CONTROL CHARTS FOR NON-RESPONSE RATES IN THE CANADIAN LABOUR FORCE SURVEY

K.P. Hapuarachchi and A. Wroński, Statistics Canada
K.P. Hapuarachchi, 11-E R.-H. Coats, Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada

KEY WORDS: control chart, longitudinal survey, autocorrelated data

1. Introduction

The purpose of the study was to decide if some form of process control methodology could be applied to monitoring of survey non-response rates.

Since the introduction of control charts by Shewhart over 50 years ago, they have been widely used in industry for studies of process capability, measure capability studies, presentation of results of designed experiments, acceptance sampling and process control (Schilling and Nelson, 1976).

In the construction of control charts, two assumptions are generally made. They are: the measurements of the quality characteristic under consideration are normally distributed, and the measurements are independent. Under these assumptions, constants required to construct control charts for various sample sizes are tabulated in standard literature on quality control such as Burr (1976) and Duncan (1986). These tables are currently used by quality control practitioners to avoid unnecessary computations in computing control limits; thereby a decision about the quality of a product can be reached in a relatively short period of time.

Several papers for constructing control charts have been written for the case where the observations are not independent and the process generating these observations is stationary. The first paper by Vasilopoulis (1974) discusses a technique to construct control charts for characteristics using an autoregressive process of order one. Later, Vasilopoulis and Stamboulis (1978) presented modifications to standard control charts for serially correlated observations assuming the model to be an autoregressive process of order 2. Spurrier and Thombs (1987) have discussed a method of constructing control charts for observations with cyclical behaviour. A similar approach for periodic data was used by Beneke et al (1988) based on the periodogram approach. An excellent summary for dealing with statistical process control when the data are autocorrelated is given by Woodall and Faltin (1993).

In many large scale surveys, resources are allocated to control non-sampling errors leading to more precise estimates derived from these surveys. Non-response is a major cause of non-sampling error and the study of non-response is, therefore, important in controlling the total non-sampling error. Because of the large amount of historical data available, the Labour Force Survey non-response process was chosen for a quality assurance methodology research study.

2. Canadian Labour Force Survey (LFS)

The primary objective of the monthly LFS is to provide estimates of the number, characteristics and activities of the employed, unemployed and persons who are not in the labour force. The secondary objective is to serve as the general survey vehicle for the collection of a wide range of information on the Canadian population by supplementary surveys.

The Labour Force Survey currently has a sample size of approximately 60,000 households per month. The survey involves interviews with about 140,000 persons per month for all persons living in these households. Each household remains in the sample for six months, and one sixth of the households are replaced each month. The sampling fraction used varies by province and by type of area (e.g., urban, rural) within provinces.

The Labour Force Survey has traditionally achieved high response rates, generally in the mid ninety percent levels. Non-response occurs in surveys due to various reasons such as refusal of respondents to give information, their being not-at-home, households being inaccessible, sample units being unable to provide required information etc. A plot of non-response rates at the national level over all types of non-responses from January 1984 to January 1994 is given in Figure 1. The non-response rate for each month is calculated as the ratio of the number of non-responding households to the total number of households in the sample. Figure 1 shows stable non-response rates prior to the 1987/1988 period and a sharp increase in non-response in the 1987 and 1988 period. Once again, this is followed by stabilized non-response rates at higher levels than the ones prior to 1987/1988. One possible explanation for the increase in rates during the 1987/1988 period is the increase in supplementary survey workloads at that time. Another very prominent feature is the seasonal fluctuations that are observed in the non-response rates. Maximum non-response rates are observed in July and followed by a minimum in October in each year.

