PREANALYSIS OF SUPERLARGE INDUSTRIAL DATASETS

DAVID L. BANKS
Department of Statistics, Carnegie Mellon University

GIOVANNI PARMIGIANI
Institute of Statistics and Decision Sciences, Duke University

DP # 91-A20
Preanalysis of Superlarge Industrial Datasets

by

David L. Banks, Department of Statistics
Carnegie Mellon University, Pittsburgh, PA 15213

Giovanni Parmigiani, Institute of Statistics and Decision Sciences
Duke University, Durham, NC 27706

Abstract: Successful analysis of superlarge datasets requires statistical procedures that automatically clean the data and uncover simple structure. The protocol we describe applies to multivariate industrial data from continuous manufacturing processes with feedback and feedforward control. Our methods form an twelve-step sequence that edits and relags the time series, as well as applying diagnostics to look for subtle data flaws. At different stages, the protocol will reject data, impute data, relag the time series, flag categories of suspicious data, and divide the dataset into more homogeneous subsets. The output is a cleaned dataset for analysis with standard statistical packages or software tools. Although there is no guarantee that every corruption has been caught and corrected, the output dataset is more thoroughly examined than traditional human-intensive methods can achieve. To assist in this preliminary analysis, we describe four graphical methods developed in studies of glass manufacture from PPG Industries’ production plants and sheet aluminium production by Alcoa.

Keywords: Imputation, Multivariate Time Series, Graphics, Data Analysis
0.1 Introduction

Our title requires two definitions. First, the term preanalysis refers to all data manipulation tasks needed to prepare a dataset for interrogation by standard statistical methods. This includes tests for validity, internal consistency, and outliers, imputation of missing or unbelievable data, and preliminary identification of local structure. Second, the term superlarge refers to datasets whose size makes preanalysis through human scrutiny impractical. Both of definitions are contextual. For example, the tasks in the preanalysis partly depend upon the model, and eventually blend into conventional investigation. Also, the size at which a dataset becomes superlarge varies according to the human resources committed to the analysis. Nonetheless, many important datasets are demonstrably superlarge and require extensive preanalysis before substantive questions can be addressed.

Modern statistical process control often involves automatic collection of superlarge datasets. For example, in the production of glass or aluminium, hourly measurements are taken on hundreds of control and quality variables for months at a time. Some of these data help maintain product quality, but much is collected in the hope that exploratory analysis will find unexpected structure which can inform future efforts to optimize the process.

Superlarge datasets present special problems. Inevitably, the data must be cleaned before statistical tools can be applied. (Our experience with datasets from Alcoa and PPG Industries indicates that even under completely automated data collection and storage, problems arise that demand special handling.) However, human-intensive data cleaning makes heavy demands for statistical training, process knowledge, and inspection of many plots and diagnostics. At the very least, traditional preanalysis requires days of expert
attention, and even then can fail to detect critical flaws.

We propose a protocol for cleaning industrial data. This takes a superlarge dataset, runs through a twelve-step process that makes minimal demands on human expertise, and then produces a corrected dataset which can immediately be examined by statistical packages or in-house software programs. The process should be tailored to the application, but overall, the steps are pertinent to a range of industrial problems. This protocol systemizes the handling of a complex problem, and concentrates human attention at the crucial points.

We also develop a suite of graphical tools to help in data cleaning and to highlight process structure. These tools apply the principle that process improvement emerges from an understanding of the extreme behavior of the production history. The value of the protocol and these graphics is illustrated with a dataset obtained from a PPG flat glass manufacturing plant in Fresno, California.

Section 2 examines general problems in precleaning superlarge datasets and introduces the PPG data. Section 3 discusses the twelve-step protocol, and graphical tools that guide its application. The points are illustrated with examples from the PPG analysis. Section 4 provides additional graphics, and shows how they led to sharper formulation of the analysis of a complex industrial datasets. It also discusses some heuristics we have learned in implementing analyses of this kind.

### 0.2 Automated Data Management in Industry

Many organizations grapple with special problems posed by superlarge datasets. The Census Bureau, the IRS, and LANDSAT project workers routinely address research issues
in automated data management. However, their accumulated experience does not bear
directly on the management of continuous manufacturing processes. Such industries gener-
ate multivariate time series, with hundreds of control variables and dozens of response (or
quality) variables. This contrasts with the LANDSAT project and government agencies, in
which observations are not sequential and no control-response mechanism exists.

Insofar as possible, our work builds on previous research with systems for automatic
data cleaning. These efforts, chiefly undertaken by the Census Bureau, have identified
three major issues: (1) data checks for inconsistent values, (2) imputation for missing data,
and (3) outlier detection and adjustment. All three problems arise in industrial datasets,
but previous solutions are difficult to extend.

