A Method for Early Identification of Loss from a Nuclear Material Inventory

DAVID E. BOOTH

Department of Administrative Sciences, Graduate School of Management, Kent State University, Kent, Ohio 44242

Received April 17, 1984

This paper presents a successful method, using generalized M estimates for AR(1) time series models, for the early detection of loss from any inventory for which the material accounting procedure is based on a material balance. In particular, this method is applicable to an inventory containing nuclear material. The method is also shown to be successful in detecting outlying points in data sets for which an AR(1) time series model is appropriate. The advantages of this procedure and its generalizations over others currently available are discussed.

INTRODUCTION

It is well-known that outliers (nonrepresentative sample points) can have a deleterious influence on the computed coefficients in any statistical procedure based on ordinary least squares.[3,19] In particular, this is true for the case of time series estimation involving ARMA models.[23] Procedures, called robust, attempt to counteract this influence by modifying the estimation procedure in such a manner that the outliers' influence on the final result is decreased. During the last few years these robust procedures have been applied to many cases of practical importance, with considerable success (see, for example, footnotes 5, 6, 9, 10, 19, and 22).

This paper gives a brief description of a modification of a procedure, first developed by Denby and Martin,[14] called the Generalized M-estimator (GM) procedure, for stationary first-order autoregressive [AR(1)] time series.[12] It then shows that this modified procedure can be effectively used to develop a model that identifies outlying points in an AR(1) series and, further, provides a presumptive statistical procedure to indicate loss of material from an inventory whose material accounting is based on a material balance. In particular, this procedure is shown to provide a successful method for the determination of losses from an inventory containing special nuclear material (SNM). The AR(1) case is considered specifically even though generalizations to AR(p) and ARMA models are available, because it is appropriate for the inventory data sets considered here, as is shown later. It must be kept in mind that each time series considered should go through model identification and diagnostic checking steps, to verify that the final model chosen is indeed appropriate.

In a seminal paper on outliers in time series, Fox[16] identified two major types of outliers, those that persist consistent with the autoregressive structure [called innovations outliers (IO)] and those that have an "additive transient character unrelated to the autoregressive specification" [called additive outliers (AO)]. The GM procedure identifies and distinguishes between these two types of outliers. This is an important point, because IOs represent continuing inventory loss, while AOs represent a one-time loss.

MODIFIED GENERALIZED M-ESTIMATOR PROCEDURE

The procedure to be used is a modification of a method first proposed by Denby and Martin[19] for robust estimation in a set of data that can be described by an AR(1) time series. The general method has been described by Martin.[23,24] The GM procedure does two things. First, it decreases the effect of any outlying points on the estimate produced by the algorithm, without having to remove any such points from the data set. Second, it can be used to identify outlying points that exist in the data set.

It is necessary to verify the assumptions that the series under consideration is both stationary and AR(1).[15,23,26] All series examined in this study were such. In case a series is encountered that does not meet these assumptions, generalizations of the procedure described herein, employing the same ideas, are available.[24,26] Necessary procedures as well as methods of achieving stationarity are described in Hull and Nie,[20] Box and Jenkins,[17] Makridakis and Wheelwright,[21] and Pankratz,[30] as well as in Martin.[23] Because outliers are possible members of data sets, it is important to apply diagnostic checks after the fitting step in order to be sure that the final fitted model explains the data well.

COMPUTING AND GM ESTIMATE

Recall that an M estimate, $\hat{\beta}$, of the parameter vector, β, in the linear regression model (for basic definitions see footnote 17)

$$\mathbf{Y} = \boldsymbol{X}\beta + \epsilon \tag{1}$$

is a solution to

$$\min \sum_{i=1}^{n} \rho[(Y_i - \boldsymbol{X}_i'\hat{\beta})/\hat{\sigma}] \tag{2}$$

with respect to $\hat{\beta}$, where $\rho(\cdots)$ is a suitable loss function and $\hat{\sigma}$ is a suitable robust scale factor for the residual vector, $\hat{\epsilon}$. The function

$$\psi(\cdots) = \rho'(\cdots) \tag{3}$$

is called the influence function corresponding to the loss function, $\rho(\cdots)$. Proceeding using the w procedure of Beaton and Tukey[2] (there are other possibilities),[17] define a set of weights by

$$W_i = \psi(\Delta_i/\hat{\sigma})/(\Delta_i/\hat{\sigma}) \tag{4}$$

where

$$\Delta_i = Y_i - \boldsymbol{X}_i'\hat{\beta} \tag{5}$$

Letting W be the diagonal matrix of weights from (5) yields the well-known solution[2]