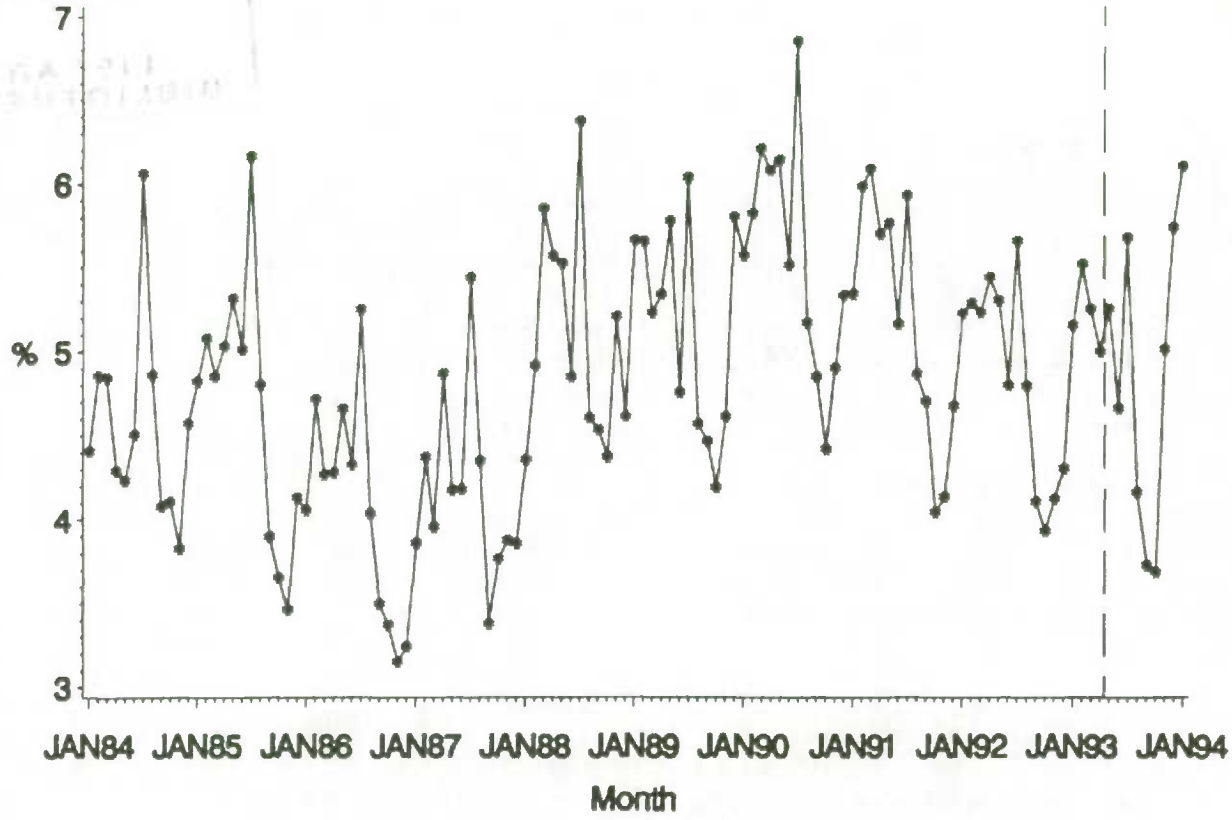

Figure 1. The non-response rate of Labour Force Survey.

3. Control Charts for Independent Observations

The general procedure in constructing any type of control chart is to first estimate the quality parameter of interest and its standard deviation, and then to set upper and lower control limits at three standard deviations from the estimate. This method is used in $\bar{x}$-charts. There are several methods proposed in the literature to estimate the standard deviation of the sample mean for these charts. One widely used method requires several random samples, each of size n. The process standard deviation is then estimated using the average range of these samples (Duncan 1986). This method is applicable when the independent samples have multiple observations.When only a single observation is taken per time period, control charts for individual observations have to be constructed. The 3-s.d.(s.d. standard deviation) control limits for an individual chart are given by $\bar{x}\pm 3s$ where $\bar{x}$ and s are the mean and the standard deviation, respectively, of the n individual observations.

Note that in the Labour Force Survey, a single sample per month is taken and as such this process does not provide data in subgroups as has been discussed in this section. Therefore, a control chart based on individual observations has to be used for non-response rates.