Automatic data checking uses software to examine records for inconsistent values; e.g.,
a census respondent who indicates that they have five children, three of each sex. Typically,
this software ascertains that all fields which should be present for a certain record are
supplied, that the values supplied are valid, and that all values are mutually consistent.
Examples of such software systems are CONCOR (see Bair, 1981), a program developed by
the International Statistical Program Center at the U.S. Bureau of Census, and SPEER,
tailored to Census surveys of industries (see Greenberg, 1981).

The second problem, data imputation, has attracted broad attention. Kalton and
Kasprzyk (1986) survey the literature, applications and theoretical issues. The strategy
is to use information from complete data records to assign probable values to incomplete
records. For census data, imputed values are usually determined by some variant of the
hot-deck method, in which partial data from an incomplete record are matched to cor-
responding data from a complete record or records, and missing values are then inferred
from the complete data records. Alternative approaches use conditional means or regression
models to impute values. When the size of the dataset is moderate, one can use the EM
algorithm, described in Little and Rubin (1987), to iteratively estimate missing values and
model parameters.

The third problem concerns outlier identification. If there is no time series structure
and dimensionality is low, then graphical methods are popular. For census data and other
applications with high dimensionality, one can use Mahalanobis distance, or similar methods
described in Hawkins (1980). These methods assume multivariate normality, and do not
extend to categorical or ordinal data. A related problem is to distinguish an outlying
record from an outlying value within a fundamentally correct record. Little and Smith
(1987) survey issues in outlier detection for data without time series structure.

Our applications arise from continuous manufacturing processes, and hence have a time
series character. In this context, little pertinent work on outlier detection has been done.
For univariate time series, Chernick, Downing and Pike (1982) propose a method based on
the use of the autocorrelation’s influence function. Khattree and Naik (1987) extend the
influence function approach to bivariate time series, but the method becomes intractable
with increasing dimensionality. Some analogous problems in time series relate to identifying
instantaneous shifts in the process, and can be treated using the technique of intervention
analysis developed by Box and Tiao (1975) and Tsay (1988).

Similar difficulties arise in extending automatic data checks and imputation methods to
multivariate applications that arise in industry. Data cross-checks for internal consistency
increase combinatorially with dimensionality; if one takes account of serial correlation in
the observations, then computing demands are even more intensive. Nonetheless, if a data
vector is in wild disagreement with its immediate predecessor and/or successor, then the logic of consistency checking requires one to flag that record. To compound the problem, record-matching imputation becomes impractical in high dimensions, since virtually every record is distinct.

All these issues, and some others, arise in a dataset collected by PPG Industries during flat glass manufacture at a plant in Fresno, California. The production technology used a refinement of the Pilkington float glass method, in which “logs” of raw material (chiefly silicates, calcium carbonate, sodium carbonate, scrap glass) are melted, the melt is then mixed in a refiner (or tank), and then continuously poured onto a float bath of molten tin. After sufficient cooling, a ribbon of hot glass channels into the annealing lehr, where final controlled cooling occurs (see Tooley (1984) for an account of modern float glass technology). The time between the entry of a piece of material and its output as finished glass is about 48 hours.

Tables 1 and 2 describe 54 control variables and 20 response variables given by PPG for analysis to the Statistical Center for Quality Improvement, a research consortium run by the University of Pittsburgh and Carnegie Mellon University. Most variables were measured hourly, but some were measured once per shift and some were measured each half-hour; variable 52 was measured less systematically, depending in part upon when glass thickness was deliberately changed. The third column of Table 1 indicates the approximate time, in shifts, between the action of the control variable and the effect on the output (this lag structure is imprecise, because of mixing within the tank and variation in glass thickness). The data were collected over a period of 20 days, beginning on November 4, 1989. As superlarge datasets go, this one is quite small.
The variable descriptions in Table 1 are incomplete; our purpose is to propose general methodology, which does not require deep exegesis of Pilkington float-bath technology. The variables in Table 1 pertain primarily to furnace characteristics; in practice, PPG measures many variables not listed here. The data were further reduced for proprietary reasons, and some variables were excluded by process engineers on other grounds.

Not all of the control variables are directly manipulated, and not all response variables are related to product quality. However, for obtaining a better understanding of process characteristics, PPG feels there is value in the entire dataset. For example, variables 3 to 28 concern observational data on the melting dynamics of the log in the tank. These provide information on the location of different portions of the log in the furnace; the ports divide the furnace into regions, and some regions have an irregular geometry that is called a Dog House. The shape of the log carries information about internal heating currents in the melt and the mixing rate. As another example of a variable that is not directly controlled, in a related study PPG recorded the ambient temperature at the beginning of the lehr phase.

