Anomaly Detection for Analysis of Annual Inventory Data: A Quality Control Approach

Francis A. Roesch and Paul C. Van Deusen

ABSTRACT

Annual forest inventories present special challenges and opportunities for those analyzing the data arising from them. Here, we address one question currently being asked by analysts of the US Forest Service's Forest Inventory and Analysis Program's quickly accumulating annual inventory data. The question is simple but profound: When combining the next year's data for a particular variable with data from previous years, how does one know whether the same model as used in the past for this purpose continues to be applicable? Of the myriad approaches that have been developed for changepoint detection and anomaly detection, this report focuses on a simple quality-control approach known as a control chart that will allow analysts of annual forest inventory data to determine when a departure from a past trend is likely to have occurred.

Keywords: sampling, control charts

Annual forest inventory designs, such as the one initiated by the US Forest Service's Forest Inventory and Analysis (FIA) program, present special challenges and opportunities. One of those challenges can be thought of in the context of managing user expectations. Given that new data are acquired each year, users naturally expect new information to be available each year. This expectation must be balanced with the realization that many variables of interest are sampled too sparsely on an annual basis to be independently estimated over small areas using data gathered exclusively from that small area during a single year. Therefore, most annually updated estimates must rely on a modeled relationship between the measurements of successive years. This requires that the analyst applies a reasonable model in the first place and that the analyst constantly monitors the data to ensure that some departure from the model has not occurred. If a departure from the model is indicated, then a more robust model that can account for the departure should be developed.

Any model (m) of the trend in a particular variable can produce an expected value for the next year's observation ($E_m[x_t]$). Occasionally, the next year's observation (x_t) will be quite different from $E_m[x_t]$. This anomaly might be due to natural sample variation in the variable or it might be due to an actual departure from the previous trend. Naturally, the early detection of a potential change in a trend would be greatly beneficial to the analysis of data from these annual inventories. Many techniques have been proposed to address similar problems, such as those designed to judge whether two samples were drawn from the same population (i.e., the two-sample t-test or the rank-sum test) and those designed to detect a change point in a trend (i.e., Lai 1995). Here we explore a useful and simple technique developed in statistical process control, known as the control chart (Shewhart 1931, Deming 1964).

In manufacturing, a graph known as a control chart is used to distinguish between inherent process variation and variation that indicates a change in the process. In its simplest form, the control chart is a graph of successive sample means or sample ranges. The chart consists of three lines, a center line, an upper control line, and a lower control line. If one of the plotted values falls outside one of the control lines, the process is judged to be out of control. In the case at hand, if the next plotted value (of difference from the modeled prediction) falls outside the control lines, we would suspect that there might have been a shift in the trend of the variable of interest. The inventory analyst would do further diagnostics to determine whether a shift may have actually occurred and what might have been the cause of the shift.

Establishment of the three control lines can be quite simple. The center control line could represent the average of k sample means for an in-control process ($\bar{y}_k$). The upper control line is then computed as $\text{UCL} = \bar{y}_k + c\sigma/\sqrt{n}$, whereas the lower control line is computed as $\text{LCL} = \bar{y}_k + c\sigma/\sqrt{n}$. The SD($\sigma$) can be estimated by using the square root of a pooled sample variance. We set c according to the relative costs of committing a type I versus a type II error. Recall that a type I error will be committed if we reject the null hypothesis when it is true, and a type II error is committed if we fail to reject the null hypothesis when it is false. For instance, the use of $c = 3$ would yield a very small chance of committing a type I error but a larger chance of committing the type II error of failing to recognize a true change in the trend. In terms of hypotheses testing, if we wish to test for a general change in the trend, we might assume the following null and alternative hypotheses:

H_0: There is no change in trend

H_1: There has been a change in trend

In general, any control chart can be devised through the formula $E(S) \pm Z_{\alpha/2}\sigma_S$, where S is a test statistic and σ_S is the SD of the test statistic (Chandra 2001). Rather than using an overall sample mean, we take the approach in the example below of plotting the detrended value $y_t = x_t - E_m[x_t]$ to compare against a horizontal line equal to

Manuscript received May 15, 2009, accepted March 3, 2010.

Francis A. Roesch (froesch@fs.fed.us), US Forest Service, Southern Research Station, 200 WT Weaver Boulevard, Asheville, NC 28804-3454. Paul C. Van Deusen, National Council for Air and Stream Improvement, Statistics and Model Development Group, 600 Suffolk St., Lowell, MA 01854.