4. Control Charts for Serially Correlated Data

As has been indicated earlier, in the construction of control charts by variables the quality characteristic is assumed to be independently normally distributed. However, the number of non-respondents (i.e. the number of individuals who do not respond in a given month) for each time period follows a binomial distribution with parameters n and p, here n is the sample size and p can be defined as the overall average non-response rate or the probability that a randomly selected individual does not respond. Note that in this survey, the sample size is large enough (about 60,000 households per month) for estimates of the non-response rate to be approximately normally distributed (a direct consequence of the central limit theorem). However, the non-response rates for the Labour Force Survey tend to be serially correlated. One reason for this is that this survey uses a rotation panel design and in many cases once someone does not respond, this person may be unlikely to begin responding in a later month. Therefore, traditional Shewhart type control charts are not applicable for these rates because of the possible

serial correlation. Furthermore, the seasonal fluctuation and trend in the above rates should also be taken into consideration in the construction of such charts.

As stated in the introduction, there are several approaches to constructing control charts for autocorrelated data. However, these procedures are not applicable to Labour Force Survey data as non-response rate is the non-stationary (i.e. there is a seasonal fluctuation and trend). Therefore, appropriate control charts should be constructed by incorporating time series models that include the autocorrelation structure, seasonality and trend.

5. Time Series Model Selection

The ARIMA modelling procedure, a widely used technique in time series analysis, was used for selecting appropriate models for the non-response rate. ARIMA modelling is a type of univariate analysis of time series data. In ARIMA analysis, we suppose that the time-sequenced observations in a data series are statistically dependent. ARIMA models are especially suited for short-term forecasting because this procedure puts heavy emphasis on the recent past rather than the distant past. They are particularly useful for forecasting data series that contain seasonal (or other periodic) variation, including those with shifting seasonal patterns.
The general mathematical formula of ARIMA models is as follows:

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

$$= (1-\theta_1 B-\theta_2 B^2- \ldots -\theta_q B^q)e_t, \quad t=1,2, \ldots, n;$$

where $X_1, X_2, \ldots, X_n$ is the time series, B is the backshift operator such that $BX_t=X_{t-1}$, n is the number of observations in the series, μ is the overall mean of the series, θ's and ϕ's are model coefficients, p and q denote the orders of the ϕ and θ coefficients respectively, d is the order of differencing to make the series stationary, e_t is the random error component assumed to be independently, normally distributed with zero mean and constant variance σ^2 and t is the time index. The goal of fitting is to choose an ARIMA model (or choose the d, q, p, θ's, ϕ's and μ parameters) that includes the smallest number of non-zero parameters needed to adequately match the patterns of available data.

The Box-Jenkins approach was used to determine suitable time series models for the non-response rate. This procedure involves detailed examination of the model identification techniques such as the sample autocorrelation function, inverse autocorrelations, the sample partial autocorrelation function, differencing data if necessary to obtain stationarity, residual analysis and various other statistical procedures. The model selected should be as simple as possible. Note that for the ARIMA analysis to be reliable, a sufficiently large number of observations (generally greater than 50) are required. For detailed discussion see Box and Jenkins (1976) or Pankratz (1983). If the model is appropriate, the residuals generated from this process will be independently normally distributed. Thus standard Shewhart charts for individual observations can be constructed for the residuals (because the residuals are not serially correlated).

The SAS ARIMA procedure was used to fit several time series models for the non-response data from January 1984 to April 1993. The following time series model which best describes the non-response rates was selected using the Box-Jenkins approach:

$$(1-B)\,(1-B^{12})X_t = (1-\theta_1 B)\,(1-\theta_2 B^{12})e_t, \quad t=1,2, \ldots, n;$$

The above model is the multiplicative seasonal ARIMA model denoted by $(0,1,1)(0,1,1)_s$. Here s is the period of seasonality, and for the non-response, data $s=12$.

In choosing an appropriate time series model for the LFS non-response data, we considered the sudden increase of these rates during the 1987/1988 period. A transfer function component was included in the original model to account for this increase. However, it was found that the transfer function component was not significant and as such it was deleted from the model. A constant term was also included in the model to account for the possibility of trend in the non-response rates, but this term was also found to be not significant.