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}) \tag{6}$$

which must be solved iteratively. Details are given in footnotes 2 and 11. The extension of these ideas to time series models was developed by Martin and co-workers.[14,23-26]

We now consider the computational algorithm for GM estimation in the special case of an AR(1) model, first given by Denby and Martin.[14] Let $\{Z_t\}$, $t = 1, 2, \ldots, n$, denote a set of time series observations. Then

$$A_t = Z_t - \hat{\mu} \tag{7}$$

0095-2338/86/1626-0023$01.50/0

is a centered series where $\hat{\mu}$ is some appropriate estimate of location (e.g., the mean of the series). A robust estimate of location should be used, and there are many possible choices. For this work, the simplest, the series median, which is itself an M estimator, was used. Now let

$$Y_t = A_{t+1} \tag{8}$$

and

$$X_t = A_t \tag{9}$$

Introducing a_t, a white noise term, yields the following AR(1) model:

$$Y_t = \beta Y_{t-1} + a_t = \beta A_t + a_t = \beta X_t + a_t \tag{10}$$

Let $\hat{\beta}$ be the GM estimate of β, which can now be computed with the definitions of Denby and Martin.[14] In this work, ψ will always denote the Huber ψ function[17] defined by

$$\psi(S) = \begin{cases} -k, & S < -k \\ S, & -k \leq S \leq k \\ k, & S > k \end{cases} \tag{11}$$

where k is a tuning constant of value 1 in this study. $\hat{\beta}$ is computed in three stages. The first stage requires centering the series and calculating what are called the location weights. The second stage requires computing what are called the autoregressive residual weights. The third stage requires computing the GM estimate iteself. The procedure is iteratively continued until a preset convergence criterion is satisfied.

Stage one begins by computing

$$S_Y = \text{median}\,(|X_i|)/0.6745 \tag{12}$$

a robust scale factor. Then, compute the location weights, W_3 (to downweight additive outliers), by

$$W_3(X_i) = \psi\left(\frac{X_i}{S_Y}\right) \Big/ \left(\frac{X_i}{S_Y}\right) \tag{13}$$

Observe that by choosing other ψ functions and continuing iteratively one can generate any desired M estimate of location. Let $\mathbf{Y}$ denote the column vector of Y's, $\mathbf{X}$ the column vector of X's, and $\mathbf{a}$ the white noise vector. Thus, we may write model 10 as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{a} \tag{14}$$

with prediction equation

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \tag{15}$$

Notice that in the AR(1) case $\hat{\beta}$ is a 1×1 matrix. Now, compute the residual vector

$$\Delta = \mathbf{Y} - \hat{\mathbf{Y}} \tag{16}$$

by using an initial estimate of β based on least squares. Then, compute the residual scale factor

$$S = \text{median}\,(|\Delta_i|)/0.6745$$

and scaled residual vector

$$\Delta_2 = \Delta/S$$

Finally, the autoregressive residual weights (to downweight innovations outliers) are computed from

$$W_2\left(\frac{\Delta_i}{S}\right) = \psi\left(\frac{\Delta_i}{S}\right) \Big/ \left(\frac{\Delta_i}{S}\right) \tag{17}$$

Letting $\mathbf{W}_2$ and $\mathbf{W}_3$ be the respective diagonal matrices of weights gives the GM estimate in final form (footnote 28, p 357):

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}_2\mathbf{W}_3\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}_2\mathbf{W}_3\mathbf{Y}) \tag{18}$$

The process is then continued iteratively until convergence (two successive estimates of β differ by less than 0.0001) is achieved.

The programs, with output, used in the present study are available from the author and can also be found in Booth.[7] An important fact to observe for use in applications of this technique is that both sets of weights, computed as described above, will be bounded by zero and one, inclusively. Thus, a value less than 1 for one or both of an observation's weights decreases the effect of that point on the computed estimate and as a consequence serves as a flag for identifying the point as an outlier (i.e., a point that is in some manner different from the other points in the data). As shown previously,[5,10] a negative-signed residual (either location or autoregressive) associated with such a point is an indicator of a potential inventory loss.