Table 1. Ecological classifications used in this study as combined from those described in McNab et al. (2005).

EcoClass	ECOSUBCD
1. Southern Appalachian Piedmont	231Aa, 231Ab, 231Ac, 231Ad, 231Ae, 231Af, 231Ag, 231Ah, 231Ai, 231Aj
2. Coastal Plains, Middle Section	231Ba, 231Bb, 231Bc, 231Bd, 231Be, 231Bf, 231Bj
3. Southern Cumberland Plateau Section and the Central Interior Broadleaf Forest Province	223Ee, 223Ef, 223Eg, 231Ca, 231Cb, 231Cc, 231Cd, 231Ce, 231Cf, 231Cg
4. Southern Ridge and Valley and the Blue Ridge Mountains	231Da, 231Db, 231Dc, 231Dd, 231De, M221Dc, M221Dd
5. Gulf Coastal Plains, Lowlands and Flatwoods Sections	232Bg, 232Bh, 232Bi, 232Bj, 232Bk, 232Bl, 232Bm, 232Bn, 232Bo, 232Bp, 232Bs, 232La, 232Lb
6. Southern Atlantic Coastal Plains and Flatwoods Sections	232Ca, 232Cb, 232Cc, 232Cd, 232Ce, 232Cf, 232Cg, 232Ja, 232Jd, 232Je, 232Jf, 232Jg

0 (assuming an unbiased model), as exploited thoroughly by Deming (1964).

The remainder of this article is arranged as follows. First, we describe an annual forest inventory sample design and the construction of a control chart for this design. Next, we present an example of the use of the chart for consideration of the next measurement, and finally we give some recommendations for analysts of data arising from similar designs.

An Annual Forest Inventory Design

For illustrative purposes, we will use the US Forest Service's FIA units' temporally rotating, panelized forest inventory sampling design as an example. Here we give a brief explanation of the design, which is well documented in Reams et al. (2005). In this design, the sample plots are located in proximity to a systematic triangular grid consisting of g mutually exclusive interpenetrating panels. The panels are spatially balanced and contain an approximately equal number of sample plots. That is, if the total sample size is n, then each panel consists of approximately n/g plots. The sequence of panels is measured in order, with one panel measured each year, after which the panel measurement sequence reinitiates. Therefore, if panel 1 was measured in 2001, it will also be measured in 2001 + g, 2001 + 2g, and so on. Panel 2 would then be measured in 2002, 2002 + g, 2002 + 2g, etc. Scott et al. (2005) describe the estimation methods that FIA uses for within-panel estimates for this design. Patterson and Reams (2005) gives an introduction to methods of combining FIA panels for the establishment of time series. Since the initiation of the rotating panel design for FIA, there have been quite a few papers focusing on using trend models for the purpose of improving annual estimates, such as Van Deusen (1996, 1999), Roesch et al. (2003), Roesch (2007), Johnson et al. (2003), and Czaplewski and Thompson (2009).

Without loss of generality, we will restrict this discussion to the case of five annual panels ($g = 5$) sampling a defined area through time. One panel will be measured each year, and starting with year 6, 20% of the plots (i.e., one panel) will be remeasured each year. Roesch (2008) gives a useful explanation of this sample frame for the interested reader.

Estimates of In-Control Variance

Assume that a process that has not been determined to be out of control prior to time t is in control at time t. We can then estimate the in-control process variance using the g prior panels measured from time $t - g$ through time $t - 1$:

$$\hat{\sigma}_0^2 = \frac{1}{g}\sum_{i=t-g}^{t-1}\frac{1}{(n_i-1)}\sum_{j=1}^{n_i}(x_{i,j}-\bar{x}_i)^2, \qquad (1)$$

where $\bar{x}_i$ is the sample mean at time i.

Example Data

The data were collected in Alabama and Georgia from 1997 through 2007. These data are available to the public online (US Forest Service 2010). In this study, we will focus our interest on estimates of basal area per acre over six broad ecological classifications for plantations of three size classes and natural stands. The ecological classifications were obtained by pooling categories within the FIA variable ECOSUBCD, described in McNab et al. (2005). Table 1 lists the ECOSUBCD classifications as they were pooled into one of six groups (EcoClass) for this study.