The maximum likelihood method was used to estimate the parameters of the model. An analysis of residuals (Actual-Forecast) was performed to investigate the assumption of normality. The Shapiro-Wilk test was used as a test of normality of residuals and the residuals were found to be normally distributed. Furthermore, this was confirmed by a normal probability plot of the residuals. A correlation check for residuals was performed and they were found to be independent (all residual autocorrelations were found to be within two standard deviation limits). The model presented in this section was found to be appropriate by applying the Portmanteau test (Q statistic) for overall goodness of fit. The maximum likelihood estimates of the parameters of the model are $\hat{\theta}_1=0.54$ and $\hat{\theta}_2=0.75$. The estimated model then becomes:

$$(1-B)(1-B^{12})X_t = (1-0.54B)(1-0.75B^{12})e_t, \quad t=1,2, \ldots, n.$$

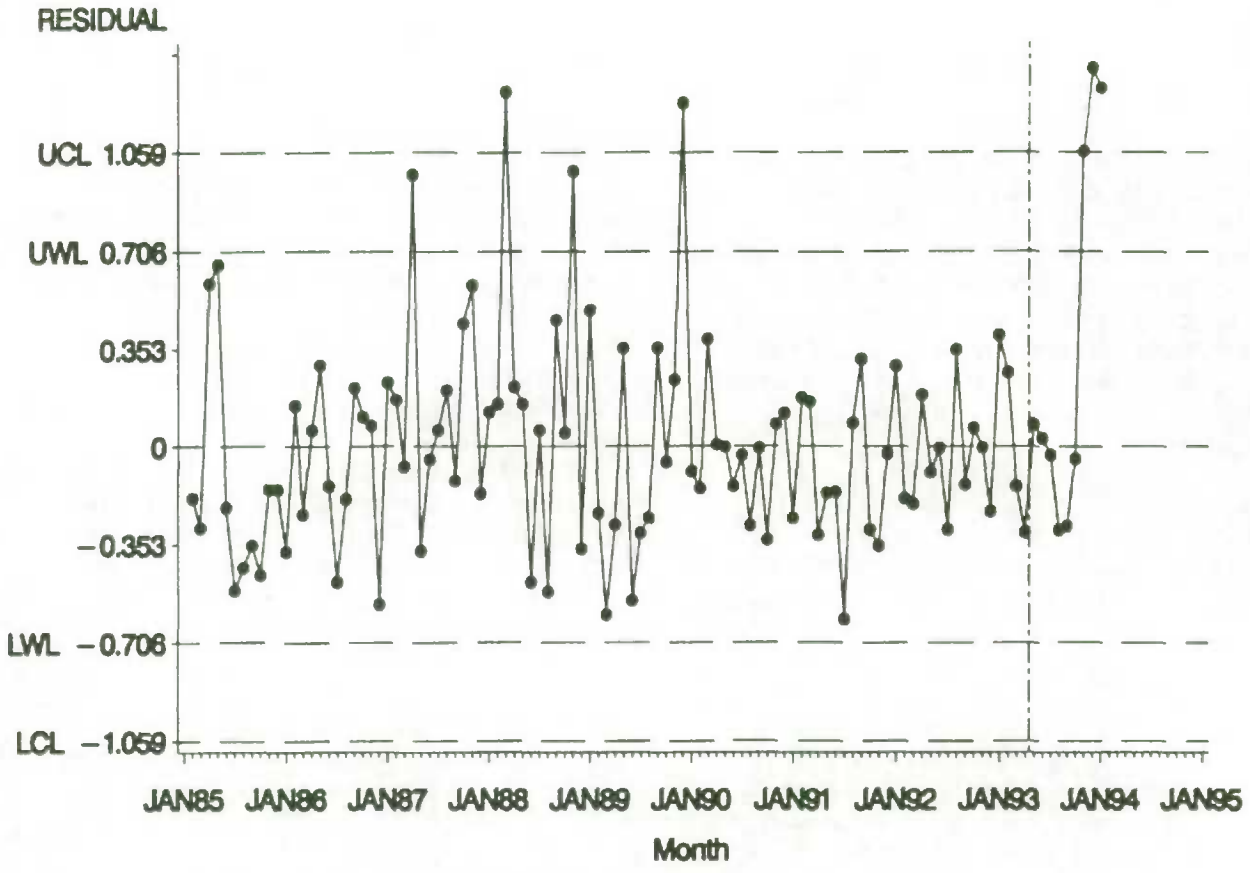

Figure 2. The Labour Force Survey non-response residuals control chart.