Table 2 lists 20 response or quality variables. These also form a multivariate time series, and are related to the control variable values at the lags indicated in Table 1. Variable 65 represents the total number of defects, standardized to account for changes in glass thickness; it offers a single summary of process quality. Other variables enable engineers to examine quality deficiencies more specifically, with the hope of relating defect types to process characteristics.
Time series structure in the PPG quality data is strong, since the product is a continuous ribbon of glass, and since manufacture involves a mixing step, with some feedback and feedforward control loops. Most control variables also show strong serial correlation. Some measurements are captured automatically, while others are entered by human operators (variables 3 to 28, and most of the quality variables, are recorded directly by inspectors).

0.3 Protocol for Automated Data Cleaning

Our data cleaning process begins with a datafile that is in machine-readable form. We expect flaws, but assume reasonable care has been taken to format the data and identify which values are tied to which process variables. The initial data structure is that of a highly multivariate time series with a complex lag structure, missing values and potentially large errors. The output data structure is relagged (synchronized) according to known processing times, and has been cleaned to reduce errors and flag issues that may trouble the analysis.

Our protocol is intended as a guideline for the cleaning and initial analysis of data that arise from complex continuous manufacturing processes. In any specific application, one expects to tailor the protocol to the problem. With this caveat, the key steps are:

1. Data Input. Reformat the data; all numerical input is read into a common floating point representation. All non-numerical input is represented consistently. The size of the representations is determined by the largest possible (signed) value.

2. Time Scale. Create or identify the time scale. All other observations are then keyed to the value of the time variable. The data record thus becomes a set of vectors indexed by time, or a multivariate time series; each vector contains the measurements
on all control or response variables at a given time point. If variables are not measured simultaneously, unmeasured values are coded to indicate that they are missing as part of normal process operation.

3. *Missing Values.* All data values known to be missing are assigned special codes. These indicate whether the data is missing due to plant shutdown, failure of a monitoring device, unknown causes or routine operation of the process inspection schedule.

4. *Sample Size Check.* Verify the numbers of observations for each variable. The input is the expected length of each time series and the expected length of each data record, accounting for known missing values. The output is a list of variables for which the expected and actual sizes do not match.

5. *Impossible Values.* List and adjust all observations with impossible values. The determination of allowable values requires guidance from a process engineer. With this list, the possible actions are: verify, approximate, code as missing data, or correct for a missing sign, zero or data-entry transposition.

6. *Synchronicity.* Reindex the database so that values are synchronous. This ensures that measurements keyed to a specific time actually pertain to common process characteristics. With unsynchronized data, control measurements recorded at a given time do not affect quality measurements until the end of the production cycle; for example, in the Fresno data, a change in variable 38 does not affect glass quality until 16 hours (2 shifts) later. To properly assess the interplay between process control and output quality, the analysis must first link corresponding values into common records.
7. *Missing Value Chart.* Produce a chronological guide to missing values. Often it is useful to subtract out known interruptions, which requires information from the process engineer regarding plant or subsystem shutdowns.

8. *Unequal Frequencies and Data Imputation.* Use data imputation procedures, smoothing, or expert knowledge to fill short gaps in the data sequence. In most cases, the missing data arise because not all variables are recorded at the same frequency. In others, gaps are caused by sensor failure or other extraneous circumstances. Sometimes it happens that the short gaps are informative, especially when the values are probably near the boundary of the historical operating region; these require expert judgment.

9. *Extreme Value Chart.* Generate a chart showing the extremes in the variation of process variables. This extreme value chart helps to flag outliers in step 10, identify data flaws that have not been previously detected, and point up patterns of variation that require examination. The chart may be viewed as a global Shewhart chart, and thus also has value for day-to-day process management.

10. *Outlier Detection.* The extreme value chart has discovered univariate outliers. These should be checked and reaffirmed, or else replaced by an estimate. Additionally, one should search for multivariate and possibly time series outliers.

11. *Descriptive Statistics.* Test the time series for nonstationarity, and examine model consistency between plant shutdowns or other natural process interruptions. Do Q-Q plots to assess departures from normality, and perhaps make normalizing transformations of the data.
12. *Exploratory Analysis.* This begins to phase into the conventional analysis. We find that some exploration is needed to plan subsequent work, and the specific methods used are sensitive to the context of the study. Rather than describing such generality, Section 4 gives examples of tools we have used in previous applications.

As a final step, we urge that one rethink all that’s been done. A process engineer should review the adjustments and ratify the cleaning process. This sounds trivial, but our experience indicates that it is inevitably useful.