We partitioned the data from stocked plantations using the FIA stand-size class code, which is a classification of the stand based on the diameter distribution (US Forest Service 2008). Stand size class 1 has at least 25% of the stocking in sawlog sized trees (dbh of at least 11.0 in. for hardwoods and at least 9.0 in. for softwoods) and 50% of the stocking in trees that are at least 5.0 in. diameter. Stand size class 2 has less than 25% of the stocking in sawlog sized trees with 50% of the stocking in trees that are at least 5.0 in. diameter. Stand size class 3 consists of stands with at least 50% of the stocking in trees with diameters less than 5.0 in.

Figure 1 gives the proportion of land in each stand classification throughout the period of interest for each EcoClass. In a production system, a figure like Figure 1 could give a quick visual indication of whether a particular population partition is large enough to support an adequate sample. Population partitions that do not appear to be large enough could be pooled with other population partitions. Alternatively, the small partition could remain autonomous and have no control chart produced, or it could have a control chart produced but ignored.

Because our interest is in basal area per acre by stand origin and size and more than one of these conditions may exist on a plot, we consider the value of the basal area per acre estimate for each condition to be proportional to the area of the condition on the plot. This suggests the use of a weighted mean for our estimate of basal area per acre in each category. That is, for the basal area per acre estimate for condition k at time i, $\bar{x}_{k,i} = \sum_{j=1}^{n_i} w_{k,i}\, x_{k,j}$, where the $w_{k,j}$ are the normalized weights such that $\sum_{j=1}^{n_i} w_{k,j} = 1$.

In the sequel, we will drop the subscript for condition. In this case, for a weighted mean, we require an adjustment to the process variance estimator (Equation 1):

$$\hat{\sigma}_{0w}^2 = \frac{1}{g}\sum_{i=t-g}^{t-1}\frac{1}{\left(1-\sum_{j=1}^{n_i} w_j^2\right)}\sum_{j=1}^{n_i} w_j (x_{i,j}-\bar{x}_i)^2, \qquad (2)$$

Note that Equation 2 reduces to Equation 1 when the weights are equal (i.e., $w_j = 1/n_i$).

