted the atypical behaviour seen in the graphs. The absence of morning peaks in the easterly series and evening peaks in the westerly series is strongly detected. The replacements suggested by the averaging technique are on the large side. The maximum peak flows which can be reasonably expected for each of the series are W496 (1000), E497 (1200), W498 (1300) and E499 (1600). During the peak hour these flows are always exceeded by the averaging technique. The results from the application of the ARIMA and influence techniques are similar. Both attempt to estimate the peak hour flows during the early part of the week but this tendency diminishes towards the end of the week. Clearly both methods are adapting to the new situation without peak hour flows. In this case this is inappropriate since peak hour flows will recommence early in the new year but in other situations this adaptive behaviour may be an advantage.

4.2 HIGHWAYS, ENGINEERING AND TECHNICAL SERVICES DATA

Much of what was said in the previous section on DOT data is immediately relevant to this data set. The averaging technique used 7*24=168 average and variability measures. The types of ARIMA models employed were similar to those for the DOT data. For the influence technique the reversed analysis was not conducted.

Of most interest in this data set is the length of the series (365 observations in each series) which spans an entire calendar year. Also there is a large gap of missing observations in both series. Thus whilst the DOT series measured the ability to detect and correct outliers, this series measures the ability to estimate long spans of missing observations.

4.2.1 Comparison of results

Figure 14 illustrates how the techniques coped with an atypical Monday followed by a missing day of data on the following Wednesday.

Figures 15-17 show how the techniques coped with a long span of missing observations. Figure 15 is immediately followed on by Figure 16 which in turn is followed on by Figure 17. In each of these series only the first and last day's original flows are present. Figures 18-20 present results for southbound traffic, which correspond to those presented for northbound traffic in figures 15-17.