6. Construction of Control Charts for Non-response Data

6.1. Non-response Rates

As has been indicated earlier the usual $\bar{x}$-charts are not applicable to non-response data and charts for individual observations have to be used. Also the procedure used in this paper is to construct control charts for residuals instead of directly obtaining charts for non-response rates. The residuals generated from this process are independently normally distributed. Thus standard Shewhart charts for individual observations can be constructed for the residuals. If the residuals are in statistical control, then the model generating these residuals can be used as a good predictor equation of the non-response rates. If one or more residuals are large (positive or negative) or any non-random patterns are observed, then at those time points the model may not accurately predict the observed non-response rates. That is, some assignable cause(s) of variation may have occurred and steps should be initiated to identify these extraneous factors. This method of modelling the autocorrelative structure in the original data and applying control charts to the residuals was proposed by Montgomery and Mastrangelo (1991).

Let $r_1, r_2, \ldots, r_n$ be the residuals for the n time periods. Define $\bar{r} = \sum r_i/n$ and $s_r = \sqrt{\sum (r_i - \bar{r})^2/(n-1)}$ as the mean and the standard deviation of the residuals. Then using standard control theory, 3-s.d. control limits for residuals are given by the upper control limit = $\bar{r}+3s_r$ and the lower control limit = $\bar{r}-3s_r$. However, note that the random error is assumed to be independently normally distributed with zero mean and constant variance. Hence for the residual chart, the central line is taken to be 0. The sample standard deviation of the residuals is $s_r=0.662$. Therefore the upper and lower control limits for the residual control chart are given by

Upper control limit = 0 + 3(0.353) = 1.059

Lower control limit = 0 - 3(0.353) = -1.059.

Figure 2 shows the residual control chart for non-response rates for the period from February 1987 to January 1994. From this chart, it is clear that two points fall outside the upper control limit and this is an indication that at these time points non-response rates are out of statistical control. When such out of control situations do occur, actions should be initiated to look for possible causes for this behaviour.