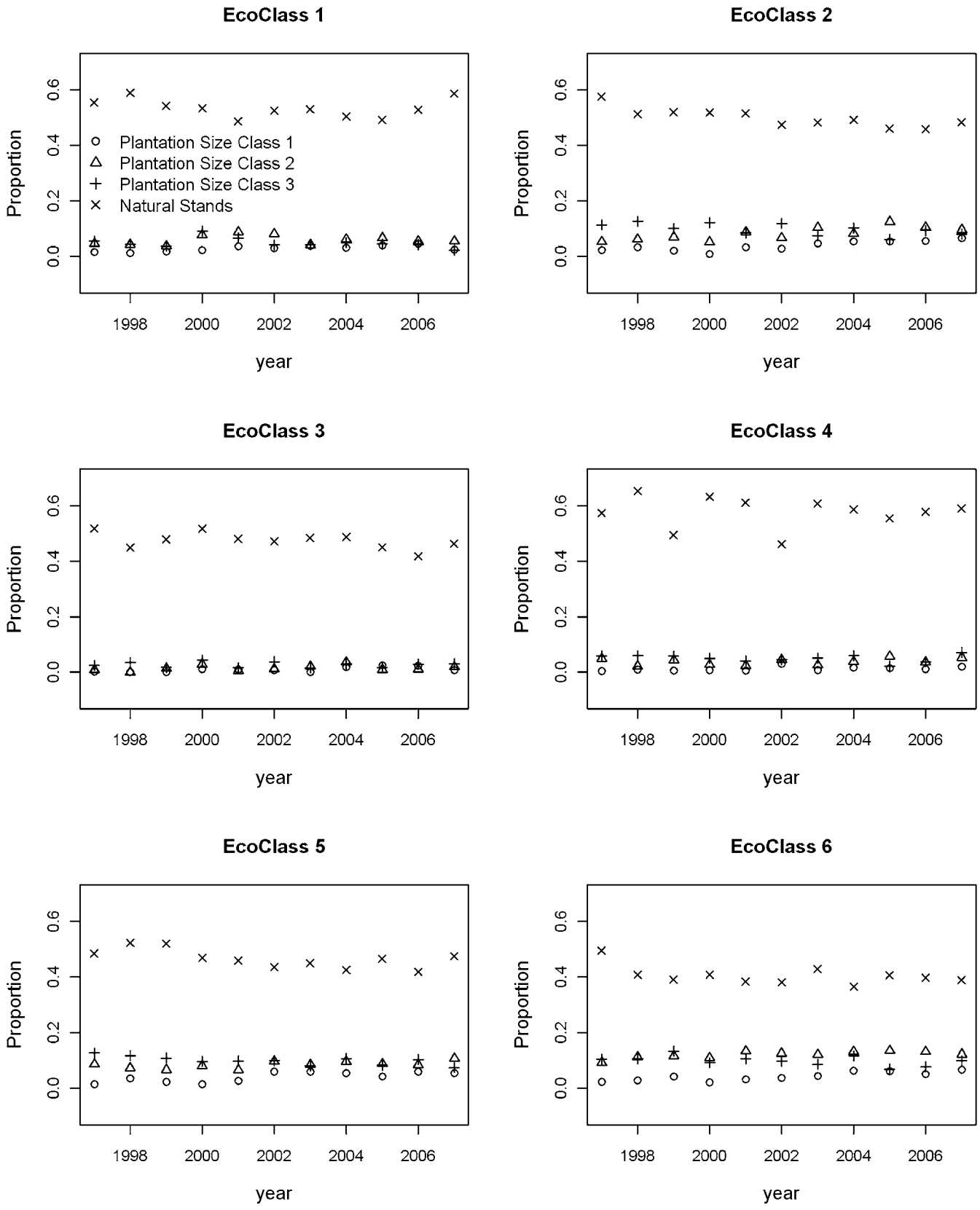

Figure 1. The proportion of land in each stand classification throughout the period of interest for each EcoClass.

Table 2. A synopsis of the control charts (with $c = 2$) as a 3-tuple of four potential outcomes for models M1, M2, and M3, respectively, for each stand classification and EcoClass combination.

	EcoClass					
Stand classification	1	2	3	4	5	6
Plantations						
Size class 1	IIi	III	OOO	iii	III	III
Size class 2	ooo	Iii	iii	OOO	III	ooO
Size class 3	iiO	IIO	iOO	III	ooo	III
Natural stands	iiO	iio	iii	iOO	iii	Iii

The potential outcomes were as follows: the process was in control for the entire observation period (I), the process was out of control for the entire observation period (O), the process came into control during the observation period (i), and the process went out of control during the observation period (o).

Table 3. A synopsis of the control charts (with $c = 3$) as a 3-tuple of four potential outcomes for models M1, M2, and M3, respectively, for each stand classification and EcoClass combination.

	EcoClass					
Stand classification	1	2	3	4	5	6
Plantations						
Size class 1	III	III	OOO	III	III	III
Size class 2	III	Iii	iii	IIO	III	IIi
Size class 3	iiO	IIo	iOO	III	III	III
Natural stands	Iii	III	III	ioo	III	III

The potential outcomes were as follows: the process was in control for the entire observation period (I), the process was out of control for the entire observation period (O), the process came into control during the observation period (i), and the process went out of control during the observation period (o).

Suppose we consider three potential models for the trend in the population mean:

M1: $\bar{X}_t = \bar{X}_{t-1}$,

M2: $\bar{X}_t = b_0 + b_1\bar{X}_{t-1}$, and

M3: $\bar{X}_t = b_0 + b_1\bar{X}_{t-1} + b_2\bar{X}_{t-2}$.

The expected value (or prediction) of $\bar{x}_{t+1}$ ($E[\bar{x}_{t+1}]$) under M1 would simply be $\bar{x}_t$, whereas the respective regression estimators would be used under M2 and M3, with the parameters being estimated from the data obtained up through the prior year.

To construct the control charts, we set $y = \bar{x}_{t+1} - E(\bar{x}_{t+1})$, under each of the three models. The mean line is drawn at $y = 0$, and the control lines are established as described above. We produced charts using a setting of $c = 2$ and $c = 3$. Setting $c = 2$ will result in a control chart that would be relatively sensitive to the commission of a type II error, as opposed to the higher setting $c = 3$, which would be more sensitive to the commission of a type I error. In practical terms, the higher setting for c will signal the possibility of a change in the trend less often, as a higher threshold has to be crossed. Fewer false alarms will have to be responded to, but there is a higher probability that a true change in the trend will be missed. In addition, the three models have different memory lengths. Therefore, we would expect to see lagged reactions between models, within the control chart, due to a true change in a variable's trend. Our purpose for introducing three different models is not to use the control chart to choose the best model but rather to show how control charts differ in sensitivity to models of varying memory length and how these differences might be used as an interpretation mechanism. We compare the models beginning in the year 2004 through the year 2007, under the assumption that there would be enough supporting data for the models in most of the conditions by that time.

