A Method for Early Discovery of Poisoning in Catalytic Chemical Processes

DAVID E. BOOTH*

Department of Mathematical and Computer Sciences, Millikin University, Decatur, Illinois 62522

THOMAS L. ISENHOUR

Department of Chemistry, Utah State University, Logan, Utah 84322

To provide effective industrial process quality control, data from the process must be collected, analyzed, and reported, both reliably and speedily. This paper presents a data analytic method, based on generalized M estimation, for the immediate discovery of the effects of poisoning in an industrial chemical reaction mixture. The model is tested with simulations based on real data. This procedure can be implemented easily on either a mainframe computer or microcomputer.

INTRODUCTION

To provide effective industrial process quality control, data from the process must be collected, analyzed, and reported both reliably and speedily. In the past, the emphasis in the development of new procedures was on reliability. Now, with microcomputers readily available, it is feasible to minimize the amount of time between the start of some deterioration in process quality and its discovery, i.e., the detection time, as well. This paper presents a technique that can easily be implemented on either a mainframe computer or microcomputer to quickly provide presumptive discovery information on the possible poisoning of a catalytic chemical process.

Catalyzed chemical reactions are of great importance to both science and industry.[19,21] A major difficulty with such processes is that almost all catalysts are capable of being deactivated by small amounts of compounds called poisons.[25] The action of a poison can be either temporary or permanent depending on the particular chemical species involved, but in either case, discovering and correcting any such poisoning are major concerns in industrial chemistry. Starkey[20] describes the problem: "Catalysts which are particularly prone to poisoning, or are so valuable as to require extensive protection, can place major constraints upon the process design. Almost all aspects of plant operation have to be examined with a view to limiting poison ingress or production.... Any catalyst modification which enhances resistance to poisoning is likely to provide process and cost benefit.... Thus, it is important to protect catalysts from poisons if at all possible." If complete protection is not possible, the best alternative is to discover the effect of a poison immediately upon its entry into a reaction system in order to minimize damage to the catalyst as well as process down time. This paper presents a data analytic method, based on generalized M estimation (GM), that allows the quick presumptive identification of poisoning activity, by identifying points that have lower percentage yields than those points from a catalytic process without poisoning. These points are identified by a procedure that indicates that they are not likely to have arisen from the same stochastic process as the points where poison activity is not occurring. Such a procedure should help the industrial chemist or chemical engineer optimize process production.

STRATEGY

The development of the proposed statistical model to discover the effects of catalytic poisoning is based on the observation that prior to any appreciable deactivation of the catalyst the series of percentage yields (or some other similar statistic) should be well described by a stationary time series, that is, a series that oscillates about a constant location pa-

rameter.[13] This is a reasonable assumption, since a nonstationary series exhibits trend and such a process would be less likely to be of industrial interest. Further, we consider a stationary model that is first-order autoregressive, AR(1). An AR(1) model is one of the form $Y_t = \beta Y_{t-1} +$ error, where the Y_t's are centered series values.[7] We consider such a model for the following reasons: (1) The data seem to be well described by such a model. (2) It is reasonable to expect that many other sets of data will be also.[7] (3) If other models [either nonstationary or other than AR(1)] are found to be necessary, the same ideas will still apply.[14]

Now, recall that any time series is generated by some underlying stochastic process (in this case, one that depends on the parameters of the chemical system). Poisoning of the catalyst will cause a decrease in the process percentage yield because it deactivates the catalyst. Thus, from the time of entry of the poison the time series will be generated by a new stochastic process that depends on new parameters of the chemical system and therefore must reflect the action of the poison. Since the new stochastic process is different from the old, we expect the time series points generated after appreciable catalyst deactivation to be innovations outliers (i.e., not representative) with respect to the series generated by the old stochastic process.[14] We make use of this in our procedure, which is based on identifying outliers in the time series. The examination of real yield data from catalytic processes (e.g., Table I, to be described fully later) indicates that it is difficult to quickly identify these points by either inspection or ordinary control chart procedures. In this study, the method of generalized M estimation (GM) developed by Denby and Martin,[9] was used as a means of identifying outlying points.