Many of these twelve points require some additional discussion. For example, the first step sounds trivial. However, consistent formatting makes all subsequent analyses easier to implement; in particular, one can move rapidly between different packages and special purpose software. Also, important flaws can be detected in this stage; for example, in a recent PPG datafile, a particular variable field did not allocate enough space for large numbers, and so adjacent values were either read as a single number (free format input), or else negative numbers became positive (column input). It took two days to pinpoint the error, and we now use a simple program that automatically searches for such flaws. Similarly, missing values often arise when changes in the data collection methods are not properly reflected in the data format. Finally, if one inspects the records visually, a regular format magnifies the ability of the human eye to discover errors.

Steps 1, 2, 3, 5, 6, 8, and 10 all create a new database from the preceding database. These new databases refine the original by reformatting, indicating the kind of missing values that occur, relagging the database time index to enable meaningful comparisons, and correcting discovered deficiencies. The result is a pseudo-database which approximates the dataset one would have liked to have been given for study in the first place.
In Step 3, the most common cause of missing data is differential measurement frequency. For example, in an Alcoa dataset some variables were measured every half-hour, whereas others were measured each hour. Thus the time index changes in half-hour increments, and some variables show missing values in every other observation. This kind of gap is viewed as missing as part of the normal operation of the process. Similarly, planned interruptions should be coded to distinguish them from failures in the data capture effort.

Step 7 produces the missing value chart, a graphic output that aids process management and informs the data cleaning process. It displays patterns of omission by category of omission. Specifically, the chart

- ignores data that is absent as part of the normal operation of the process.
- shows the number of observations missing at a given time period.
- shows the number of consecutive time periods with missing values.
- flags time periods in which certain subsets of missing values occur, such as those pertaining to specific subsystems in the production process.
- flags time periods according to known causes of missing data, such as plant shutdowns or sensor failure.

In specific applications, other kinds of missing data may occur. For example, data may be missing because the values have fallen outside the range of sensitivity of the measuring instrument.

An example of this chart is shown in Figure 1. The chart is artificial, since we want to display a range of missing value behavior in a single table over a short time span.
However, the types of behavior depicted have all occurred in datasets from PPG and Alcoa. The vertical axis represents 16 generic variables (the numbering does not refer to Tables 1 or 2); more variables could be included, by compressing the vertical scale. It is usually convenient to segregate the control and quality variables, and to consecutively list control variables in the same subsystem. The horizontal axis represents time.

(Insert Figure 1 about here.)

This graphic displays missing value patterns by icon patterns. Dots indicate missing values attributable to planned causes, such as plant shutdown or subsystem shutdown; exclamation points indicate data missing for unknown reasons; vertical strokes represent data missing for unplanned reasons, but for which the cause is known. Additional icons might be necessary in some applications.

Features highlighted by Figure 1 include (a) a plant shutdown late in the epoch, (b) a regular pattern of missing data in variable 14 (when a similar pattern was found in the PPG data, subsequent inquiry discovered that one of the shift foremen had misunderstood the reporting requirements), (c) a sustained subsystem shutdown in variables 4, 5, 7 and 8, and (d) a brief subsystem failure, for unknown reasons, in variables 1, 2 and 3 (when a similar pattern was found in Alcoa data, subsequent inquiry discovered that one measurement was missing due to sensor failure, and its output was needed in calculating the other missing values). Also, there are a sprinkling of other missing values with unknown causes. The plant manager could use this information to focus improvement in process monitoring.

In realistically large applications, the time span is too long and the numbers of variables too great to be represented as above. Instead, one should use a color monitor; white pixels
mark normal values, colored pixels code the categories of missing data. This compact display accommodates about 750 variables and 1000 time points on a standard workstation. For greater time spans one can scroll the screen, divide the time span into intervals or collapse intervals without missing data.

Step 8 corrects the data base for missing values that arise from normal process operation, and, perhaps, for the sprinkling of isolated unpatterned missing values that occur for unknown reasons. In our work with Alcoa and PPG, most adjustment accounts for differential sampling rates. There are sophisticated strategies for interpolating or imputing such missing values. If one has knowledge of the underlying time series model, or if the process history permits tight record matching, then one could use either a form of forecasting/backcasting or hot deck imputation, respectively. However, our experience indicates that situations with this degree of structure are rare, so we prefer to estimate values by linear interpolation of the data.

The advantages of linear interpolation, as opposed to the more sophisticated smoothing routines described by Friedman, Grosse and Stuetzle (1983), are both computational and theoretical. First, linear interpolation is easy to program and quick to implement, in almost any software system that might be used. Second, linear interpolation makes a strictly local adjustment of the data; distant values have no influence on the correction. This seems a desirable feature when cleaning data without a well-defined probability model. Even when there is information or understanding that would enable slight improvement over simple linear interpolation, it may not be cost-effective to pursue such refinement; i.e., if one's ultimate inference from the mountain of available data is sensitive to the rule for imputing missing values, then the entire solution is surely unstable. Spending effort on optimal
imputation takes resources for other phases of the analysis, making this a case of the best
being the enemy of the good.