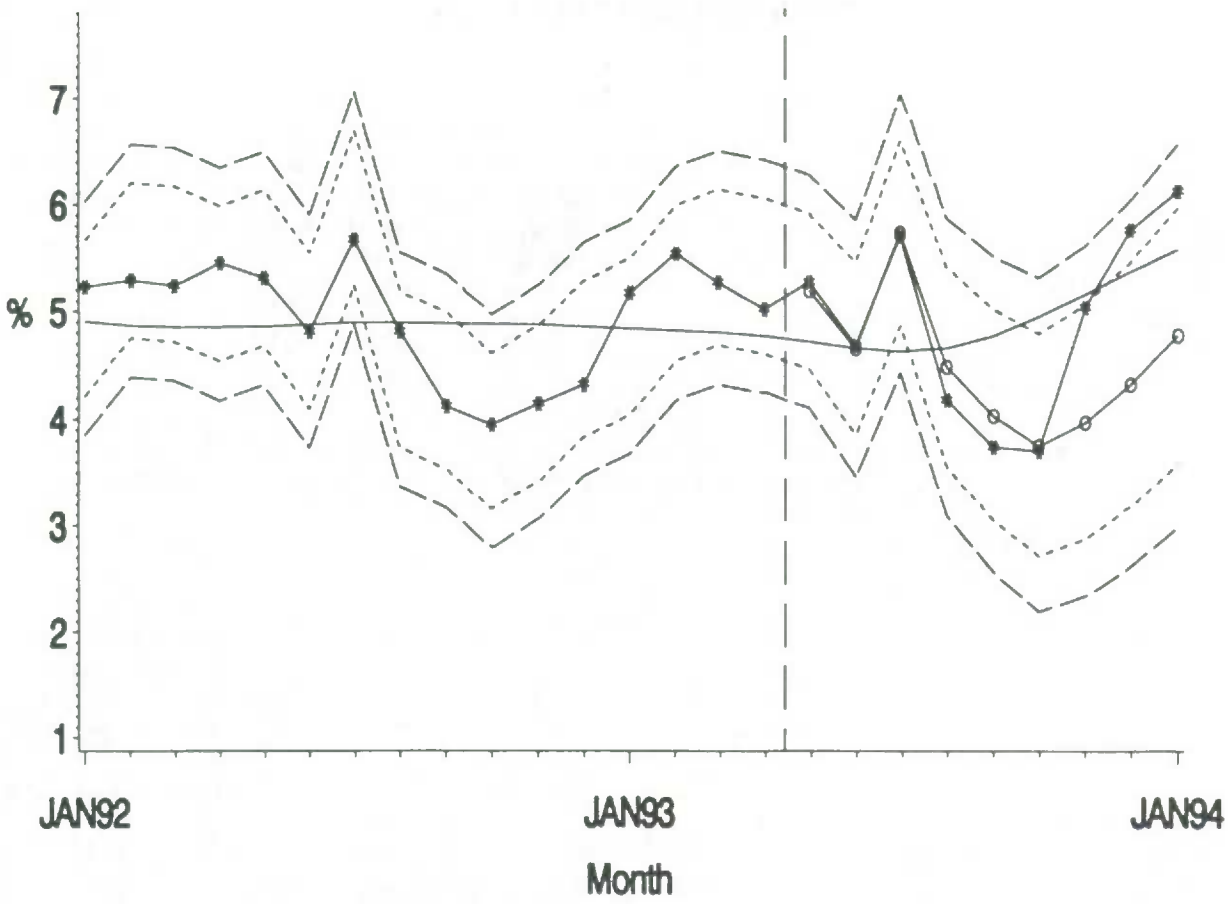

Figure 3. The Labour Force Survey non-response modified control chart.

6.2. Modified Control Chart

A modified control chart for non-response rates for the same period is presented in fig. 3 (from January 1992). The purpose of this exercise is to provide users with actual observations and their predicted values so that it is easy for them to understand and interpret these charts. Another advantage of this chart is that they can be constructed even if the residuals are not normally distributed (ARIMA calculates its prediction limits that are equal to control limits if residuals are normally distributed). The modified chart is obtained by adding forecasted or fitted values to the residuals. The solid line with stars represents the original data. The smooth solid line represents trend. Note that the trend estimates for the last two months have to be treated as preliminary estimates. The most interesting part of the graph is on the right of the broken vertical line dividing the original series and the forecast. The data after April 1993 have been plotted on the chart. Their position with respect to the forecast (circles) and the control and warning limits (broken lines) can be inspected. This chart shows forecasted values up to January 1994 and can be used to plot and analyze the non-response data until then. If the actual observation falls above or below the control limits it would be a warning sign. The analyst interested in the series should look for possible causes in the way the survey is conducted.

7. On-going Use of Residual Control Chart

In using the proposed procedure in this paper, it is assumed that the model structure and the parameters stay the same during the forecast period (i.e. for the future time periods). This implies that the forecast-generating process is in control. If this assumption is correct, the forecast errors are normally distributed with mean zero and constant variance such that both the mean and the variance remain constant over time. Therefore, approximately 95% of the forecast errors should fall within two standard deviation limits. The stability of the process (non-response rates) can be examined (as illustrated above) by (a) developing an appropriate time series model using the first several observations and (b) constructing a control chart based on these observations. This model can then be used to obtain forecasts and the residuals for the subsequent observations. These residuals can be plotted on the

control chart and if any of the points fall outside the control limits, action may be initiated to investigate the out of control situation. This might be due to a shift in the process, either in the parameters (may be due to some assignable cause) or a change in the actual structure of the model. In this case, the forecast-generating ARIMA model should be adjusted.