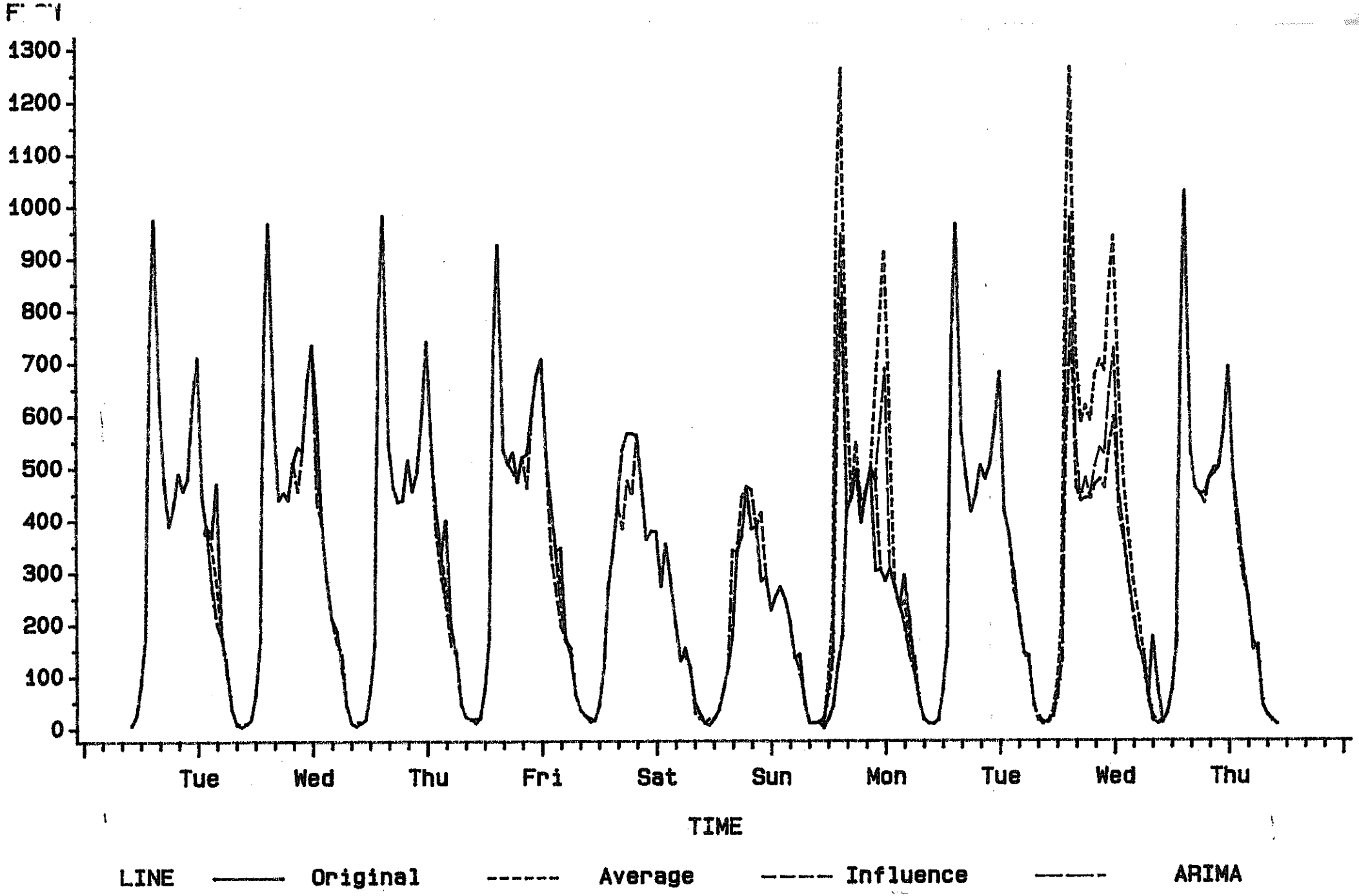

Figure 14 - Outliers detected and replacements for northbound traffic

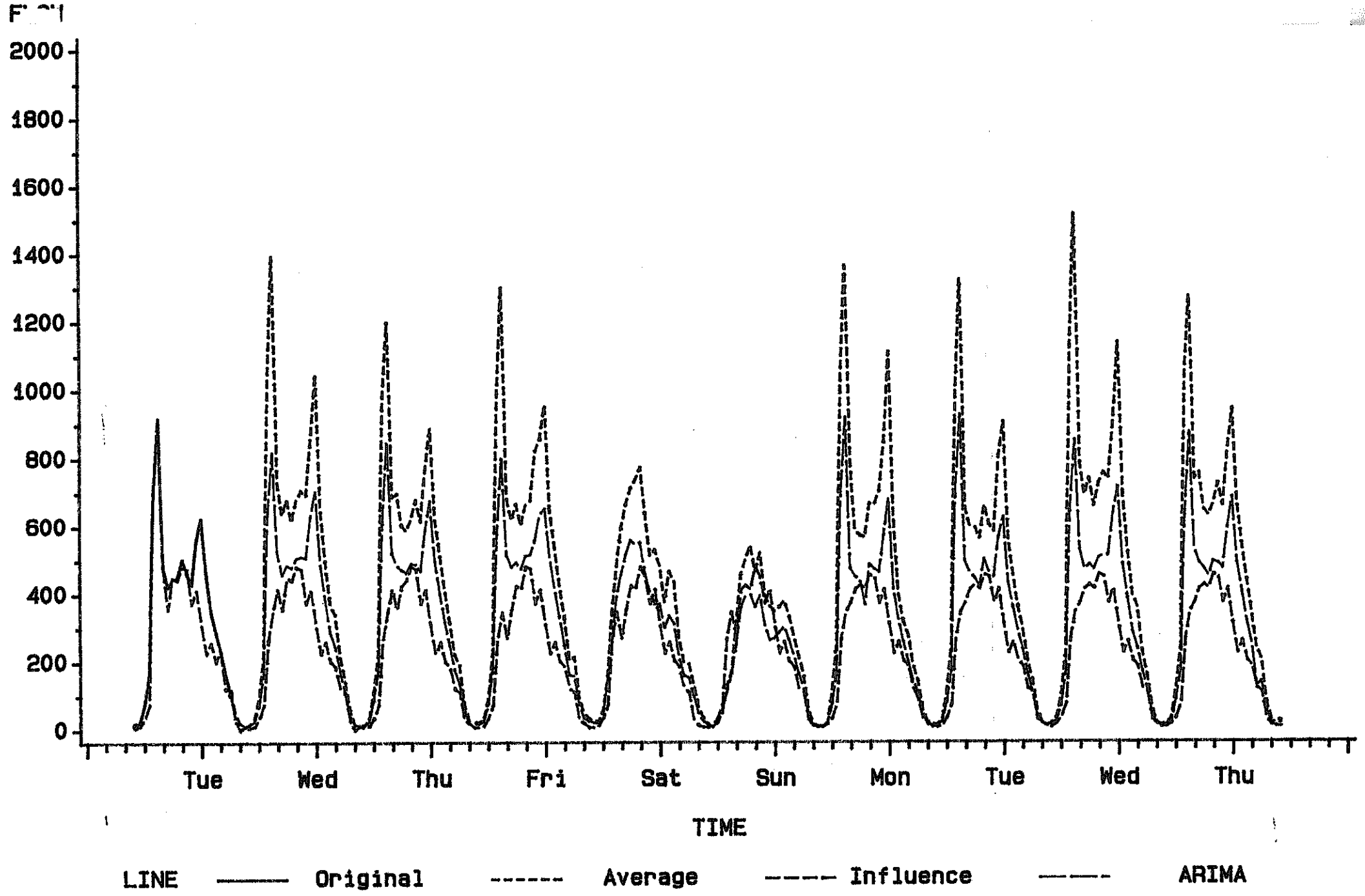

Figure 15 - Replacements for northbound traffic - part 1

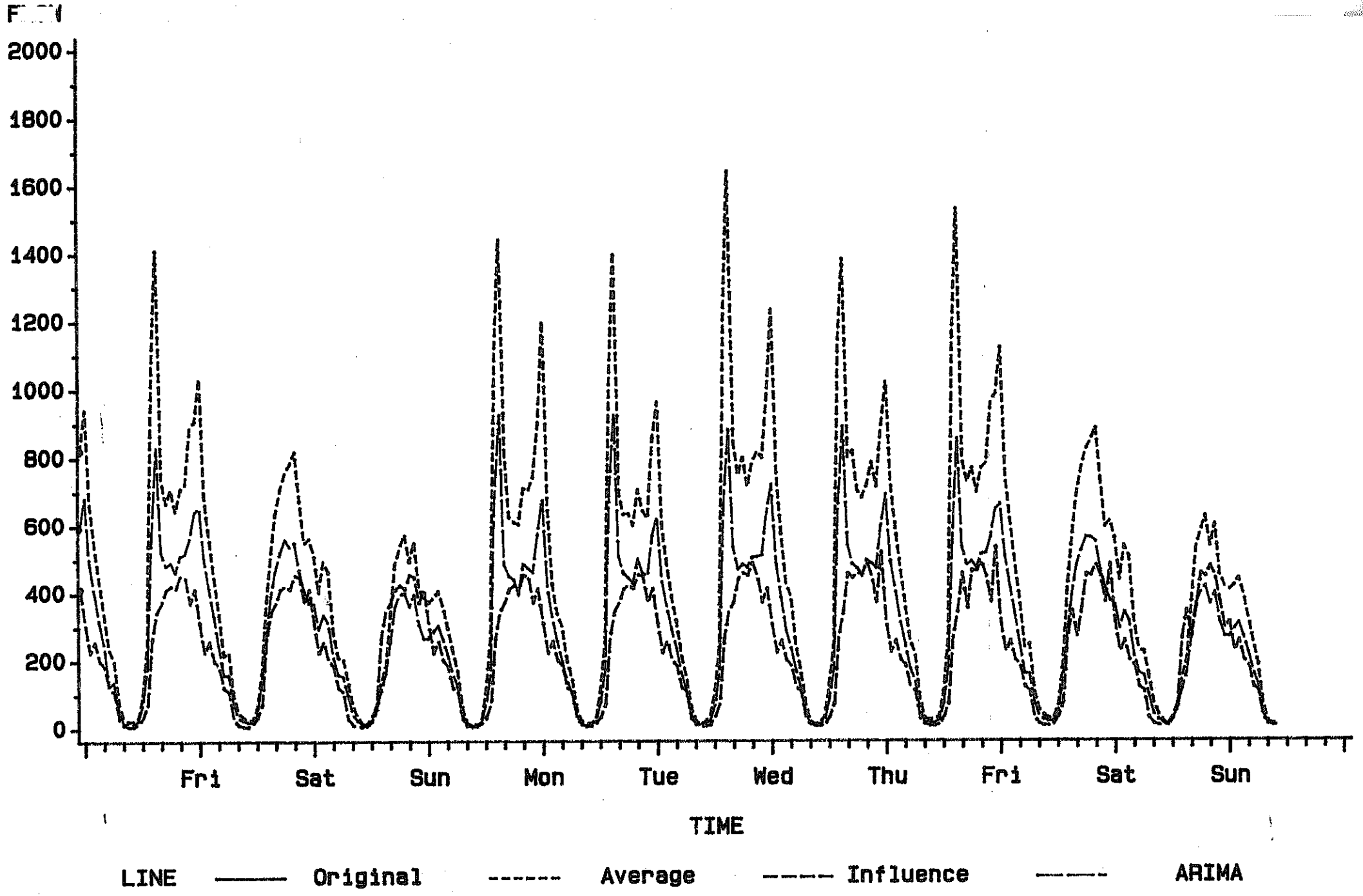

Figure 16 - Replacements for northbound traffic - part 2

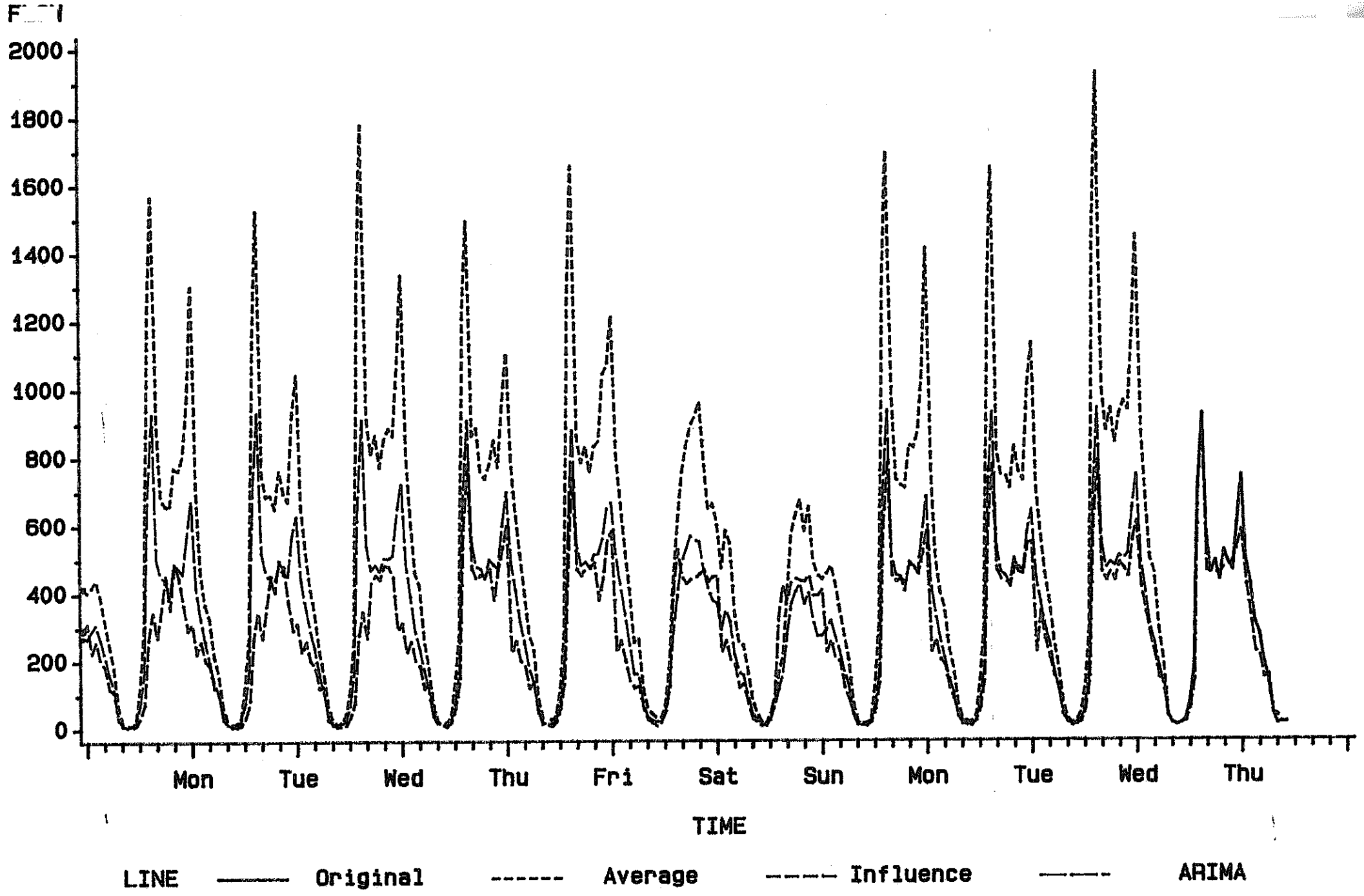

Figure 17 - Replacements for northbound traffic - part 3

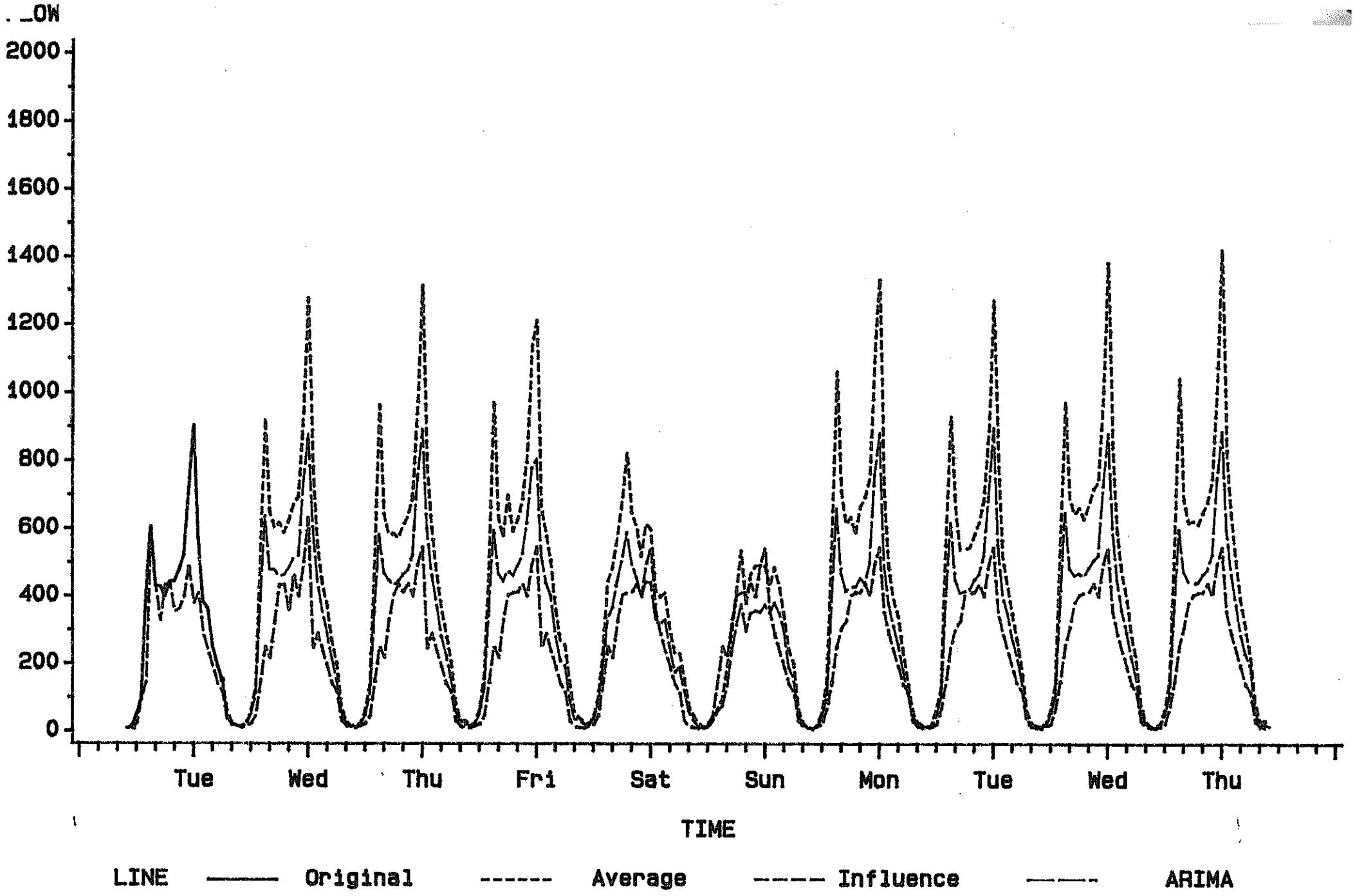

Figure 18 - Replacements for southbound traffic - part 1

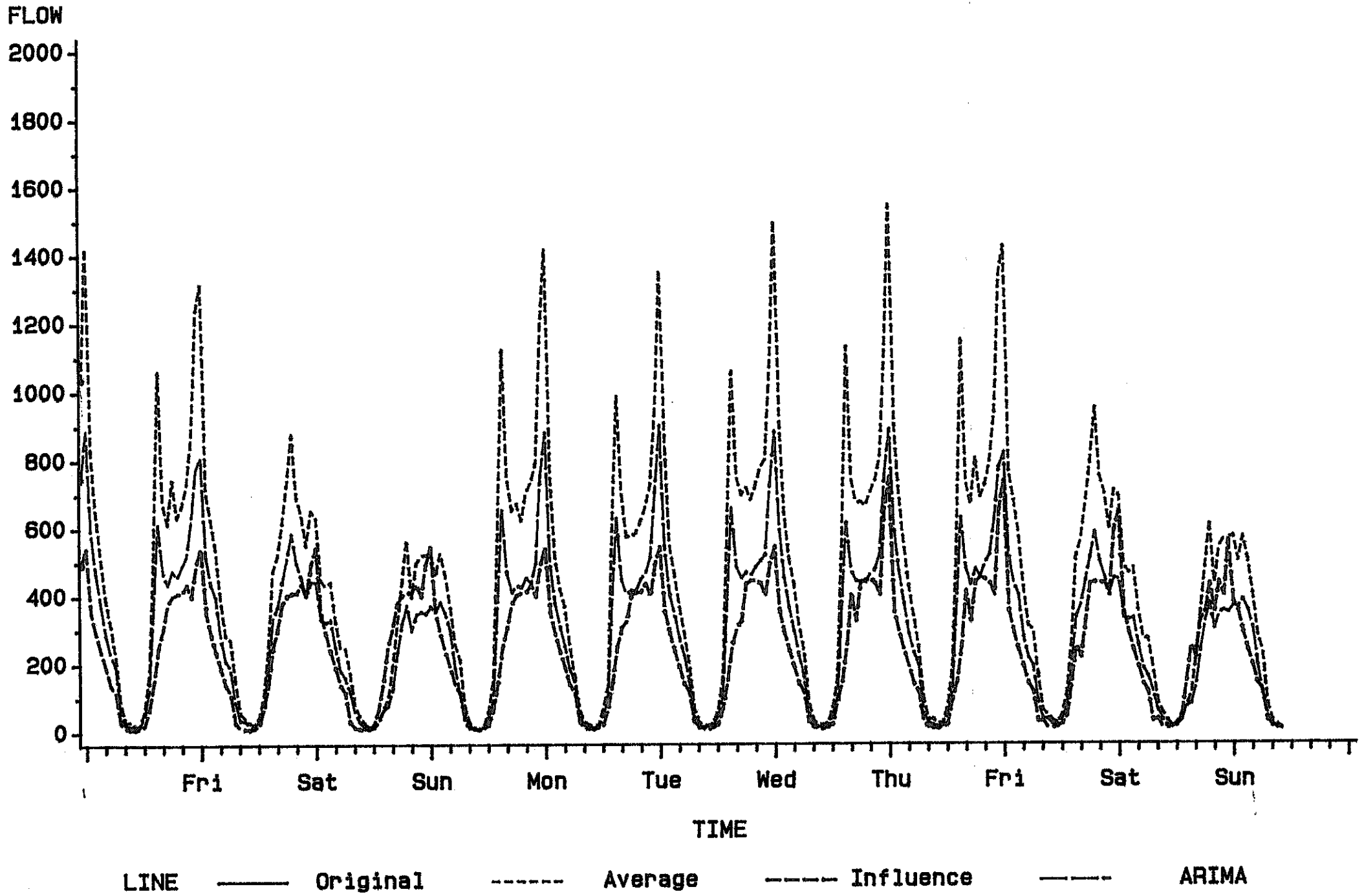

Figure 19 - Replacements for southbound traffic - part 2

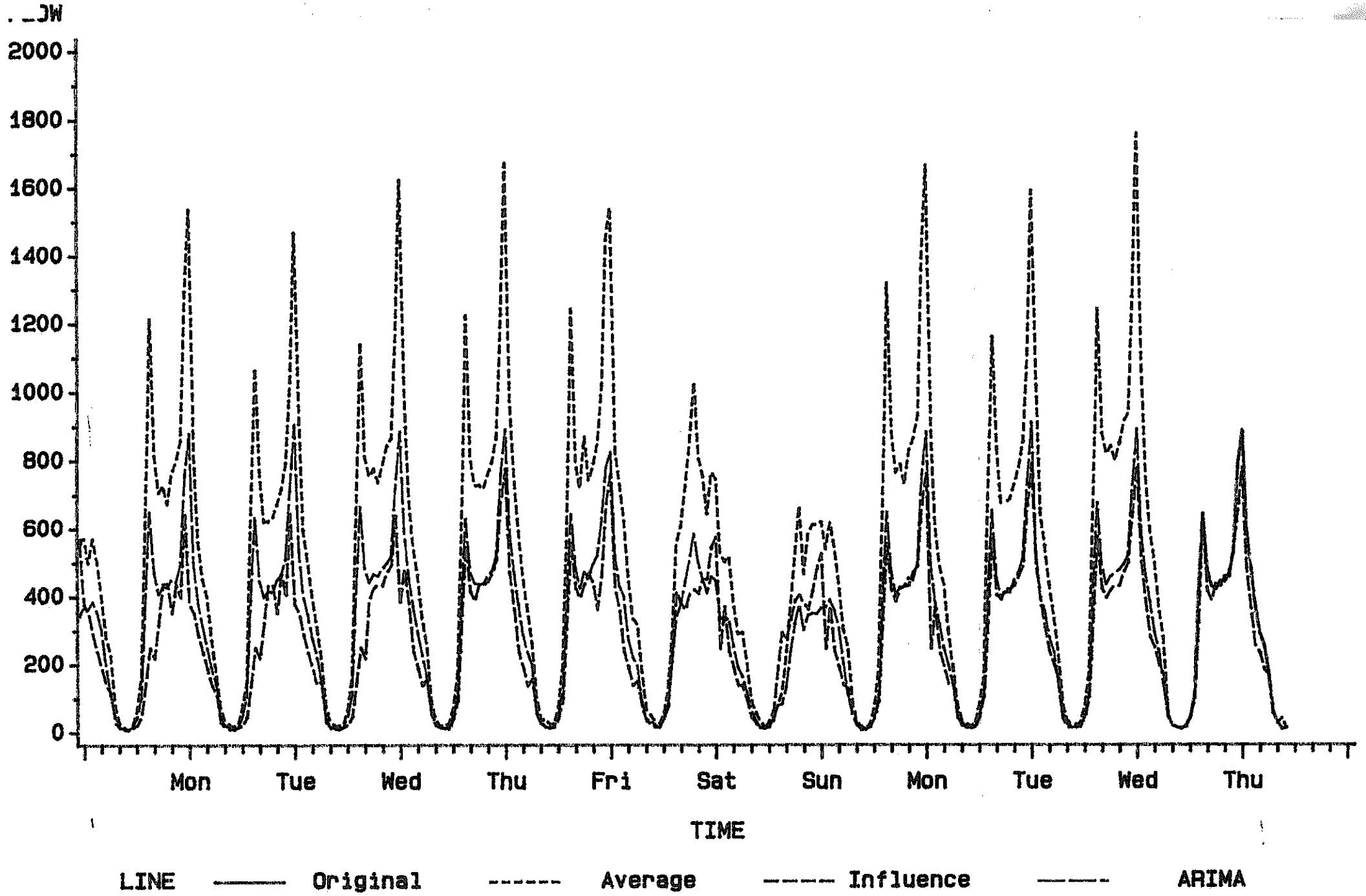

Figure 20 - Replacements for southbound traffic - part 3

As has been seen previously in this study, the averaging technique has tended to produce overestimates for the missing observations. Indeed these estimates increase over time, with the maximum peak hour flow (usually a Wednesday) going up from 1450 at the start to 1970 at the end of the period of northbound missing observations. The influence technique has failed to consistently reproduce the typical weekday flow profile. Only in the last week of the four weeks of missing data is a strong peak flow present in the estimates, elsewhere the morning or evening peaks are missing. This is a result of the z_{t+k} term used to estimate the replacement value in equation (15) of appendix 1c. Thus the *quality* of the estimate at time t depends upon the quality of the observation at time t+k. The ARIMA method has performed best with reasonable replacement values for the typical flow profile over a day.

4.3 LEICESTER DATA

A large number of Leicester series were analysed, 320 in total, with all three methods.

4.3.1 Averaging technique

The application of this technique to the Leicester data required the maintenance of only one average and variability measure and three observations to prime these measures. Of the three techniques considered this one tended to produce the largest number of outliers in a series. One serious drawback to this technique is the tendency to lose control and replace long runs of observations with exponentially increasing replacements which are entirely inappropriate (see Figures 26 and 27). This is most likely to occur from the start of the series or when there are long runs of missing or outlying values. Doubling the priming period to six observations should eliminate the problem from occurring at the start of the series (compare figure 26 with figure 27).