A MODEL FOR EARLY DETECTION OF LOSS FROM A NUCLEAR MATERIAL INVENTORY

The management of nuclear material inventories is a difficult problem. As Goldman et al.[7] observed, "...the nuclear community places great emphasis on security and prevention of theft or loss of special nuclear material (SNM)...." Following Chernick et al.,[13] we will refer to the problem of detecting such theft or loss as the nuclear safeguards problem. The importance of this problem has been discussed previously (see, for example, footnotes 1, 13, 15, 17, and 32). Clearly, the theft or accidental loss of nuclear material could have a significant negative impact on any population of living organisms, including humans. Therefore, early detection of any such loss becomes of paramount importance to the management of a nuclear inventory. As Chernick et al. have remarked about currently existing procedures for dealing with this problem (other than the one they proposed), "A major drawback of previous methods is their inability to detect the theft until several periods after the loss has occurred." Clearly, the sooner we are able to detect such a loss, the sooner we can do something about it. As will be shown in this study, the GM-estimation procedure provides a simple and immediate method of detecting any such loss.

The principle fact that allows use of the GM procedure with nuclear inventories is that SNM inventory accounting is based on the concept of material balance.[17] Thus, inventory differences are computed periodically to detect any losses that may have occurred.[13] These differences are, of course, based on analytical chemical determinations of the amount of material present and may also involve statistical sampling procedures. The sequence of differences should, however, be a stationary dependent process (with zero mean) if no loss has occurred (footnote 33, p 213).[13,15,17] After the model identification, parameter estimation, and diagnostic checking steps, discussed later, an AR(1) model was found to be reasonable for all series considered in this paper. In the absence of loss, the random error component should arise only from experimental error in the determination of the amount of nuclear material present (and sampling error, if a sampling procedure is also used). A loss then can be characterized as an outlying point with respect to this process, and thus the weights of the GM-estimation procedure (recall that the purpose of the weights was to decrease the effect of outlying points on the estimate of the model parameter) should provide a method for detecting outlying points and hence provide a flag for such a loss. A negative residual, coupled with a low weight from either of the two weighting schemes, will indicate that a possible loss has occurred, thus hopefully triggering a major investigative effort to determine the exact circumstances involved. Note that the further the weight value is from one, the greater the possibility that loss has occurred. A large

Table I. Simulated Loss Data (Data Set 8)

time	observation	time	observation
1	1.50	7	−0.75
2	−1.00	8	0.80
3	1.00	9	0.70
4	−0.20	10	0.60
5	−0.30	11	−0.50
6	0.50	12	−4.00

experimental or sampling error could also give such weight values. But if such were the case, it would still be useful information since it would indicate that changes in one or the other of these procedures is necessary to accurately monitor the inventory material. Indeed, it is possible to say more. As Denby and Martin remark,[14] the location weights are designed to discount additive outliers (AO), that is, outliers that affect only a single observation, while the autoregressive residual weights are designed to discount innovations outliers (IO), those that also affect subsequent observations. Thus, the GM method provides a presumptive test for whether a loss is one time only (AO) or continuing (IO). The GM method is the only currently available procedure that allows such a determination. Graphical extensions of this method also exist.[11,25]

The present study tested this model on a series of eight data sets. Data sets 1–7 have been reported previously by Chernick et al.[13] These data were taken from U.S. Department of Energy reports involving various isotopes, from various nuclear facilities. Data set 8 (given in Table I) is a simulated one, in which the last observation is an obvious loss. In the case of each data set, the hypothesis that it was a stationary AR(1) process was checked by plotting the series, as well as by computing autocorrelations and partial autocorrelations, using the SPSS Procedure, Box-Jenkins, see footnotes 20, 21, and 30. Because outliers may be present, these procedures are suspect,[13] and thus, diagnostic checking of the final fitted model is extremely important. Therefore, after the series was fitted with the GM-estimation procedure, the autoregression residuals were examined as described by Hull and Nie,[20] Box and Jenkins,[12] Markadakis and Wheelwright,[21] and Pankratz.[30] These checks included plots of the autoregressive residual series (both raw and scaled), as well as examination of their autocorrelations and partial autocorrelations, both computationally and graphically. Further, the autoregressive residuals were plotted vs. the lagged centered series values (i.e., the predictor variable) as suggested by Neter and Wasserman.[29] To add emphasis, the analysis of the final residuals is extremely important in this case because outlying points can seriously affect the model identification step. All of these procedures indicated that the models fitted by the GM-estimation procedure met all required assumptions and were thus appropriate for all eight series under consideration. As an example of the checks, scaled autoregressive residual plots, including the correlation studies, from data set 7 can be found in reference 7. Copies of the programs for series identification and diagnostic residual checking as well as the plots themselves may be obtained from the author.