8. Revision of the model

As indicated earlier, if there is a change in the model, the distribution will shift; in particular, its mean may change. As a consequence, a large proportion of forecast errors will lie outside two standard deviation limits. Periodically (e.g., once a year) the adequacy of the model must, therefore, be examined using the Box-Jenkins approach. If another time series model seems to fit the new data, then this model should be incorporated in constructing control charts for residuals generated from non-response rates.

9. Conclusions

The control chart procedure described in this paper may be a useful technique for examining if the estimates such as the non-response rate or any other similar estimate in a series of longitudinal surveys is under statistical control. This is achieved by calculating the forecast errors and examining whether these errors fall within pre-specified control limits. As has been indicated before, the non-response rate derived from such longitudinal surveys which use panels are correlated and as such, standard control charts are not applicable because of possible serial correlation. Therefore, to construct control charts for serially correlated data the following procedure has been proposed in this paper:

(a) Identify the appropriate time series model that best describes the non-response rates.

(b) Estimate the parameters of the model in (a).

(c) Incorporate the model in (a) to construct an appropriate control chart for residuals generated from this model.

(d) As observations for new time points are available, calculate the residuals and plot them in the residuals control chart. If any of the residuals fall outside the control limits, examine why an out of control situation has occurred.

10. Further Study

Every month, Statistics Canada conducts a number of longitudinal surveys, e.g., Labour Force Survey (LFS), Survey of Employment, Payroll and Hours (SEPH), Monthly Wholesale and Retail Survey (MWRT). Various parameters from these surveys are computed and published. Because of the large amount of historical data available from these surveys, they can be used to investigate the possibility of applying the control chart methodology discussed in this paper. We have already begun analyzing other non-response rates (e.g., for new entrants) and the refusal rates, coefficients of variation, slippage rates, turnover rates and vacancy rates at the national level for the Labour Force Survey. We intend to report the findings of this analysis in a future paper.

11. References

1. Beneke, M., Leemis, L.M., Schlegel, R.E. and Foote, B.L. (1988). Spectral Analysis in Quality Control: A Control Chart Based on the Periodogram, Technometrics, Vol. 30, No. 1, pp. 63-70.
2. Burr, I.W. (1976). Statistical Quality Control Methods, Marcel Decker, New York.
3. Box G.E.P. and Jenkins G.M. (1976). Time Series Analysis: Forecasting and Control, Holden-Day.
4. Duncan, A.J. (1986). Quality Control and Industrial Statistics, Fifth edition, Irwin, San Francisco.
5. Montgomery, D.C. and Mastrangelo, C.M. (1991). Some statistical Process Control Methods for Autocorrelated Data, Journal of Quality Technology, Vol. 23, No. 3, PP.179-204.
6. Pankratz A. (1983). Forecasting with Univariate Box-Jenkins Models: Concepts and Cases, John Wiley New York.
7. Schilling, E.G. and Nelson, P.R. (1976). The Effect of Non-normality on the Theoretical Limits of $\bar{x}$-Charts, Journal of Quality Technology, Vol. 8, pp. 183-188.
8. Spurrier, J.D. and Thombs, L.A. (1987). Control Charts for Detecting Cyclical Behaviour, Technometrics, Vol. 29, No. 2, pp. 1630171.
9. Vasilopoulis, A.V. (1974). Second Order Autoregressive Model Applied to Quality Control, Ph.D. Dissertation, New York University.
10. Vasilopoulis, A.V. and Stamboulis, A.P. (1978). Modification of Control Limits in the Presence of Serial Correlation, Journal of Quality Technology, Vol. 10, No. 1, pp. 20-30.
11. Woodall, W.H. and Faltin, F. W. (1993). Autocorrelated Data and SPC, ASQC Statistics Division Newsletter, Vol. 13, No. 4, pp. 18-21.