4.3.2 ARIMA technique

An ARIMA(0,1,2) model was found to be the most consistent performer across all the 320 series which were analysed. Often the θ_2 term was (just) insignificant and could easily have been eliminated from the model without losing much of the model's descriptive power. Only on those occasional cases where its presence stopped the convergence of the estimation procedure was the parameter removed from the model. The interpretation of this model form arises from the fact that an ARIMA(0,d,1) model is equivalent to an ARIMA(p,d,0) with an infinite number of autoregressive terms, ie p is large. The parameter value associated with these high order autoregressive terms will decrease, provided θ_1 is less than 1.0. In essence such models use as much past data as is available to predict the correct observation.

The length of the series did tend to influence how many outliers were detected. For the days of short series (26/04/91, 01/05/91 and 03/05/91) only in the region of 2 or 3 observations were detected as outliers in the 40 series (0.2% of observations), whilst for the longer series, in the region of 40 observations were detected (0.7%).

4.3.3 Influence technique

The influence technique used the first eight autocorrelations from the series to detect outliers and suggest replacements. The technique tended to produce more outliers than the ARIMA technique but just less than those detected by the averaging technique.

4.3.4 Comparison of results

The results which follow provide a fair representation of the findings. Figure 21 demonstrates all three methods on a long time series. Most replacements are made in the first part of the series which is reproduced in more detail in Figure 22. Figure 23 shows results for another long series.