Whatever imputation method one uses, the output from Step 8 is a new, artificial
database that spans data gaps

- to remove very short sequences of missing data whose absence is deemed random and
  uninformative.

- to obviate difficulties in subsequent analysis caused by data missing as part of the
  usual operation of the process.

The first adjustment is appropriate when clerical error or sensor failure causes missing
data, but the process is under control and nearby values are likely to be similar to those
that are missing. The second is useful when not all observations are recorded at the same
sampling rate; here one can interpolate the least frequent (data inflation) or average the
most frequent (data reduction). If one variable is measured daily and another is measured
hourly, we recommend interpolating the daily values to create hourly proxies, since the
alternative, compressing the hourly data to daily averages, destroys information.

A caution is needed. When large amounts of missing data are estimated, the variance
of the values in the artificial database tends to be too small. Thus the outlier tests in Step
10 will be oversensitive, and may require some tuning.

Step 9 produces an extreme value chart; in analogy with the missing value chart in
Figure 1, it looks for patterns over time in the least and largest values of the control
and quality variables. Figure 2 gives an example obtained from the PPG data shown in
Tables 1 and 2. The vertical axis indicates the identification number of the corresponding
control and quality variables; in many applications, it is useful to group these according to process sequence or process subsystems. The horizontal axis is the time index of the data, with 32756 denoting November 4, 1989, and 32776 denoting November 24, 1989 (the time scale reflects a PPG plant coding system). For a given variable at a given time, we plot blank space if the datum is within ±2.5σ of the mean for that control or quality variable. Otherwise, we plot a stroke if the the datum exceeds the upper 2.5σ bound, and a dot if it falls below the lower 2.5σ bound.

(Insert Figure 2 about here.)

In our experience, a ±2.5σ rule captures data oddities without overloading the visual display; for other applications, the customary ±3σ zone might serve better. For datasets with very large numbers of variables, it is more compact to code the direction of extreme deviations with color pixels, and display very large numbers of variables simultaneously upon a color monitor.

From the management perspective, Figure 2 clearly shows how the process drifted out of control at about 32767, and how the process engineers then “chased” the process. Even at the end of the time span covered by this dataset, it is not certain that stability has returned.

More generally, Figure 2 shows both isolated data points and runs of data (often affecting downstream variables) that fall outside the zone of normal operating values. The former look like potentially influential data flaws, and should be checked to ensure that no coding glitch has occurred; even if no identifiable cause can be found, it may improve the analysis to replace these outliers by interpolated values. In contrast, runs of extreme values carry a
great deal of information about process capability beyond the usual operating region. These should be preserved in the database, and closely studied in the subsequent analysis phase.

The extreme value chart enables managers to identify which variables are prone to fall out of control, and which quality variables are sensitive to such departures. These charts can also discover faulty control loops or differences in performance between operating shifts. For some applications, it merits note that the extreme value chart responds to planned changes in control setpoints, so one should calculate control means only over intervals between process innovations.

Step 9 has produced a chart that guides Step 10. Here one identifies outliers, and remedies these in the way most appropriate to the problem. One should verify their values, and then either replace them by less extreme estimates (i.e., treat them as missing values), or else prepare a robust analysis. To increase the complexity, there is also concern about multivariate outliers and outliers with respect to the time series structure.

Multivariate outliers occur when two or more variables take values that are jointly improbable. Hawkins (1982) describes methods of multivariate outlier detection. We feel that simple Mahalanobis distance is broadly sensitive and often adequate. The Mahalanobis distance for the total process and the process subsystems may be included as rows in the extreme value chart. If such outliers are found in those rows, they should be checked by process experts, and either retained or replaced by interpolation.

Univariate and multivariate time series outlier detection is difficult, but potentially informative. Univariate time series outliers can often be recognized as runs of outside values on the extreme value chart, and multivariate outliers can be detected as patterns of such runs. No useful theory is available for the multivariate case. From an exploratory
standpoint, inspection of the extreme value chart may be the most practical way to identify multivariate outlier behavior; in our application this showed up visually as the process engineers chased the drifting system.

Step 11 is conventional methodology; too much detail sidetracks the purpose of this paper. In brief, Step 11 moves the preanalysis beyond data cleaning and lays the foundation for addressing the substantive questions. One has finally constructed an artificial database that resembles what classroom lectures assume one receives in the first place, and now it becomes appropriate to begin applying textbook tools. For example, if one contemplates an analysis based upon normal theory methodology, such as fitting a linear model or extracting the principal components, then it is proper to undertake Q-Q plots and look for normalizing transformations.