Results and Discussion

Tables 2 and 3 give a synopsis of the resulting 144 control charts into one of four outcomes:

1. The process was in control for the entire observation period (I)
2. The process was out of control for the entire observation period (O)
3. The process came into control during the observation period (i)
4. The process went out of control during the observation period (o).

Note that the fourth outcome (o) is the one that would trigger an alarm. The alarm would signal a possible change in trend or the possible rejection of the assumed trend model when combining the next year's data, for a particular variable, with data from previous years. In the two tables, the outcomes are given in a 3-tuple for M1, M2, and M3, respectively, for each stand classification and EcoClass combination. That is the 3-tuple Iii indicates that outcome I occurred for M1, whereas outcome i occurred for M2 and M3. Table 2 gives the results when c was set equal to 2, and Table 3 gives the results when c was set equal to 3.

We had 144 potential charts (four stand classes in six EcoClasses for three models and two settings of c). We grouped the models in a chart for each stand class and EcoClass. Of the charts resulting from a setting of $c = 2$, 24 were in control for the entire observation period of 2004–2007, 14 were not in control for the observation period, 25 started out of control but came into control before the end of the observation period, and the rest went out of control during the observation period.

In comparing Table 2 with Table 3, note the outcome o occurred nine times in Table 2, when c was set equal to 2, and only three times in Table 3, when c was set equal to 3. Setting $c = 3$ is a very common recommendation in the process control field because there is usually a high cost associated with a triggered alarm, such as shutting down an entire production line. In the case at hand, we are more concerned with making the invalid assumption that the next year's data are compatible with previous years' data in the same way (or under the same model) as has been assumed true in the past. This concern is better addressed by setting c lower, as we have done for the results summarized in Table 2.

For brevity, we present only the control charts for the setting of $c = 2$ in EcoClasses 5 and 6, both on the coastal plain with a mix of natural stands and plantations. We see in Figure 2 for EcoClass 5 that the three models give very similar results in all three size classes in the plantations. In the charts for plantation size class 3 and natural stands, plotted values do occur slightly outside the control zone. Note, however, that the three models show coincident results in the out-of-control region in both cases. In lieu of any prior or subsequent differentiation of the model results, an undifferentiated out-of-bounds result would probably be due to sampling variation as opposed to a true change in trend. However, because there was, in both of these cases, some prior within-bounds differentiation between the models, a deeper investigation into a potential change in trend is warranted.

Figure 3 plots the results for EcoClass 6. Note that these results for plantation size class 2 give a much clearer indication of a change in trend than the more subtle results in Figure 2. In this figure, the longest memory model (M3) is clearly differentiated and out of bounds.