The averaging technique tends to detect trough outliers, which neither of the other two techniques do, and replace them with peak outliers (see Figure 21 and Figure 23 beyond observation 35). The influence technique detects a large number of outliers in the early part of the series and suggests replacements which make the peak resemble the non-peak flow during the remainder of the day. The estimates for the shorter periods of missing values are all reasonable. The ARIMA technique functions well, most of the outliers it detects are the strong peak five or ten minute flows at the start of the series. The replacements for outliers and estimates of missing values are reasonable, with the possible exception of observation 48 in figure 24.

Figures 24 and 25 show how the techniques perform on shorter series. Figure 26 shows how the three techniques deal with a spell of eight missing values. The replacements produced by the ARIMA and influence techniques tend to be less *jagged* than the surrounding series. The average method has performed badly, with the replacement values increasing exponentially over the patch of missing values and this effect being carried through to replacements for subsequent outliers. A priming period of six observations was used for this exercise. If the priming period is reduced to three observations then Figure 27 results. Problems start at observation 5 and thereafter the replacements increase exponentially. The replacement value at the end of the patch of missing values (observation 101) is 226,116.

5. SUMMARY AND CONCLUSIONS

The application of three fundamentally different techniques to three sets of transport data has produced interesting results. An averaging method tended to detect unreasonable outliers and also suggested replacements which were consistently larger than could reasonably be expected. A technique based upon a point's influence would detect outliers well but the replacement values were suspect when there was a long run of such outlying values. This techniques also produced poor estimates when a long run of missing values needed estimating. The technique which appeared to perform best both in detection, replacement and estimation was that based on an ARIMA model for the data. Few spurious outliers were detected and the replacement and estimated values were always reasonable.