MODIFIED GENERALIZED M ESTIMATOR PROCEDURE

The procedure to be used is a modification of a method first proposed by Denby and Martin[9] 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.[14,15] This 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). Necessary procedures as well as methods of achieving stationarity are described in Hull and Nie,[12] Box and Jenkins,[7] and Makridakis and Wheelwright,[13] as well as in Martin.[15] 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

Table I. Data Sets

	day (% yield)							contaminant
case	1	2	3	4	5	6	contaminant	concn
1A	87.6	95.1	92.6	87.0	85.0	77.5	acetylene	0.09%
2A	89.4	96.8	90.8	96.0	85.0	77.5	acetylene	0.09%
3A	89.0	96.3	94.7	87.0	89.0	77.5	acetylene	0.09%
1P	87.6	95.1	92.6	87.0	85.0	71.0	phosphine	0.07 ppm
2P	89.4	96.8	90.8	96.0	85.0	71.0	phosphine	0.07 ppm
3P	89.0	96.3	94.7	87.0	89.0	71.0	phosphine	0.07 ppm
1S	87.6	95.1	92.6	87.0	85.0	94.3	hydrogen sulfide	0.04%
2S	89.4	96.8	90.8	96.0	85.0	94.3	hydrogen sulfide	0.04%
3S	89.0	96.3	94.7	87.0	89.0	94.3	hydrogen sulfide	0.04%
1N	87.6	95.1	92.6	87.0	85.0			
2N	89.4	96.8	90.8	96.0	85.0			
3N	89.0	96.3	94.7	87.0	89.0			

the final fitted model explains the data well.

COMPUTING THE GM ESTIMATE

Recall that an M estimate, $\hat{\beta}$, of the parameter vector, β, in the linear regression model (for basic definitions, see footnote 11)

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

is a solution to

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

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),[11] define a set of weights by

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

where

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

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

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

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.[9,14,15]

We now consider the computational algorithm for GM estimation in the special case of an AR(1) model, first given by Denby and Martin.[9] Let $\{Z_t\}$, $t = 1, 2, ..., n$, denote a set of time-series observations. Then

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

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.[9] In this work, ψ will always denote the Huber ψ function[11] defined by

$$\psi(l) = \begin{cases} -k, l < -k \\ l, -k \leqslant l \leqslant k \\ k, l > k \end{cases} \tag{11}$$

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 itself. 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

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

Observe that by choosing other ψ functions and continuing iteratively one can generate any desired M estimate of location. Let Y denote the column vector of Y's, X the column vector of X's, and 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

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

Letting $\mathbf{W}_2$ and $\mathbf{W}_3$ be the respective diagonal matrices of weights gives the GM estimate in final form:

DISCOVERY OF POISONING

J. Chem. Inf. Comput. Sci., Vol. 25, No. 2, 1985   83

$$\hat{\beta} = (X'W_2W_3X)^{-1}(X'W_2W_3Y) \qquad (18)$$

The process is then continued iteratively until convergence (two successive estimates of β differ by less than 0.0001) is achieved. 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 0 and 1, inclusively. Thus, a value less than 1 for one or both of an observations' 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).

DATA SETS

As Taylor et al.[24] suggest, detailed information on industrial catalytic processes is likely to be regarded as a trade secret and thus be unavailable to the outside researcher. Thus, because an original complete set of industrial poisoning data was unavailable, the next best thing, a simulation, was performed. During World War I, the U.S. feared that its supply of nitrates from Chile might be cut off. Accordingly, the Bureau of Mines undertook extensive studies, and developed both a pilot plant and a full-scale plant for the oxidation of ammonia to nitric acid. These studies and their results were extensively described in the literature of the period,[18,22-24] including operational data, on which this simulation was based. Parsons[18] gave a detailed description of operational data of the full-scale plant, which produced nitric acid by using a platinum-gauze catalyst. Taylor and Capps[22,23] and Taylor, Capps, and Coolidge[24] have described the pilot studies (including poisoning data) for this project. With the data reported in these studies, percentage yields were selected by means of a random-number generator to form the simulated data sets given in Table I. The contaminants to the process were acetylene (impure), phosphine (both shown to be poisons), and hydrogen sulfide (a contaminant, but not a poison). The observations for days 1–5 were selected from the sequential daily data (uncontaminated) reported by Parsons.[18] The day 6 observations (contaminants) were selected from the data reported by Taylor and Capps.[22,23] Thus, the resulting simulated data sets (except 1N, 2N, and 3N) each contain five observations (days 1–5) selected at random from a similar 5 days of operation of the uncontaminated process, while the day 6 observation is chosen from the yields from the contaminated pilot plant studies. 1N, 2N, and 3N represent a control selected from the 5 days of uncontaminated operation.

COMPUTER PROGRAMS

Computer programs (written in Basic) to perform the GM algorithm have been implemented on both the DEC Vax 11-780 and the Radio Shack TRS-80 Model III. The TRS-80 version runs easily in 16K of RAM. Both versions are available from the authors.