RESULTS

The results of the GM-estimation procedure for the eight data sets are given in Tables II–IX. The computations were performed with a value of 1 for the tuning constant in the GM procedure. There are several important things to notice in these tables. To begin, the GM-estimation procedure finds all of the outliers reported by Chernick et al.[13] using their procedure. These outliers are identified in the GM method by the fact that they have at least one GM procedure weight value that is less than 1. A particular value for the cutoff depends on the choice of tuning constant. In this study a rule identifying an outlier by a weight value ≤ 0.75 is reasonable. Thus, it may be concluded that the GM method is successful in identifying outliers in AR(1) time series models. It is further observed that in data sets 6 and 8 the final observation is an outlier. Thus we see that an analyst can detect an outlier (i.e., a potential loss) as soon as new inventory accounting data

Table II. Summary of Data Set 1[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		0.836364
3	1	1		0.536364
4	1	1		0.857143
5	0.739518	1		−4.92338
6	0.724028	0.707595	*	−5.02857
7	1	0.457855		3.63766
8	1	1		1.83766
9	1	1		−3.18571
10	0.514682	1	*	−7.07403
11	1	0.389177		2.42078
12	1	1		−3.27792
13	1	0.864838		−0.35065
14	0.762434	1		−4.77532
15	1	0.566076		2.01429
16	1	1		−2.11948
17	1	1		−1.21429
18	1	1		−1.25455
19	1	1		1.14545
20	0.831403	1		4.37922
21	1	0.691871		0.311688
22	1	1		2.5961
23	1	0.972943		−0.0883114
24	0.526377	1	*	6.91688
25	1	0.420732		−2.97922
26	1	1		−3.5
27	1	0.889548		−2.49091
28	1	0.798312		2.82013
29	1	1		0.996753

[a] Median of original series = −1.3. Number of iterations to converge = 6. GM $\hat{\beta}$ = 0.4026. S = 3.641. S_Y = 3.113. [b] An asterisk (*) in the Chernick column in this and subsequent tables indicates that the observation was reported to be an outlier by Chernick et al.[11]

Table III. Summary of Data Set 2[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		−0.682467
3	0.312801	1		−6.44091
4	0.753296	0.277984		−2.67403
5	1	0.77352		−0.434419
6	1	1		−0.417533
7	1	1		0.276623
8	1	1		−0.0824671
9	1	1		0.394156
10	1	1		0.0233772
11	1	1		1.8
12	1	0.988386		−1.0948
13	1	1		−0.570132
14	1	1		0.170779
15	1	1		1.01169
16	1	1		−0.641557
17	0.891861	1		2.25909
18	0.794892	0.77352		2.53442
19	1	0.74129		0.840263
20	0.432438	1		−4.65909
21	0.152921	0.378531		−13.1747
22	1	0.137914		1.94609
23	0.277588	0.658924		7.2578
24	0.145108	0.250577		−13.8851
25	0.349595	0.124412		5.76427
26	0.0258919	0.26956	*	−77.8143
27	1	0.0227506		1.62976
28	0.0823592	0.286951	*	24.4623
29	1	0.0738214		0.208474

[a] Median of original series = −1.4. Number of iterations to converge = 10. GM $\hat{\beta}$ = −0.0584. S = 2.0197. S_Y = 1.779. [b] See Table II.

Table IV. Summary of Data Set 3[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		2.95848
3	1	1		0.255232
4	1	1		6.42609
5	1	1		2.95848
6	1	1		-21.4448
7	1	1		-2.87462
8	1	1		8.5867
9	0.644873	1		-44.1
10	0.653736	0.685819		-43.5047
11	1	0.40434		-25.4186
12	0.372294	0.385282	*	-76.3929
13	0.445414	0.228953	*	-63.856
14	1	0.191908		16.64
15	1	0.317696		12.2581
16	0.645555	0.546919	*	-44.0567
17	0.591469	0.363081	*	-48.0866
18	0.271156	0.282133		104.874
19	1	1		-10.3378
20	1	1		-27.5674
21	0.319618	1		88.9767
22	0.565691	0.405968		-50.2685
23	1	1		10.6549
24	0.750936	1		-37.8706
25	1	1		-16.9203
26	0.478658	0.806523		59.4117
27	1	0.922092		-15.0763
28	1	1		12.4809
29	1	1		-4.38649

[a] Median or original series = -11.6. Number of iterations to converge = 6. GM $\hat{\beta}$ = 0.7096. S = 28.44. S_Y = 30.24. [b] See Table II.