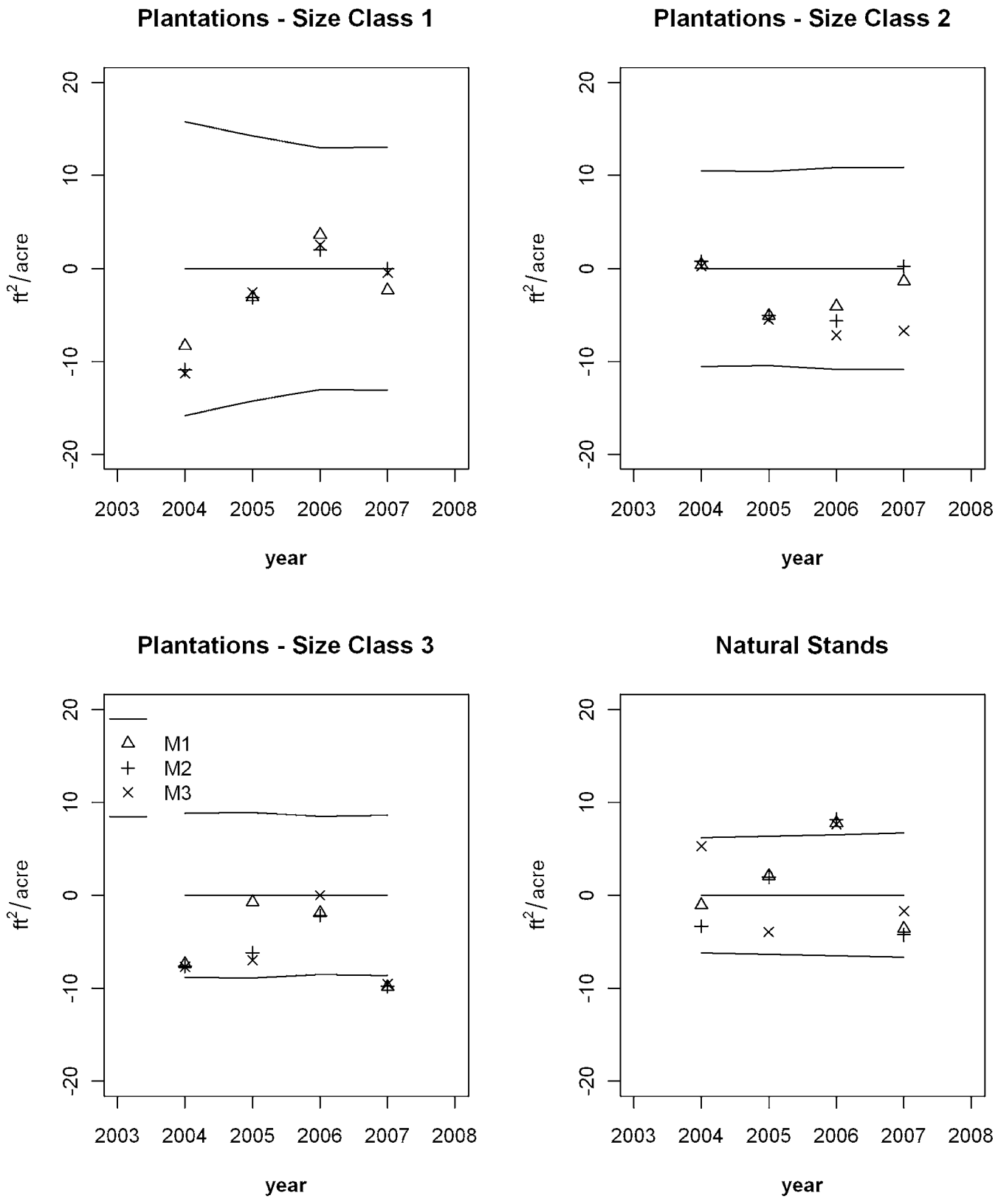

Figure 2. The control chart for the detrended basal area estimates for the three models (M1, M2, and M3) in each of the stand classifications for EcoClass 5. Values falling above the top line (the upper control limit) or below the bottom line (the lower control limit) are out of control.