Step 12 moves the preanalysis firmly towards conventional analysis. Although methods depend on the application, we have found it useful to build two types of charts, both aimed at detecting data structure in the region of optimal operation. These tools provide useful guidance in planning an analysis when there is little physical understanding of the process model. Details on these charts, and our experience in implementing them, are described in the following section.

In reviewing the protocol, it may seem that the burden of human scrutiny has hardly shifted. We have proposed the regular examination of two charts, and will soon describe two more. Nonetheless, compared with current practice, this is a major reduction in the amount of time on the preanalysis. One does not have to look at each variable separately, again and again, making corrections and seeking connections. Instead, the charts and the protocol focus the analyst’s attention where it is needed, while delegating support tasks to
software and assistants.

As a final point, we emphasize that the 12-step protocol is a guideline. In a formal sense, the preanalysis is part of the total analysis; if one has a good model, then it should drive one’s decisions during the preanalysis phase. Since well-defined models are rare for complex processes, it usually happens that the chief concern is to be sure that the preanalysis does not entail any alterations of the data that will hinder subsequent examination of central research questions.

0.4 Implementation Experience

This section describes our experience with four superlarge datasets; three were from PPG Industries, one was from Alcoa. Three datasets are confidential; the fourth is the PPG data from Fresno, described in Tables 1 and 2. We are permitted to relate the methodological insights gleaned from the confidential datasets, but cannot present numerical results or identify specific variables. All four datasets were analyzed by the Statistical Center for Quality Improvement (SCQI).

The structures of the processes that created the four datasets are remarkably similar, conforming to a general paradigm for continuous and batch-continuous manufacturing. Besides glass and aluminium, this pattern includes chemical and petroleum production, and many other cases. Some common features are:

1. very large numbers of control and quality variables, measured over long periods of time,

2. process maturity, meaning that catastrophic disturbances are rare, considerable ex-
pertise with the process is available, and the historical record is concentrated in a small region of the space of possible controls.

3. process subtlety, meaning that engineers understand the most important main effects, so that competitive advantage is likely to accrue from exploitation of interactions,

4. process instability, so that constant control is needed to prevent production from drifting into poor quality,

5. management interest in the physical model focuses on behavior in the region of the control space that produces top quality output,

6. feedback and (more rarely) feedforward control loops, tending to enforce stationarity on the observed time series,

7. quality measurements taken at time \( t \) correspond to earlier control variables, introducing a lag structure,

8. serial correlation in all or most of the variables, usually from mixing, thermal inertia and other forms of process momentum.

All of these factors had to be taken into account during the preanalysis phase.

Both companies wanted to use SAS and Fortran programs in examining their historical data. Also, the SCQI was interested in comparing MARS (cf. Friedman (1991)), AVAS (cf. Tibshirani (1988)), and certain multivariate time series routines for the response surface research. Therefore the preanalysis aimed at enabling diverse software applications. Also, the preanalysis wanted to discover simple patterns of covariation among the variables, to facilitate more focused investigation.
Although the data were delivered on tape, e-mail and floppy disks, all contained various formatting problems. Data values were run together, or omitted signs, or some identical categorical data had multiple labels, and so forth. At first, untangling this was largely accomplished through global search-and-replace commands, but with later datasets we found it time-effective to write special purpose programs for editing. Free-format input and SAS commands enabled convenient reformatting.

Subsequent steps in the cleaning protocol were implemented substantially as described in Section 3. Not all datasets required all steps in the protocol (usually because industry statisticians had already addressed them), but it offered a valuable checklist in organizing discussions with area experts. Most of the steps were automated by means of ad hoc programs, or took advantage of already existing software. Other steps required statistical judgement and expert opinion, especially when identifying outliers, coding missing values, deciding upon the lag structure and finding and correcting unbelievable values. These steps typically entailed joint work with the process engineers.

A few departures from the protocol are worth pointing out. First, Steps 3 and 4 were repeated several times in the cleaning effort. Some of the outliers discovered late in the protocol were recoded into a missing value category, and on two occasions a process engineer realized that the original missing value codes could be extended more informatively. Second, the synchronization in Step 6 was often approximate; several phases in the production involved local mixing, making precise determination of the lag of the downstream effect imprecise, and subject to expert judgment. In future work, a Fourier decomposition might identify this structure more precisely. Finally, one PPG dataset turned out to have two major process interventions, and the outlier detection methods in Steps 9 and 10 had to be
performed separately in each of the three inter-interruption epochs.