Table V. Summary of Data Set 4[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		0
3	1	1		0.5
4	0.916075	1		-0.0809211
5	1	1		-0.0190789
6	0.232321	1	*	-0.319079
7	0.130868	0.211797	*	0.566448
8	0.34249	0.211797		-0.216448
9	1	1		-0.05
10	1	1		-0.0190789
11	0.916075	1		0.0809211
12	1	1		-0.0309211
13	1	1		0
14	1	1		0.05
15	1	1		-0.0309211
16	1	1		0
17	1	1		0
18	1	1		0
19	0.296516	1		-0.25
20	0.187479	0.296516	*	-0.395394
21	1	0.13478		-0.00986746
22	0.317389	0.211797	*	-0.233552
23	1	0.164731	*	-0.0717097
24	1	0.211787		0.066488
25	1	0.494193		-0.0572366
26	1	0.494193		-0.0572366
27	1	0.494193		0.0427634
28	1	1		0.0309211
29	1	1		0

[a] Median of original series = -0.05. Number of iterations to converge = 6. GM $\hat{\beta}$ = 0.6184. S = 0.0741. S_Y = 0.0741. [b] See Table II.

become available. Finally, the tables show that all possible outliers are distinguished by the GM procedure not just those that satisfy a preset probability criterion, as in the case of the Chernick et al. procedure. This makes the GM procedure conservative in the sense that it is less likely that a loss or theft will go unnoticed. Further, the autoregressive residuals (observed series value minus predicted value) are a measure of

Table VI. Summary of Data Set 5[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		-0.0062654
3	1	1		0.05
4	1	1		0.0437346
5	1	1		-0.206265
6	0.673259	1		-0.474938
7	1	0.593032		0.062654
8	0.532915	1		0.6
9	1	0.494193		-0.0751848
10	1	1		-0.3
11	1	0.988386		-0.162408
12	1	1		0.225062
13	1	1		-0.0250616
14	0.532915	1		-0.6
15	0.441166	0.494193		-0.724815
16	1	0.370645		0.150246
17	0.321766	1	*	0.993735
18	0.98274	0.296516		-0.325308
19	1	1		-0.0749384
20	0.77508	1		0.412531
21	0.491811	0.74129		-0.650123
22	0.752738	0.494193		-0.424815
23	1	0.593032		-0.137346
24	0.170539	1	*	-1.87494
25	1	0.156061		0.238085
26	1	1		-0.2
27	1	1		0.0250616
28	1	1		0.3
29	1	0.988386		0.0124076

[a] Median of original series = -0.1. Number of iterations to converge = 3. GM $\hat{\beta}$ = 0.1253. S = 0.3197. S_Y = 0.2965. [b] See Table II.

Table VII. Summary of Data Set 6[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		0.0290798
3	1	1		0.0290798
4	1	1		0.17908
5	1	0.847188		0.00355868
6	1	1		0.17092
7	0.363851	0.847188		-0.69441
8	0.17558	0.204494	*	-1.44331
9	1	0.111893		-0.19123
10	1	1		-0.0209202
11	0.770049	1		-0.32908
12	1	0.456178		-0.0780376
13	0.305648	1		-0.82908
14	0.384203	0.179707		-0.659634
15	0.875718	0.282396		0.289324
16	0.96047	0.395355		-0.263803
17	1	0.456178		0.121962
18	0.639177	0.847188		-0.396441
19	0.512636	0.348842		-0.494357
20	1	0.348842		-0.0443568
21	1	1		0.0290798
22	1	1		-0.0209202
23	1	1		0.0209202
24	1	1		0.0290798
25	1	1		0.0290798
26	0.078675	1	*	-3.22092

[a] Median of original series = -0.075. Number of iterations to converge = 10. GM $\hat{\beta}$ = -0.1632. S = 0.2534. S_Y = 0.1483. [b] See Table II.

loss size in mass units and are thus a measure of the physical significance of a loss. Should approximate tests of statistical significance of loss size be desired, they are routinely available (footnote 12, p 135; footnote 4, p 401; footnote 27, pp 76–78, 193, 223; footnote 19; footnote 29, pp 233, 326–328; footnote 30, Chapter 10). On the basis of these observations it therefore can be concluded that the GM procedure gives a successful method for the early detection of loss in inventories with material accounting based on a material balance scheme, and in addition, by consideration of which of the weights is low,