RESULTS

The results of the GM procedure computations for the Table I data sets are given in Table II. We observe that in all of acetylene and phosphine cases (the poisons) the day 6 observations show an autoregressive residual weight of less than 1 and a decrease in percentage yield, thus indicating that there is an active poison in the reaction mixture. In the case of hydrogen sulfide (a contaminant, but not a poison), we observe only one autoregressive residual weight less than 1 (case 2S). This, however, reflects the fact that the 94.3 datum is greater than expected from the other observations, not less, and is not an indicator of poisoning. Such an observation can occur for many reasons, e.g., experimental error, change in process conditions, etc. In cases 1N, 2N, and 3N there is no evidence

Table II. GM Results

case	observation no.	autoregressive residual wt	location wt
1A	2	0.658 779	1.0
	3	1.0	0.722 283
	4	1.0	1.0
	5	1.0	1.0
	6	0.593 785	1.0
2A	2	1.0	1.0
	3	1.0	1.0
	4	1.0	1.0
	5	1.0	1.0
	6	0.940 031	1.0
3A	2	0.756 715	1.0
	3	1.0	0.781 908
	4	1.0	1.0
	5	1.0	1.0
	6	0.480 35	1.0
1P	2	0.658 731	1.0
	3	1.0	0.722 283
	4	1.0	1.0
	5	1.0	1.0
	6	0.335 761	1.0
2P	2	1.0	1.0
	3	1.0	1.0
	4	1.0	1.0
	5	1.0	1.0
	6	0.589 445	1.0
3P	2	0.756 715	1.0
	3	1.0	0.781 908
	4	1.0	1.0
	5	1.0	1.0
	6	0.306 89	1.0
1S	2	1.0	1.0
	3	1.0	1.0
	4	1.0	1.0
	5	0.910 691	1.0
	6	1.0	1.0
2S	2	1.0	1.0
	3	1.0	1.0
	4	1.0	1.0
	5	0.641 408	1.0
	6	0.897 707	0.648 015
3S	2	1.0	1.0
	3	1.0	0.908 677
	4	1.0	1.0
	5	1.0	0.908 677
	6	1.0	1.0
1N	2	0.482 205	1.0
	3	1.0	0.513 961
	4	1.0	0.770 941
	5	1.0	1.0
2N	2	1.0	1.0
	3	1.0	1.0
	4	1.0	1.0
	5	1.0	1.0
3N	2	0.740 013	1.0
	3	1.0	0.406 186
	4	1.0	0.520 204
	5	1.0	1.0

of poisoning on day 5. Notice that a point must be examined without addition of any later observations since later observations may affect the weights of previous points. It is possible to "tune" the procedure to different processes by adjusting the parameters in (11). This tuning will adjust for different yield variations that can be exhibited by different production processes. In addition, yield decrease is flagged by residual signs.[5]

Three other procedures for the identification of time-series outliers have been proposed in the recent past,[1,8,10] which could perhaps also be used in the present context. Two of these procedures, due to Fox[10] and Abraham and Box,[1] have recently been reviewed.[3] Both of these procedures are more difficult to apply than GM, though as Martin discusses[14] GM is to a degree a descendent of the Fox method. The Fox procedure itself consists of two separate tests for the two most common

outlier types, while GM does both at the same time. The Abraham and Box procedure is less likely to be useful than GM because it was designed to consider only one type of outlier.[1] The relationship between GM and the Chernick et al.[8] procedure is discussed elsewhere.[4,5]

CONCLUSIONS

On the basis of these results it can be concluded that the GM procedure is an effective method for the quick discovery of the effects of catalytic poisons, as long as care is taken to ensure that all model assumptions are met prior to drawing any conclusions about the process based on this method. The computational procedure may be implemented quickly on a laboratory microcomputer so that results become immediately available to a quality-control laboratory. Related procedures, involving quality-control charts, which may be useful in some circumstances for catalytic processes, are discussed in Booth and Isenhour.[6]