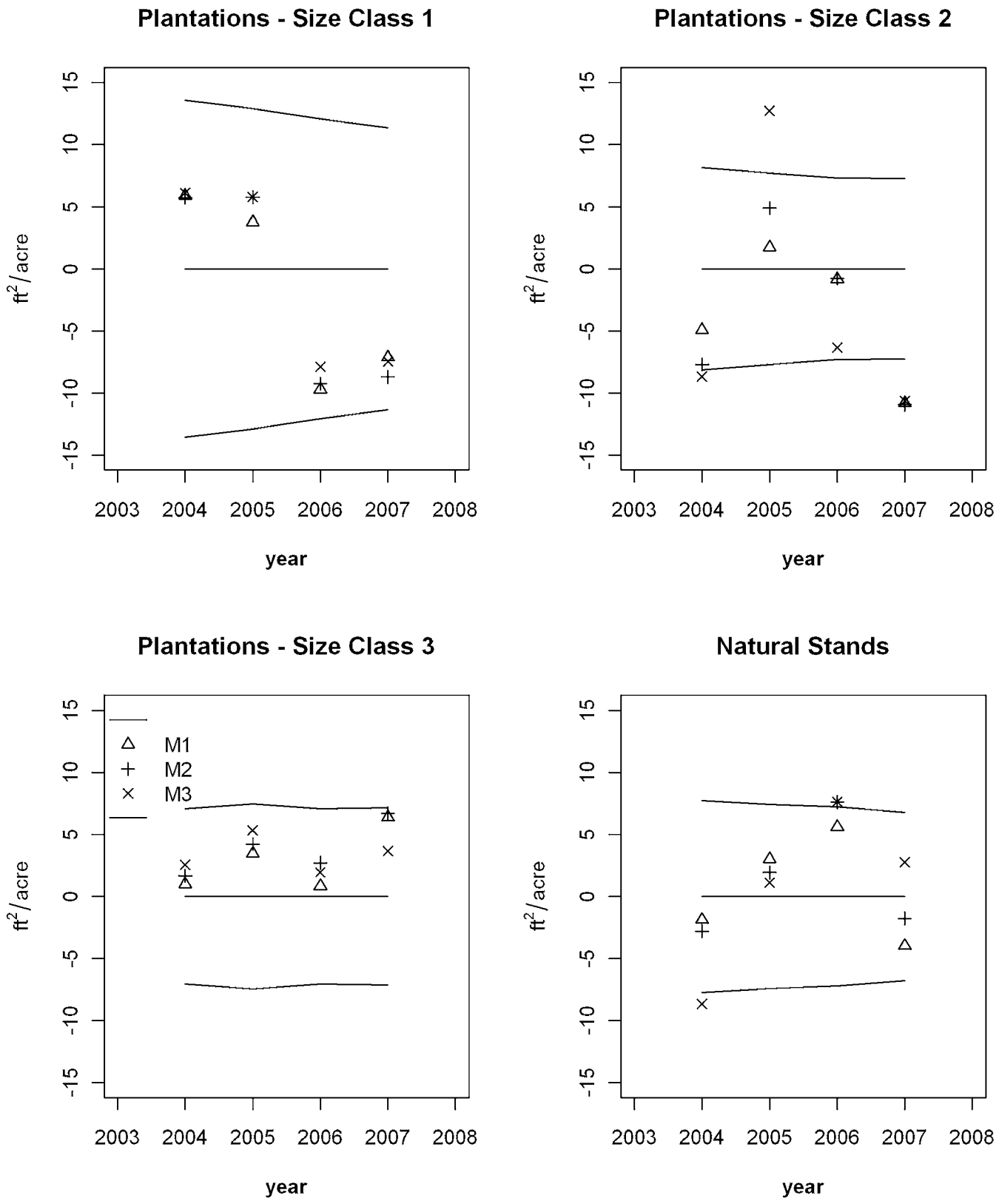

Figure 3. The control chart for the detrended basal area estimates for the three models (M1, M2, and M3) in each of the stand classifications for EcoClass 6. Values falling above the top line (the upper control limit) or below the bottom line (the lower control limit) are out of control.

Conclusion

The control chart is a simple and useful way for practitioners to determine whether a change in a past trend is likely to have occurred. As in most time series approaches, decisions will be made with more confidence as the length of annual observations increases. As with the advantage of increasing sample size, this effect is asymptotic in nature. The control chart approach for detecting changes in trend is only helpful for well-observed trends, i.e., those that are expected to be in control most of the time. When a chart is never in control, it is usually assumed to indicate that not enough data are available. In this example, this can be the result of either high within-condition variance, especially in the natural stands, or a small amount of a condition within an EcoClass.

The interested reader can reproduce all of the charts by downloading the data from the website given above and following the procedure that we describe. Charts that stay in control through the entire observation period show no evidence of a change in trend and do not need to be considered further. Charts that are out of control for the entire observation period result from conditions for which there are too few observations and should therefore be ignored. This is in keeping with the historical use of control charts: the charts are not considered valid or useful until the process is in control (that is, we have enough information to begin monitoring the chart). Once the process is in control, the charts provide an occasional momentary distraction for the engineer (or scientist) but are otherwise ignored until the process goes out of control. Analogously, in annual forest inventories, an analyst would keep using a particular estimator for combining the next year's data with the previous years' data as long as the control chart based on the estimator's underlying assumptions does not suggest that a change in those assumptions is appropriate. Once the chart goes out of control, the inventory analyst, much like the engineer who would shut down the production line until the problem is resolved, would reconsider the optimal use of the next year's data and develop a model that is appropriate for the new conditions.

Literature Cited

CHANDRA, M.J. 2001. *Statistical Quality Control.* CRC Press. New York. 284 p.