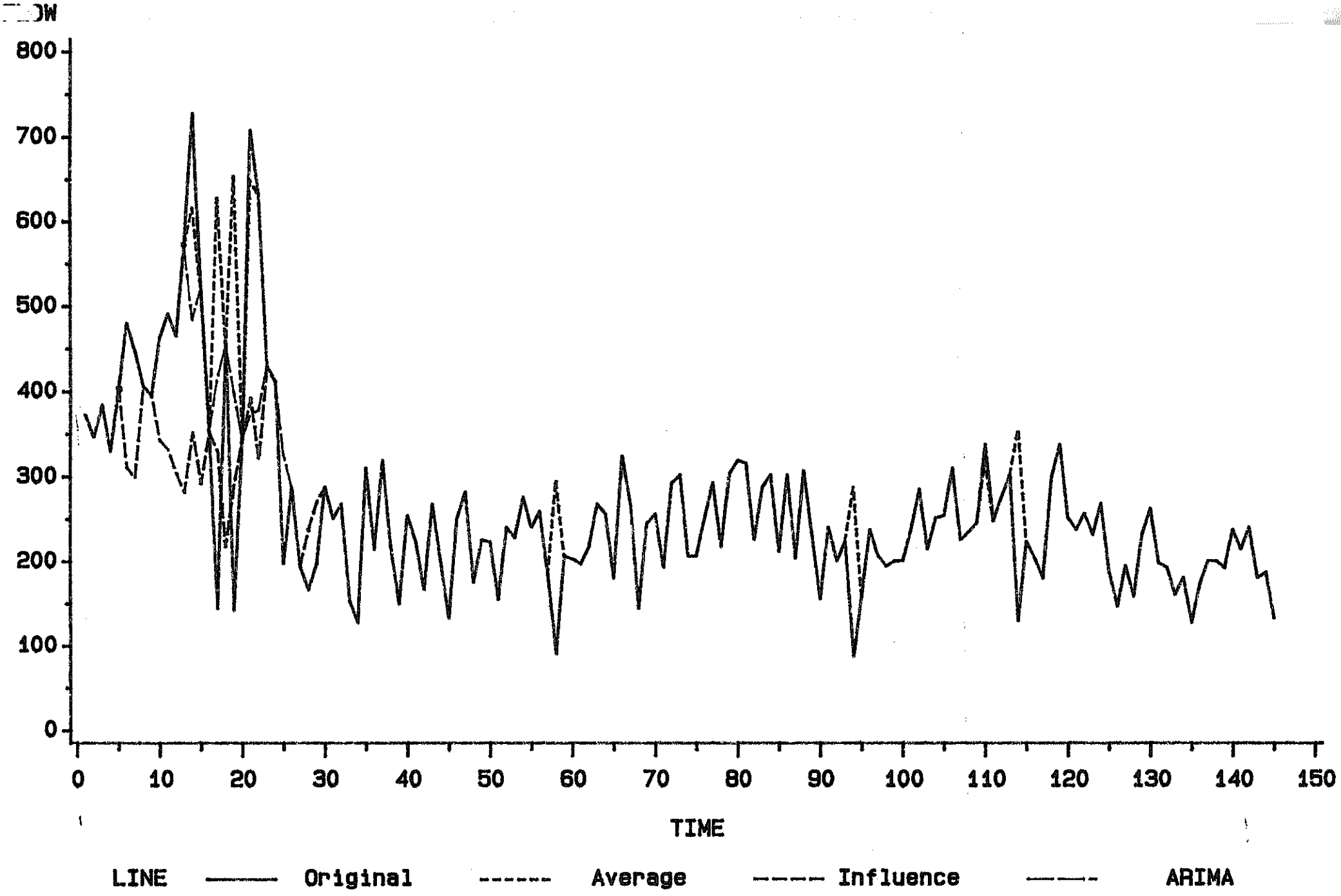

Figure 21 - Link 1146C on 08/05/91

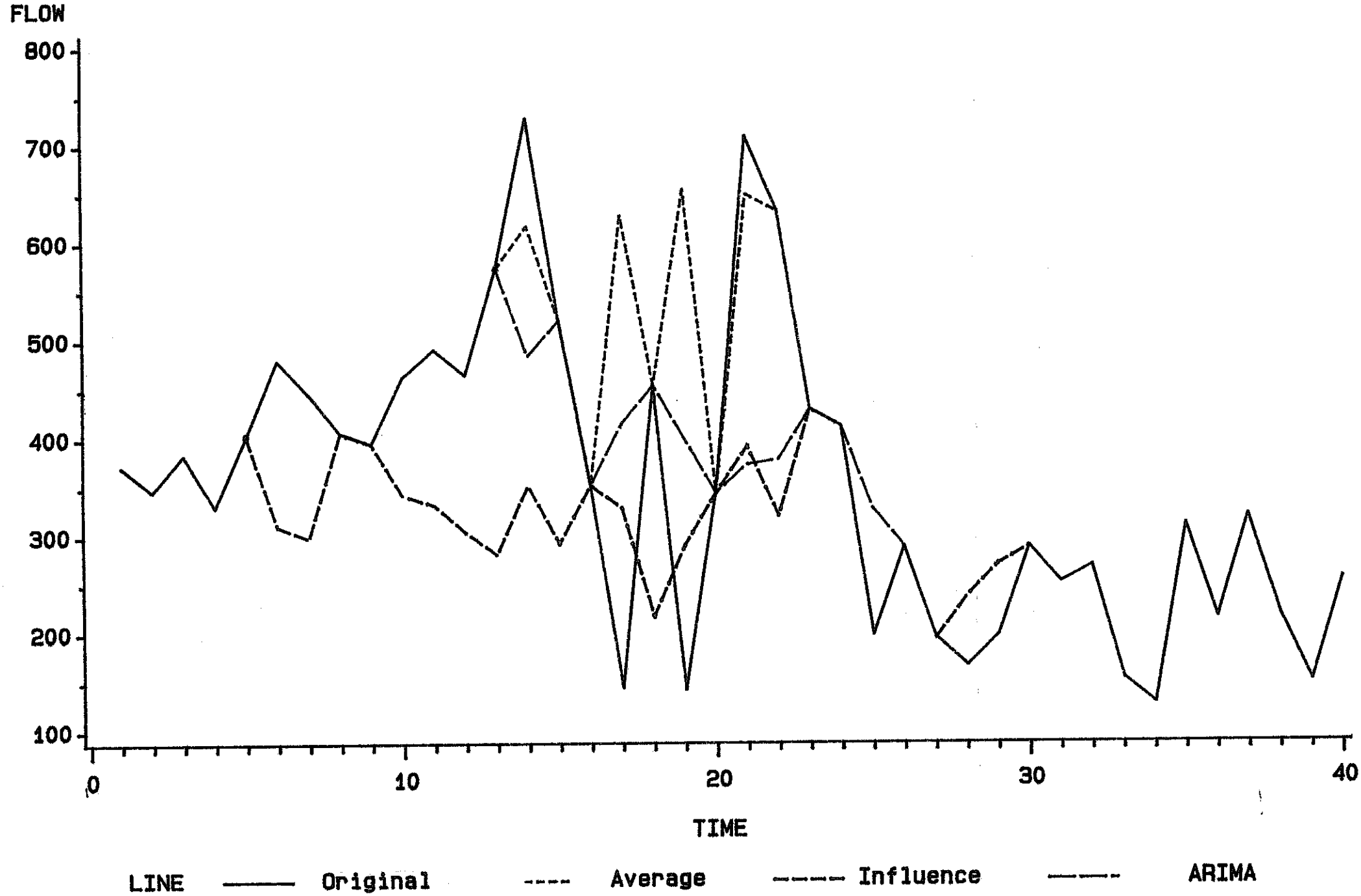

Figure 22 - Link 1146C on 08/05/91 - First 35 observations

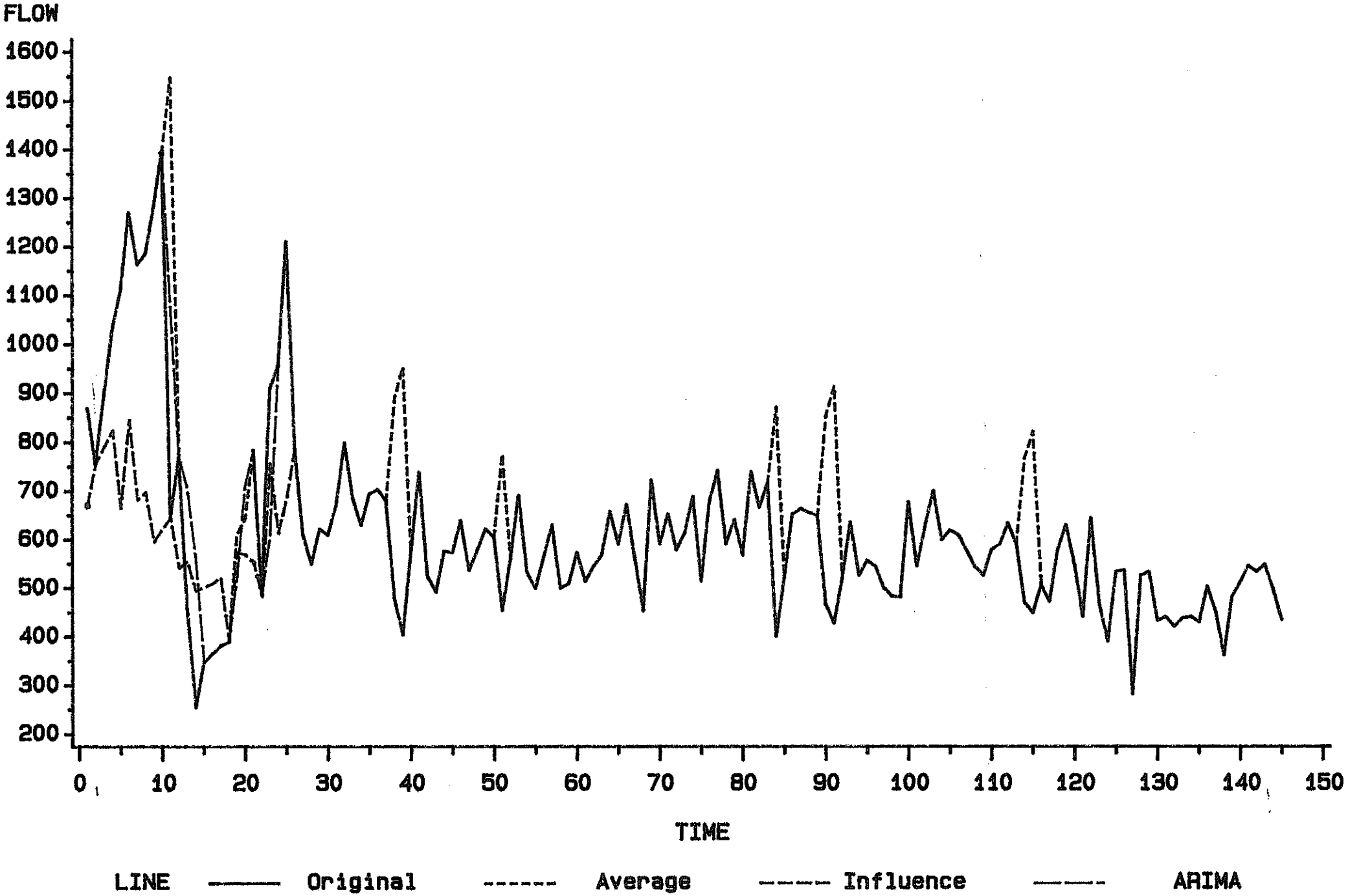

Figure 23 - Link 1142A on 08/05/91

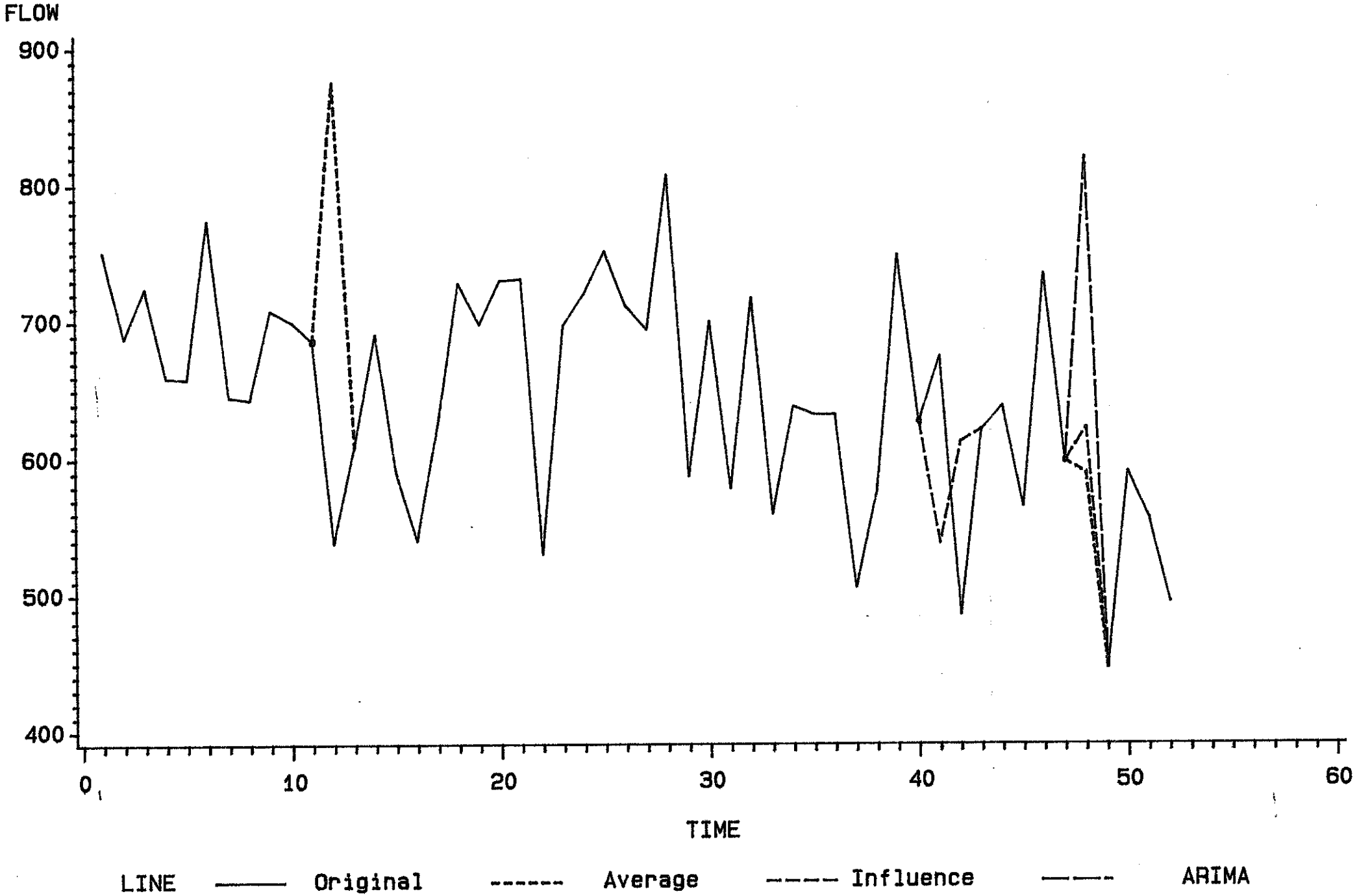

Figure 24 - Link 1222M on 03/05/91

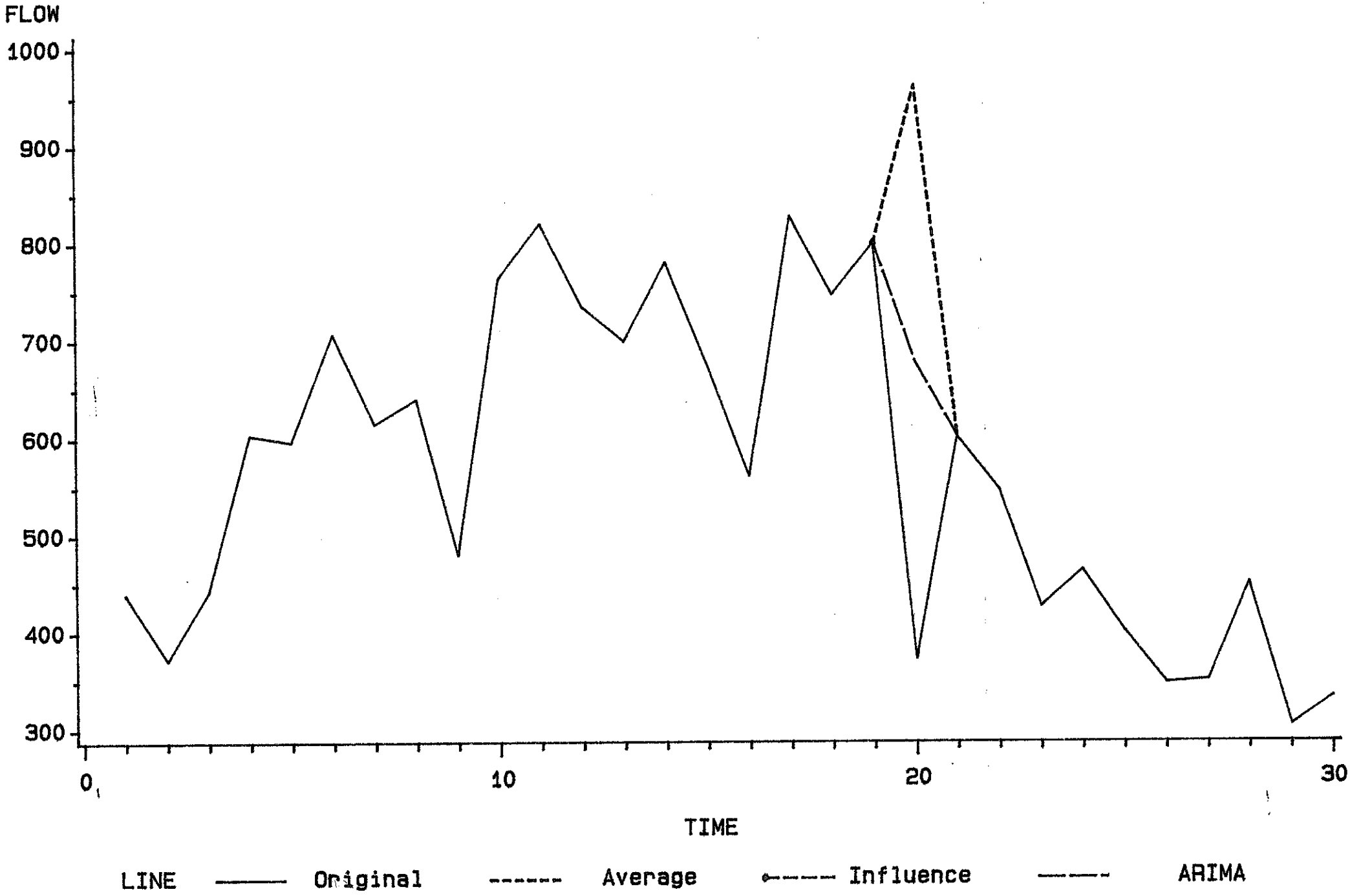

Figure 25 - Link 1321S on 26/04/91 - Shorter priming

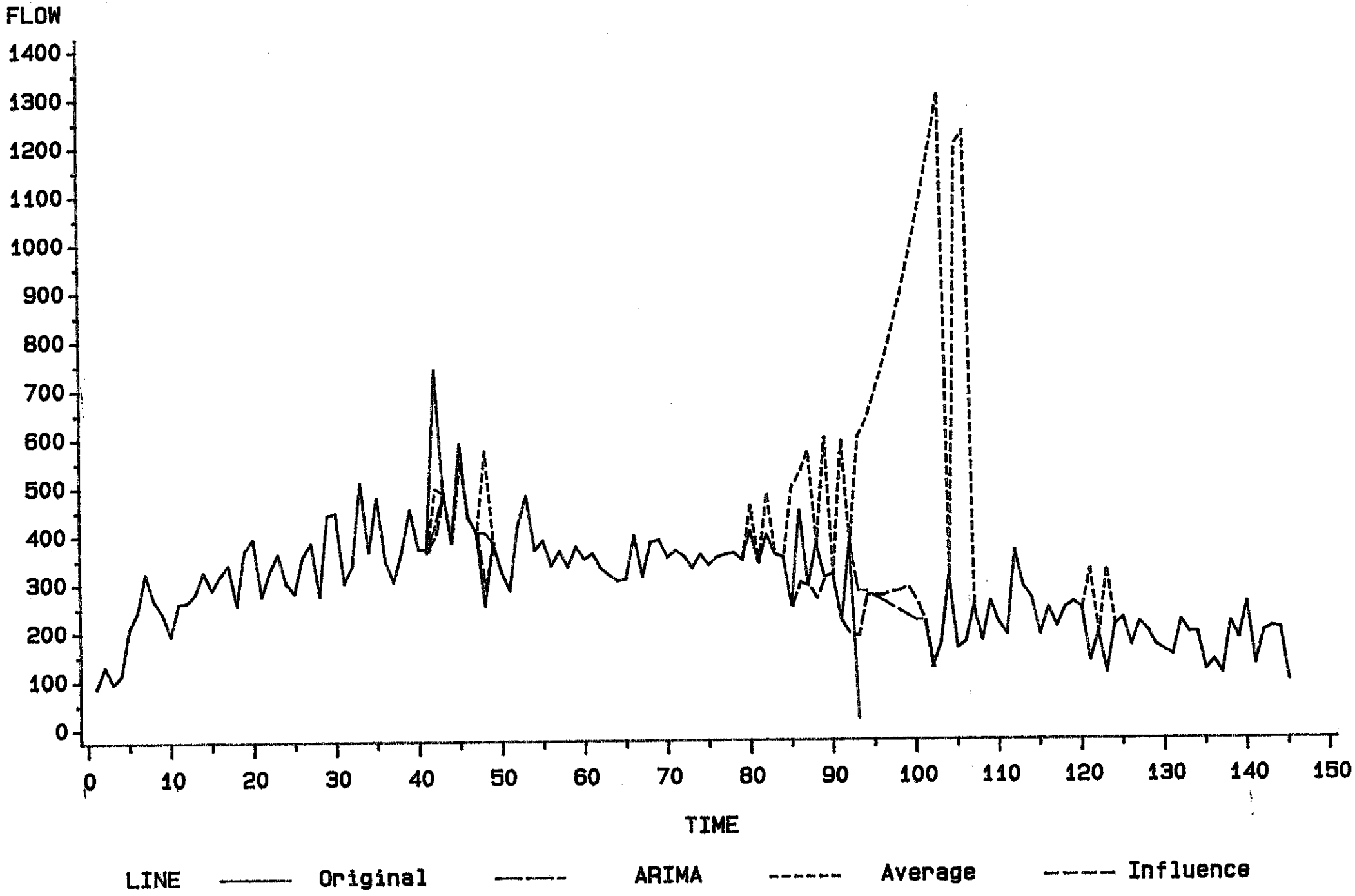