REFERENCES AND NOTES

(1) Abraham, B.; Box, G. E. P. "Bayesian Analysis of Some Outlier Problems in Time Series". *Biometrika* **1979**, *66*, 229–236.
(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, R. D. "Outlier..........s". *Technometrics* **1983**, *25*, 119–149.
(4) 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.
(5) Booth, D. E. "A Method for the Early Detection of Loss from a Nuclear Material Inventory". Submitted for publication in *J. Chem. Inf. Comput. Sci.*
(6) Booth, D. E.; Isenhour, T. L. "An Application of Robust Time Series Analysis to the Interpretation of Quality Control Charts". Submitted for publication in *J. Quality Technol.*
(7) Box, G. E. P.; Jenkins, G. M. Time Series Analysis: Forcasting and Control; Holden-Day: San Francisco, CA, 1976; revised edition, p 9.
(8) Chernick, M. R.; Downing, D. J.; Pike, D. H. "Detecting Outliers in Time Series Data". *J. Am. Stat. Assoc.* **1982**, *77*, 743–747.
(9) Denby, L.; Martin, R. D. "Robust Estimation of the First Order Autoregressive Parameter". *J. Am. Stat. Assoc.* **1979**, *74*, 140–146.
(10) Fox, A. J. "Outliers in Time Series". *J. R. Stat. Soc., Ser. B* **1972**, *34*, 340–363.
(11) Hogg, R. V. "Statistical Robustness: One View of Its Use in Applications Today". *Am. Stat.* **1979**, *33*, 108–115.
(12) Hull, C. H.; Nie, N. H. "SPSS Update 7-9"; McGraw-Hill: New York, 1981.
(13) Makridakis, S.; Wheelwright, S. "Forecasting Methods and Applications"; Wiley: New York, 1978; p 633.
(14) Martin, R. D. "Robust Methods for Time Series". In "Applied Time Series Analysis II"; Findley, D. F., Ed.; Academic Press, New York, 1981.
(15) 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.
(16) Montgomery, D.; Johnson, L. "Forecasting and Time Series Analysis"; McGraw-Hill: New York, 1976.
(17) Mosteller, F.; Tukey, J. "Data Analysis and Regression"; Addison-Wesley: New York, 1977.
(18) Parsons, C. L. "Commercial Oxidation of Ammonia to Nitric Acid". *J. Ind. Eng. Chem.* **1919**, *11*, 541–552.
(19) Pearce, R.; Patterson, W. R. "Catalysis and Chemical Processes"; Wiley: New York, 1981; p xv.
(20) Starkey, P. E. "Methods of Operating Catalytic Processes". In "Catalysis and Chemical Processes"; Pearce, R.; Patterson, W., Eds.; Wiley: New York, 1981.
(21) Stinson, S. C. "Catalysts: A Chemical Market Poised for Growth". *Chem. Eng. News* **1983**, *61*, 19–25.
(22) Taylor, G. B.; Capps, J. H. "Effect of Acetylene on Oxidation of Ammonia to Nitric Acid". *J. Ind. Eng. Chem.* **1918**, *10*, 457–459.
(23) Taylor, G. B.; Capps, J. H. "Effect of Phosphine and Hydrogen Sulfide on the Oxidation of Ammonia to Nitric Acid". *J. Ind. Eng. Chem.* **1919**, *11*, 27–28.
(24) Taylor, G. B.; Capps, J. H.; Coolidge, A. S. "The Production of Nitric Acid from Nitrogen Oxides". *J. Ind. Eng. Chem.* **1918**, *10*, 270–274.
(25) Twigg, M. V. "The Catalyst: Preparation, Properties, and Behaviour in Use". In "Catalysis and Chemical Processes"; Pearce, R.; Patterson, W. R., Eds.; Wiley: New York, 1981.

Method for Estimating the Human and Environmental Exposure Potential of Chemicals Having Designated Uses[†]

WENDY L. BYER

Information Ventures, Inc., Philadelphia, Pennsylvania 19102

This paper describes a methodology that was designed to assist the Environmental Protection Agency in setting priorities for regulating chemicals in accordance with requirements of the Toxic Substances Control Act by indicating the extent of exposure potentially associated with a particular use of a chemical. Development and application of the methodology to actual chemical uses proceeded in four stages: (1) development of an exposure classification scheme, (2) development of a chemical use list, (3) assignment of exposure category scores to chemical uses, and (4) merging of exposure scores into ranked groups, creating the Index of Exposure. The purpose of the Index was to identify those chemical uses having the greatest potential for human and environmental exposure. Theoretically, the objectivity of the scoring methodology is such that, given the same data for a chemical use, different scorers will be able to arrive at the same exposure score for that use.

INTRODUCTION

The Toxic Substances Control Act (TSCA), which became effective on January 1, 1977, makes the Environmental Protection Agency (EPA) responsible for regulating chemical substances "whose manufacture, processing, distribution in

commerce, use or disposal may present an unreasonable risk of injury to health or the environment."[1] Under TSCA, manufacturers, processors, and importers may be required to submit certain data to EPA prior to the commercial production, processing or distribution of a new chemical or of a previously produced chemical intended for a new use. EPA is also authorized to collect the same data for existing chemicals in existing uses. Such data may include the chemical's molecular structure, its intended use, estimated production volume for each intended use, the byproducts resulting from