CZAPLEWSKI, R.L., AND M.T. THOMPSON. 2009. Opportunities to improve monitoring of temporal trends with FIA panel data. In *Forest Inventory and Analysis (FIA) Symposium 2008; October 21–23, 2008; Park City, UT. Proc. RMRS-P-56CD*, McWilliams, W.H., G.G. Moisen, and R. L. Czaplewski (comps.). US For. Serv., Rocky Mountain Res. Stn., Fort Collins, CO. 1 CD.

DEMING, W.E. 1964. *Statistical adjustment of data*. Dover Publications Inc., New York. x + 261 pp. (Corrected republication of the 1943 John Wiley and Sons printing.)

JOHNSON, D.S., M.S. WILLIAMS, AND R.L. CZAPLEWSKI. 2003. Comparison of estimators for rolling samples using forest inventory and analysis data. *For. Sci*. 49(1):50–63.

LAI, T.L. 1995. Sequential changepoint detection in quality control and dynamical systems. *J. R. Stat. Soc. Ser. B (Method.)* 57(4):613–658.

MCNAB, H.W., D.T. CLELAND, J.A. FREEOUF, J.E. KEYS, JR., G.J. NOWACKI, AND C.A. CARPENTER. 2005. *Description of "Ecological subregions: Sections of the conterminous United States" first approximation*. US For. Serv., Ecosystem Management Coordination. Washington, D.C.

PATTERSON, P.L., AND G.A. REAMS. 2005. Combining panels for Forest Inventory and Analysis estimation. P. 69–74 in *The enhanced Forest Inventory and Analysis program: National sampling design and estimation procedures*, Chap. 5. Bechtold, W.A., and P.L. Patterson (eds.). US For. Serv. Gen. Tech. Rep. SRS-80. Asheville, NC.

REAMS, G.A., W.D. SMITH, M.H. HANSEN, W.A. BECHTOLD, F.A. ROESCH, AND G.G. MOISEN. 2005. The Forest Inventory and Analysis sampling frame. P. 11–26 in *The enhanced Forest Inventory and Analysis program: National sampling design and estimation procedures*, Chap. 2. Bechtold, W.A., and P.L. Patterson (eds.). US For. Serv. Gen. Tech. Rep. SRS-80. Asheville, NC.

ROESCH, F.A. 2007. Compatible estimators of the components of change for a rotating panel forest inventory design. *For. Sci.* 53(1):50–61.

ROESCH, F.A. 2008. An alternative view of continuous forest inventories. *For. Sci.* 54(4):455–464.

ROESCH, F.A., J.R. STEINMAN, AND M.T. THOMPSON. 2003. Annual forest inventory estimates based on the moving average. P. 21–30 in *Proc. of the third annual Forest Inventory and Analysis symposium; 2001 October 17–19; Traverse City, Michigan*, McRoberts, R.E., G.A. Reams, P.C. Van Deusen, and J.W. Moser (eds.). US For. Serv. Gen. Tech. Rep. NC-230. US For. Serv., North Central Res. Stn., St. Paul, MN. 208 p.

SCOTT, C.T., W.A. BECHTOLD, G.A. REAMS, W.D. SMITH, J.A. WESTFALL, M.H. HANSEN, AND G.G. MOISEN. 2005. Sample-based estimators used by the Forest Inventory and Analysis National Information Management System. P. 43–67 in *The enhanced Forest Inventory and Analysis program: National sampling design and estimation procedures*, Chap. 4. Bechtold, W.A., and P.L. Patterson (eds.). US For. Serv. Gen. Tech. Rep. SRS-80. Asheville, NC.

SHEWHART, W.A. 1931. *Economic control of quality of manufactured product*. Van Nostrand Reinhold, Princeton, NJ. 501 + xiv p.

US FOREST SERVICE. 2008. The Forest Inventory and Analysis database: Database description and users manual version 3.0 for phase 2. Forest Inventory and Analysis Program, US For. Serv., Washington, DC.

US FOREST SERVICE. 2010. *FIA DataMart.* Available online at 199.128.173.17/fiadb4-downloads/datamart.html; last accessed May 24, 2010.

VAN DEUSEN, P.C. 1996. Incorporating predictions into an annual forest inventory. *Can. J. For. Res.* 26:1709–1713.

VAN DEUSEN, P.C. 1999. Modeling trends with annual survey data. *Can. J. For. Res.* 29(12):1824–1828.