Figure 26 - Link 1146Q on 06/04/91 - Patching of missing observations

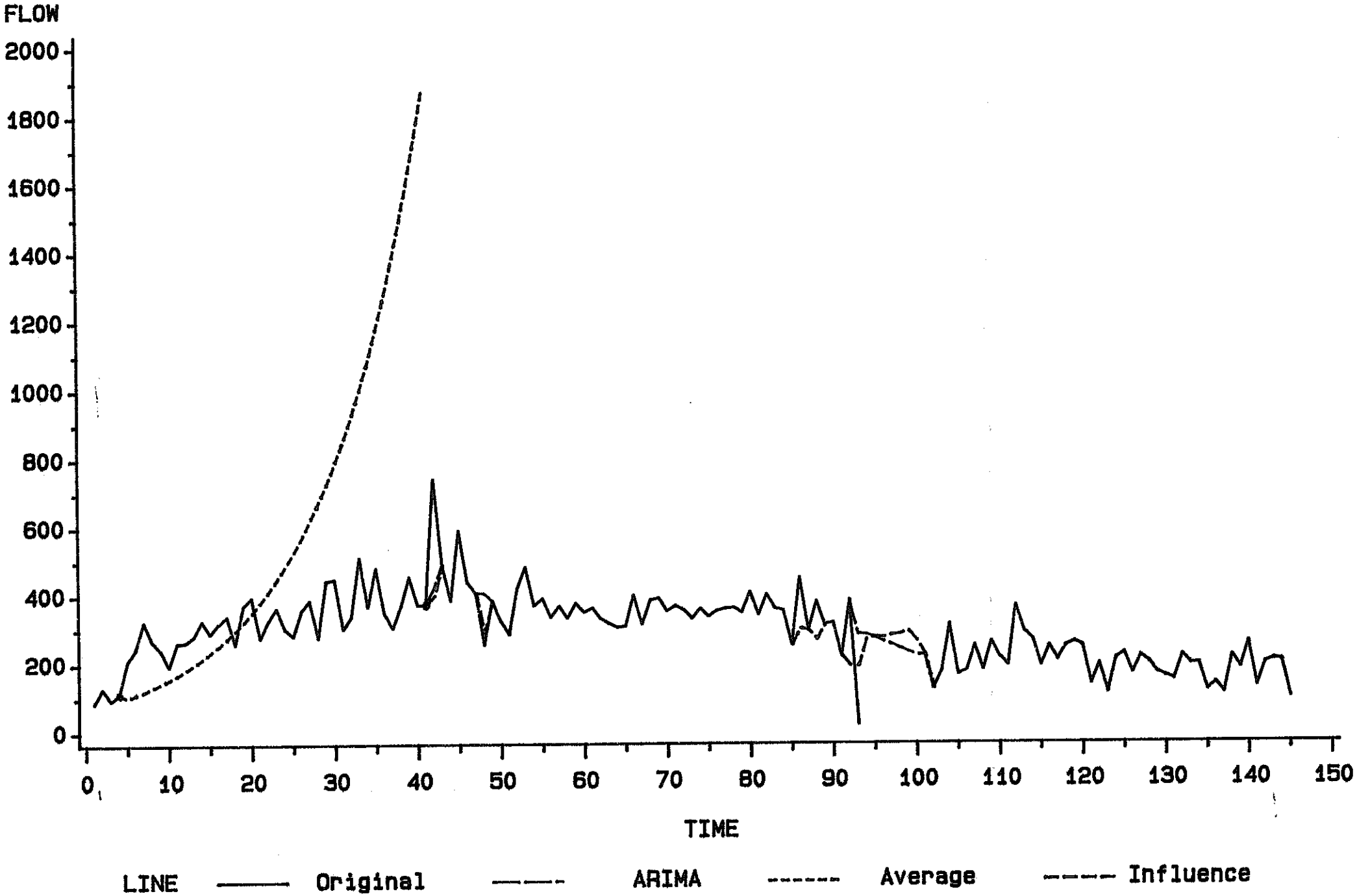

Figure 27 - Link 1146Q on 06/04/91 - Shorter priming

REFERENCES

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

Chernick, M.R., Downing, D.J. and Pike, D.H. (1982). "Detecting Outliers in Time Series Data". J.A.S.A., 77, 743-747;

Redfern, E.J., Watson, S.M., Clark, S.D., Tight, M.R., Payne G.A. (1992). "Modelling outliers and missing values in traffic count data using the ARIMA model". ITS Working Paper, University of Leeds;

Tsay, R.S. (1992). "Model Checking via Parameter Bootstraps in Time Series Analysis". J.R.S.S. C, 41. 1-15;

Watson, S.M. (1987). "Non-normality in Time Series Analysis". Unpublished Ph.D. Thesis, Trent Poly;

Watson, S.M., Clark, S.D., Redfern, E.J., Tight, M.R. (1992a). "Outlier Detection and Missing Value Estimation in Time Series Traffic Count Data". Proceedings of 6th World Conference on Transport Research;