Table VIII. Summary of Data Set 7[a]

observation	autoregressive residual wt	location wt	Chernick[b]	autoregressive residual
2	1	1		0.707121
3	1	1		0.707121
4	1	1		0.657121
5	1	1		0.269904
6	1	1		-1.02783
7	0.829725	1		2.0301
8	0.915442	0.823655		1.83981
9	1	0.6446		-0.588024
10	1	1		-1.5
11	1	0.988386		-0.316506
12	0.0914367	1	*	-18.421
13	0.590201	0.0797086		2.85532
14	0.545877	0.780305		3.08576
15	0.691595	0.570223		2.43528
16	0.890108	0.478252		-1.89255
17	1	1		-0.0187713
18	1	1		-0.923301
19	1	1		-1.24434

[a] Median of original series = -1. Number of iterations to converge = 19. GM $\hat{\beta}$ = 0.2557. S = 1.684. S_Y = 1.483. [b] See Table II.

Table IX. Summary of Data Set 8[a]

observation	autoregressive residual wt	location wt	autoregressive residual
2	1	0.713835	-0.577127
3	1	0.83798	0.361997
4	1	1	0.0106979
5	1	1	-0.598523
6	1	1	0.159042
7	1	1	-0.751477
8	1	1	0.268085
9	1	1	0.825828
10	1	1	0.683393
11	1	1	-0.459042
12	0.193345	1	-4.42583

[a] Median of original series = 0.15. Number of iterations to converge = 6. GM $\hat{\beta}$ = -0.424. S = 0.8556. S_Y = 0.9637.

the procedure distinguishes presumptively between a continuing loss (IO) and a single loss (AO).

Advantages of the GM-Estimation Procedure. The GM-estimation procedure, as described herein, has a number of advantages over previously suggested solutions to the nuclear safeguards problem. Currently, only the Chernick et al.[13] procedure identifies outlying points (possible material losses) immediately on making an inventory measurement. As seen in Tables II–IX, the GM method shares this advantage, while identifying all the outliers that the Chernick et al. method does. In addition, the GM method has advantages over other possible procedures.

(1) The procedure is easy to implement and use because it is based on the commonly used technique of regression analysis. The only requirements are a computer, a program to implement the GM estimates (available from the author or Booth[7]), and access to programs for diagnostic checking (available in most statistical computer packages, e.g., SPSS, SAS, IMSL, and Minitab). The determination of possible loss points (i.e., outliers), either at past times or at present times, is particularly easy. One simply executes the GM-estimation program and looks for observations that have weight values less than 1, for either set of weights. The Chernick et al. procedure requires arranging the computational results in a special tabular format and then looking for a special pattern. It is also observed that once the GM procedure has been implemented all that needs to be done to check new observations is to add them to the data set, execute the GM program, and examine as before.

(2) The Chernick et al. procedure uses the sample autocorrelations to compute the influence function that then allows the analyst to compute the statistics actually used in the outlier test. These autocorrelations are sensitive to the presence of outliers as the authors themselves remark,[13] and thus, the actual hypothesis test is approximate. The GM procedure does not suffer from this sensitivity because all of the estimates used in the procedure are robust.

(3) The GM-estimation procedure provides more information per computer run than other possible procedures that use sets of deleted observations.[10] Further, the GM procedure distinguishes between a loss (characterized by residual with negative sign coupled with a low weight from the algorithm) and an overage (the corresponding case with positive residual sign).[5,10] This information is provided immediately by the GM procedure. It is also observed that if GM estimation is being used as the estimation method in a Box–Jenkins inventory forecasting procedure,[23,31] only minimal modification of the procedure would result in all of this information being received essentially gratis from the forecasting procedure's output. Also note that since the procedure does use the full data set and not one with some observations deleted, as some other outlier detection methods[3] do, it can distinguish between one-time and continuing losses.

(4) The GM procedure is applicable, with no modification required, not only to nuclear material inventories but to any inventory system whose material accounting system can be based on a material balance scheme. The drug industry immediately suggests itself as a possible application. Other applications of the GM procedure to quality control are discussed elsewhere.[11]

(5) The GM procedure provides weights for all observations and, thus, as shown in the tables, indicates all outliers. The procedure is conservative because all possible loss points are indicated, not just those that satisfy a hypothesis test at some arbitrary preset level. As indicated previously, the GM procedure does allow such tests if they are desired. Further, it allows the analyst to distinguish between one-time and continuing losses by distinguishing between additive and innovations outliers.[22,25] GM is the only currently available procedure that has this property. This increase in the amount of available information is important since the goal of the method is to detect and therefore allow the stoppage of losses from inventories of *weapons grade nuclear material.*