Regarding the transition from preanalysis to preliminary analysis, the industries had asked us to examine time series models and multiple regression. Our analysis indicated that conventional time series modelling was inadequate to process description. Grossly different models were needed to fit different time spans of the same variable. Process experts found none of the time series models to be plausible, workable or even suggestive. Similarly, although both industries had made heavy use of regression in developing an understanding of historical data, it turned out that the predictive power of these, as assessed with cross-validation from the source data, was regrettably small (even after Step 11 had found normalizing and variance stabilizing transformations).

Two methods that virtually all of the industry engineers deemed more valuable in setting the stage for a detailed, process-specific analysis were

1. producing an process covariation chart,

2. producing an SIB (Sliced Inverse Boxplot) chart.

Both methods rely on the intuition that there are many ways that a complex industrial process can go wrong, but only one way to make high quality product. Rather than spend attention and degrees of freedom in quantifying poor performance, it is more reasonable, and more locally linear, to build models for the region of high quality output.

The process covariation chart highlights local main effects and two-way interactions in the region of highest quality output. Also, it looks for antagonisms and protagonisms among the quality variables. The underlying strategy for the chart is similar to the motivation behind sliced inverse regression, developed by Li (1991) and Duan and Li (1991).

23
Figure 3 shows a process covariation chart for the Fresno PPG plate glass plant. The numerical values 1 through 6 on the side correspond to quality variables 59 through 64, respectively, as described in Table 2. Similarly, the alphabetical values a through n correspond to control variables 17 through 28, respectively, followed by 38 and 42; each is described in Table 1 (38 and 42 are out of sequence because we wanted to include them in this example; subsequent SIB chart analyses show interesting features). The quality variables selected for Figure 3 reflect the interests of PPG scientists; the control variables were chosen by building the matrix for sets of 14 control variables taken in consecutive order (omitting the first two variables, which are the date and shift), and then choosing the matrix that offered an range of detail for interpretation.

(Insert Figure 3 about here.)

To read Figure 3, start with the $6 \times 6$ upper lefthand block. This shows the interactions among the 6 quality variables. The sparser the hatching, the greater the evidence that the quality variables are antagonistic. Black squares indicate near perfect antagonism—in 90% or more of the data vectors in which the quality variable in the row scored in the best proportion $p$ (here $p = .1$), the quality variable in the column was also in the best $p$ proportion (diagonals are necessarily black). When squares in this upper lefthand block are black or densely hatched, one expects that it is possible to control the process so as to simultaneously achieve good levels of quality with respect to both variables. In contrast, when a square is white or lightly hatched, this indicates antagonism between the quality objectives. For a white square, less than .1% of the vectors have both row and column quality values in the top $p$ proportion, suggesting it is difficult to simultaneously satisfy
both quality goals.

We chose \( p = .1 \) after discussion with PPG experts. The hatching scale codes the percentage of data in which extreme values of the two variables appeared together in the same vector: black indicates more than 90\%, heavy hatching indicates 20\% to 90\%, moderate hatching indicates .5\% to 20\%, light hatching indicates .1\% to .5\%; white indicates less than .1\%. These cutpoints were chosen to be visually symmetric about independence of the two variables, in which case one expects 1\% of the best values to co-occur. Perfect symmetry would require information on the marginal distributions of each of the variables, and is unnecessary for our purposes.

Reading the upper lefthand block, we see that these quality variables are generally very strongly protagonistic. A process engineer might want to examine in greater detail a relatively weak protagonism that appears between low small bubble counts (59, represented in the matrix by 1) and low large bubble counts (60, represented by 2); perhaps bubbles in a section of glass tend to be counted as one kind or the other, so that positive large bubble counts force small bubble counts to be zero, and conversely. Similar study might explain the other weak protagonisms, seen between variables 63 and 64 (coded as 5 and 6, respectively) and between 64 and 62 (coded as 6 and 4, respectively).

The lower lefthand and upper righthand rectangular blocks look for main effects between quality and control variables in the region of high quality production. The lower lefthand block pertains to control values falling in the largest proportion \( p \) of the historical values; the upper righthand block pertains to control values in the lowest proportion \( p \). We do not know in advance whether either extreme is associated with good quality product.

As before, the hatching density indicates how often the best values of a given quality vari-

25
able co-occur with extreme values of the control variable; the hatching scale is unchanged. If both upper righthand and lower lefthand blocks are moderately to lightly hatched, this suggests that the setpoints for the control values in the process are near their optimal locations. However, if some of the squares are densely hatched, then manipulating those control variables may lead to improvement in the corresponding control variables.