Watson, S.M., Clark, S.D., Redfern, E.J., Tight, M.R. (1992b). "Development of an influence statistic for outlier detection with time series traffic data". ITS Working Paper 366, University of Leeds;

Appendix 1a - Averaging technique

An automatic procedure used by the British Department of Transport is to validate each observed data value against corresponding data from the same site 1, 2, or more weeks previously. An average (μ) and standard deviation (σ) for each site*day of week*period of day*combined vehicle category is updated using an exponentially weighted moving average (EWMA), where the lagged values, x_{t-k}, are the previous estimates from the corresponding time of day and week for the same site and vehicle category:

$$\mu_{s,t} = (1-\theta)\,\mu_{s,t-1} + \theta\, x_t$$
$$\sigma^2_{s,t} = (1-\theta)\,\sigma^2_{s,t-1} + \theta\,(x_t-\mu_{s,t})^2 \qquad (1)$$

where s is the seasonal component and θ is the smoothing parameter

Typical values used for the smoothing parameter are about $\theta = 0.3$. Any observation more than 4 standard deviations away from the current mean is rejected and replaced. Missing or rejected data, $x_{s,t}$, is estimated by the exponentially weighted average:

$$x_{s,t} = \theta\, x_{s,t-1} + \theta(1-\theta)\, x_{s,t-2} + \theta(1-\theta)^2\, x_{s,t-3} + \ldots \qquad (2)$$

Appendix 1b - Box-Jenkins technique

Box-Jenkins (1976) is a widely applied approach and has been used with some success in modelling transport data. The principal method involves building AutoRegressive Integrated Moving Average (ARIMA) models using the Box-Jenkins (1976) modelling methodology. Routines for fitting and validating the models are readily available in the major statistical packages such as SAS, BMDP and SPSS. The form of the model is given in the following shorthand : ARIMA(p,d,q) where p is the number of autoregressive terms, q the number of moving average terms and d the degree of differencing required.
Autoregressive elements of the model are of the form shown by the following examples:

$$ARIMA(1,0,0) \quad : X_t = \phi_1 X_{t-1} + \varepsilon_t$$
$$ARIMA(2,0,0) \quad : X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t \qquad (3)$$
$$ARIMA(1,0,0)^{13} \quad : X_t = \phi_{13} X_{t-13} + \varepsilon_t$$

Where ϕ_i is the i'th Autoregressive parameter
$\{\varepsilon_t\}$ is a noise series

Whilst the moving average form is for example:

$$ARIMA(0,0,1) \quad : X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1}$$
$$ARIMA(0,0,2) \quad : X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \qquad (4)$$
$$ARIMA(0,0,1)^{13} \quad : X_t = \varepsilon_t - \theta_{13} \varepsilon_{t-13}$$

where θ_j is the j'th Moving average parameter
$\{\varepsilon_t\}$ is a noise series

And the differencing is denoted as:

$$Z_t = \Delta^d X_t = (X_t - X_{t-d}) \qquad (5)$$

The full form is given thus:

$$Z_t = \phi_1 Z_t + \ldots + \phi_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \quad (6)$$

where $Z_t = \Delta^d X_t$

An alternative form of the model is to use the backshift operator B where $B^d X_t = X_{t-d}$. This allows the equation form to be more succinctly expressed. Thus equation (6) becomes:

$$(1 - \phi_1 B - \ldots - \phi_p B^p)\ Z_t = (1 - \theta_1 B - \ldots - \theta_q B^q)\, \varepsilon_t \quad (7)$$

$$\Phi(B)\, Z_t = \Theta(B)\, \varepsilon_t$$

where $\Phi(B)$, $\Theta(B)$ *are polynomials of* B

Extensions to include seasonal components are now easily made. The general form of such an ARIMA(p,d,q)(P,D,Q)^s model will be:

$$(1 - \phi_1 B - \ldots - \phi_p B^p)(1 - \Phi_s B^s)\ Z_t = (1 - \theta_1 B - \ldots - \theta_q B^q)(1 - \Theta_s B^s)\, \varepsilon_t \quad (8)$$

where $Z_t = \Delta^d \Delta^D X_t$

Thus for an ARIMA(1,1,1)(1,1,1)^{13} model the equation form is:

$$(1 - \phi_1 B)(1 - \Phi_{13} B^{13})\ Z_t = (1 - \theta_1 B)(1 - \Theta_{13} B^{13})\, \varepsilon_t \quad (9)$$

where $Z_t = \Delta\Delta^{13} X_t$

The form of model which is appropriate to the data is identified using the autocorrelation function (ACF):

$$r_k = \frac{\Sigma\ (Z_t - \bar{Z})\,(Z_{t+k} - \bar{Z})}{\Sigma\ (Z_t - \bar{Z})} \quad (10)$$

and the related functions of the partial autocorrelation function (PACF) and the inverse autocorrelation function (IACF). The behaviour of these functions is compared with the theoretical behaviour given in the following table:

Model	AR	MA	ARMA
ACF	Decay slowly	Drops to zero after lag q	Decay slowly
PACF	Drops to zero after lag p	Decay slowly	Decay slowly
IACF	Drops to zero after lag p	Decay slowly	Decay slowly

Table 3 - ARIMA model identification

It is worth noting that an ARIMA(0,d,1) model is equivalent to an ARIMA(p,d,0) model with the maximum number of feasible autoregressive terms, through the relationship given in (11).

$$
\begin{aligned}
X_t &= (1 - \theta_1 B)\,\varepsilon_t \\
\varepsilon_t &= (1 - \theta_1 B)^{-1} X_t \\
\varepsilon_t &= X_t + \theta_1 X_{t-1} + \theta_1^2 X_{t-2} + \ldots \\
X_t &= -\theta_1 X_{t-1} - \theta_1^2 X_{t-2} - \ldots - \varepsilon_t
\end{aligned}
\qquad (11)
$$

provided $|\theta_1| < 1$

Appendix 1c - Influence technique

This approach to outlier detection is based on the influence of an observation on the autocorrelation estimate, r_k, k=1, ... L which is used extensively in time series modelling and analysis. It can be argued that observations which have an undue influence on r_k should be identified as these may affect the success of the modelling process. Using the influence function matrix of Chernick et al (1982), Watson (1987) proposes a quantitative outlier detection statistic, IS_t, based on the influence of the t'th observation on r_k.

Firstly a matrix of influence values ($I_{t,k}$) is computed:

		LAG k =	1	2	3	4	5	.	.	.	L
OBS	t =	1	$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	$I_{1,4}$	$I_{1,5}$	.	.	.	$I_{1,L}$
		2	$I_{2,1}$	$I_{2,2}$	$I_{2,3}$	$I_{2,4}$	$I_{2,5}$	.	.	.	$I_{2,L}$
		3	$I_{3,1}$	$I_{3,2}$	$I_{3,3}$	$I_{3,4}$	$I_{3,5}$	.	.	.	$I_{3,L}$
		.	.	.	.	.	.				.
		.	.	.	.	.	.				.
		.	.	.	.	.	.				.
		N	$I_{N,1}$	$I_{N,2}$	$I_{N,3}$	$I_{N,4}$	$I_{N,5}$	.	.	.	$I_{N,L}$

$$
\text{where } I_{t,k} = Y_t Y_{t+k} - r_k \frac{(Y_t^2 - Y_{t+k}^2)}{2} \qquad (12)
$$

The influence statistic for the t'th observation is then defined by

$$
IS_t = \frac{1}{p} \left(\Sigma_{Lt} I^2 + \Sigma_{Dt-1} I^2 \right) \qquad (13)
$$

where $\Sigma_{Lt}I^2$ indicates summation of $I_{t,k}$ over the L elements in the t'th row and $\Sigma_{Dt-1}I^2$ indicates summation of $I_{t,k}$ over all available elements in the preceding (t-1)'th diagonal of the influence function matrix ($p = L_t + D_{t-1}$). In the derivation of theoretical moments and critical values for IS_t, ρ_k is assumed to be constant. In practice a global "summary sample estimate" is needed for ρ_k. The algorithm allows several different estimates to be used, although the following measure, r*, was found through empirical results to be appropriate:

$$
r^* = \frac{|\max r_k| + |\min r_k|}{2} \qquad (14)
$$

Where $|\max r_k|$ and $|\min r_k|$ are the absolute values of the maximum and minimum sample autocorrelation values respectively. Further manipulation of IS_t suggests a possible replacement or

estimation procedure for use with outlying or missing observations. The L replacement values are given by:

$$(z_t^k\ s) + y \quad \text{where } z_t^k = \frac{z_{t+k}}{r_k}\ (1 - \sqrt{1 - r_k^2}) \tag{15}$$

y and s being the mean and standard deviation of the original series, z_t representing the transformed data. Several replacement values are possible according to the k lags considered. Empirical and simulated results (Watson et al, 1992b) suggest that the replacement value associated with the largest r_k value is the most appropriate.