(6) Should an inventory analyst care to do so, approximate hypothesis tests or alternatively confidence intervals may be constructed, with the autoregressive residuals, to determine if a statistically significant difference between the expected value and observed value exists. These tests are standard, and formulae exist in the literature [e.g., footnotes 10 (p 135), 4 (p 401), 27, 29 (pp 233, 326–328), and 30 (Chapter 10)]. Thus, the autoregressive residuals give an immediate estimate of the amount of any such loss, a very important quantity, and confidence intervals can be constructed from these estimates if desired. The hypothesis tests considered by Chernick et al. concern sums, differences, and products of series observations and as such do not provide estimates of properties of immediate physical interest as the GM procedure does.

CONCLUSIONS

On the basis of the results of this study it can be concluded that (1) the GM-estimation procedure is a successful method for identifying outliers in an AR(1) process, (2) the procedure is a useful method for the early detection of loss in any inventory whose material accounting procedure is based on material balance, in particular nuclear materials, and (3) the procedure provides several advantages that existing procedures do not. Among these are (a) the GM procedure provides direct estimates of the physical amount of material lost, (b) the procedure detects all outliers (i.e., losses), not just those at some preset probability level, and (c) the procedure allows an analyst

to distinguish between one-time and continuing losses.

ACKNOWLEDGMENT

I thank Professor T. L. Isenhour, Professor K. N. Berk, and the reviewers for many valuable suggestions.

REFERENCES AND NOTES

(1) "Atlantic Council's Nuclear Fuels Policy Working Group, Nuclear power and Nuclear Weapons Proliferation"; Atlantic Council of the U.S.: Washington, DC, 1978.
(2) Beaton, A. E.; Tukey, J. W. "The Fitting of Power Series Meaning Polynomials, Illustrated on Band-Spectroscopic Data". *Technometrics* **1974**, *16*, 147-186.
(3) Beckman, R. J.; Cook, D. "Outlier...s". *Technometrics* **1983**, *25*, 119-149.
(4) Berenson, M. L.; Levine, D. M.; Goldstein, M. "Intermediate Statistical Methods and Applications"; Prentice-Hall: Englewood Cliffs, NJ, 1983.
(5) Booth, D. E. "Regression Methods and Problem Banks"; COMAP, Inc.: Lexington, MA, 1985; Module No. 626.
(6) Booth, D. E.; Montasser, S. "Robust Discriminant Analysis and the Periods of Modern Egyptian Economic Developments". *Ind. Math.* **1985**, *35* (1), 81-91.
(7) Booth, D. E. "Some Applications of Robust Statistical Methods to Analytical Chemistry". Doctor of Philosophy Dissertation, The University of North Carolina at Chapel Hill, 1984.
(8) Booth, D. E. "A Model for the Early Detection of Loss in Nuclear Material Inventories", Tabor School of Business and Engineering, Millikin University: Decatur, IL, 1983; Faculty Working Paper 83-4.
(9) Booth, D. E. "The Analysis of Outlying Data Points Using Robust Regression: A Multivariate Problem Bank Identification Model". *Decis. Sci.* **1982**, *13*, 71-81.
(10) Booth, D. E. "The Analysis of Outlying Data Points by Robust Regression: I. A Model for the Identification of Problem Banks". *Ind. Math.* **1981**, *31* (2), 85-98.
(11) Booth, D. E.; Isenhour, T. L. "An Application of Robust Time Series Analysis to the Interpretation of Quality Control Charts". Submitted to *J. Quality Technol.*
(12) Box, G. E. P.; Jenkins, G. M. "Time Series Analysis Forecasting and Control"; Holden Day: San Francisco, 1976; revised ed.
(13) Chernick, M. R.; Downing, D. J.; Pike, D. H. "Detecting Outliers in Time Series Data". *J. Am. Stat. Assoc.* **1982**, *77*, 743-747.
(14) Denby, L.; Martin, R. D. "Robust Estimation of the First-Order Autoregressive Parameter". *J. Am. Stat. Assoc.* **1979**, *74*, 140-146.
(15) Downing, D. J.; Pike, D. H.; Morrison, G. W. "Analysis of MUF Data Using ARIMA Models". *Nucl. Mater. Manage.* **1978**, *7* (4), 80-86.
(16) Fox, A. J. "Outliers in Time Series Data". *J. R. Stat. Soc., Ser. B*, **1972**, *34*, 340-363.
(17) Goldman, A. S.; Picard, R. R.; Shipley, J. P. "Statistical Methods for Nuclear Material Safeguards: An Overview". *Technometrics* **1982**, *24*, 267-274.
(18) Hillier, F. S.; Lieberman, G. J. "Operations Research"; Holden-Day: San Francisco, 1974; 2nd ed.
(19) Hogg, R. V. "Statistical Roubustness: One View of Its Use in Applications Today". *Am. Stat.* **1979**, *33*, 108-115.
(20) Hull, C. H.; Nie, N. H. "SPSS Update 7-9"; McGraw-Hill: New York, 1981.
(21) Makridakis, S.; Wheelwright, S. "Forecasting Methods and Applications"; Wiley: New York, 1978.
(22) Mallows, C. "Robust Methods—Some Examples of Their Use". *Am. Stat.* **1979**, *33*, 179-184.
(23) Martin, R. D. "Robust Methods for Time Series". In "Applied Time Series Analysis II"; Findley, D. F., Ed.; Academic Press: New York, 1981.
(24) Martin, R. D. "Robust Estimation for Time Series in Autoregressions". In "Robustness in Statistics"; Launer, R. L.; Wilkinson, G., Eds.; Academic Press: New York, 1979.
(25) Martin, R. D.; Zeh, J. E. "Determining the Character of Time Series Outliers". "Proceedings of the American Statistical Association"; Business and Economics Statistics Section, American Statistical Association: Washington, DC, 1977.
(26) Martin, R. D. "Time Series: Model Estimation, Data Analysis, and Robust Procedrues". In "Modern Statistics: Methods and Applications"; American Mathematical Society: Providence, RI, 1980.
(27) Montgomery, D.; Johnson, L. "Forecasting and Time Series Analysis"; McGraw-Hill: New York, 1976.
(28) Mosteller, F.; Tukey, J. "Data Analysis and Regression"; Addison-Wesley: New York, 1977.
(29) Neter, J.; Wasserman, W. "Applied Linear Statistical Models"; Irvin: Homewood, IL, 1974.
(30) Pankratz, A. "Forecasting with Univariate Box-Jenkins Models"; Wiley: New York, 1983.
(31) Preston, D. B. "Robust Forecasting". In "Applied Time Series Analysis"; Findley, D. F., Ed.; Academic Press: New York, 1981.
(32) "Nuclear Energy and National Security"; Research and Policy Committee, Committee for Economic Development: New York, 1976.
(33) Wagner, H. M. "Principles of Management Science"; Prentice-Hall: Englewood Cliffs, NJ, 1970.

Cambridge Crystallographic Data Centre. 7. Estimating Average Molecular Dimensions from the Cambridge Structural Database

ROBIN TAYLOR* and OLGA KENNARD

Crystallographic Data Centre, University Chemical Laboratory, Cambridge CB2 1EW, England

Received June 11, 1985

The Cambridge Structural Database contains the atomic coordinates of some 40 000 organocarbon crystal structures. It is therefore likely to be a major source of data in future determinations of average molecular dimensions. From a statistical point of view, there is no single "optimum" method of obtaining such averages. Practical guidelines are suggested here on the basis of computer-simulation results.

INTRODUCTION

The Cambridge Crystallographic Data Centre (CCDC) maintains and distributes the Cambridge Structural Database (CSD), which currently contains the results of about 40 000 organocarbon crystal structure determinations. Previous papers in this series[1-3] document the development of CSD, concentrating mainly on the organization and content of the database and its associated search and retrieval software. This software is still under active development in Cambridge, but increasingly, the CCDC is addressing the problems of database *utilization*.

The value of CSD as a research tool is well-known,[4] and there are currently over one-hundred published papers describing CSD utilization projects. Moreover, it is now recognized as an important facility in molecular graphics.[5,6] One of the most significant areas in which CSD can be of use is in the estimation of average molecular dimensions. This is a major objective of many chemical and crystallographic research projects. It is also of fundamental importance in molecular graphics, e.g., in the construction of "fragment libraries".

The statistical problems involved in estimating average molecular dimensions were examined from a theoretical point of view in two earlier papers.[7,8] In the present paper, we have two objectives. First, we use the theoretical results obtained earlier to devise practical guidelines for estimating average molecular dimensions from CSD. Second, we outline the computer-simulation algorithm used to obtain these guidelines,

0095-2338/86/1626-0028$01.50/0