For example, in the upper righthand block, the black squares indicate that good values of ream knots, stones and top drip (61, 62 and 63, coded as 3, 4 and 5, respectively) tend to co-occur with low values of the total bubbler flow (42, coded as n); specifically, in at least 90% of the cases in which each of these quality measures is in the best 10% of its values, the total bubbler flow is in the lowest 10% of its operating region. Similarly, for the lower lefthand block, the black squares indicate that good values of ream knots, stones and top drip co-occur with high values of the left combustion air flow (38, coded as m).

Finally, the 14×14 lower righthand block looks for potential two-way interactions among the control values in the region of best quality product. The diagonal squares in this block are meaningless. The subdiagonal squares code the number of times that the largest values (those in the upper deciles) of two control variables co-occur among the best 10% of a selected quality variable (this requires the analyst to focus upon a single quality measure of primary interest). Similarly, the superdiagonal squares code the number of times that the least values (those in the smallest deciles) of two control variables co-occur among the best 10% of a selected quality variable. The hatching scale is modified to reflect the fact that under independence, the expected percentage of such co-occurrences is only .1%; the ad hoc adjustment is to multiply the percentages by 10, so that the significance of a particular hatching is approximately commensurate with the other blocks in the matrix. In
particular, black or densely hatched squares suggest interactions that the process manager may be able to exploit. Light hatching is less informative, because the emphasis is on product improvement.

In Figure 3, the lower righthand block uses ream knots (61) as the selected quality variable. This reflects the immediate concern of the PPG engineers who brought us the problem. One sees that there are potentially useful interactions between the #5 bottom at the #1 doghouse at 3 and 6 shift offsets (variables 17 and 25, coded as $a$ and $i$) and between the #8 Bottom in the downtank Refiner and the #1 bottom between the End Wall and #1 port (variables 20 and 21, coded as $d$ and $e$). Process experts suggested that the first reflects good product obtained from stable process control, while the second interaction captures an energy differential that drives a thermal plume in the refining tank which helps control the mixing rate.

The lower righthand block may be extended in two ways. One way is to show another block, which codes the number of co-occurrences of large values for one variable with small values of the other control variable, all among the best quality product. The other extension produces a lower righthand block for each of the quality variables. This depth of detail goes beyond the intent of this section, but the industries whose data we analyzed often wanted to consider issues that required aspects additional depth of examination.

The process covariation chart is a tool for finding antagonistic quality goals and detecting possible univariate and bivariate improvements in process setpoints. By reducing the size of the squares (probably one should use red and green color rather than gray scales, to retain visual impact) one can simultaneously represent very large numbers of control and quality variables. The greatest drawback of the chart is that it ignores time series structure; in
some applications, this causes it to exaggerate the evidence in the data. On the other hand, it offers an easy way to ignore complex structure in regions of inferior glass, while looking for simple structure in the control region giving best quality output.

The SIB chart has an affinity to the sliced inverse regression work of Li (1991) and Duan and Li (1991) and to the regression box-plots of Tukey (1977, p. 100). It offers a closer scrutiny of variables highlighted in the process covariation chart, and also enables rough assessment of the influence of serial correlation upon one's conclusions.

Figure 4 shows an example of a SIB chart; it consists of two graphics. The upper graphic shows ten standard box-and-whisker diagrams; these display an approximate inverse regression based on the deciles of the quality variable. The lower graphic contains time series plots for the standardized values of the indicated control and quality variables (the quality variable plot is shifted up 3 standard deviations for better comparison). Figure 4 examines the black square in the upper right rectangle of Figure 3, which indicated that low values for the left combustion air flow (38) were associated with low (i.e., good) values of ream knots (61). The appearance of the SIB chart reinforces the indications that this connection is potentially useful for product improvement.

(Insert Figure 4 about here.)

To read the SIB chart, look first at the set of 10 box-and-whisker diagrams. The leftmost diagram is a box-and-whisker rendering of all values of control variable 38 that correspond to the best 10% of quality variable 61. The adjacent diagram is a box-and-whisker summary of the values of the control variable that correspond to quality levels in the second best 10%. This continues, so that the rightmost diagram reflects the control values for the worst 10%
of variable 61. As one surveys the general trend in the median and spread of the diagrams, there is confirmation for the idea that lower values for 38 are associated with improvement in the values of 61.

Now examine the time series chart beneath. The plot is a straightforward overlay of sketches of the control variable (light) and quality variable (dark) against time; to facilitate comparison, both are standardized and the mean of the quality variable is shifted up by three units. Close inspection shows that the change over time in variable 38 seems to follow rather than precede change in quality variable 61. This may undermine the case for making adjustments to variable 38 in order to control the ream knot defect rate; however, we remind the reader that the lag structure in this data is approximate, and so we cannot be sure of precedence. In any case, the comovement of the two time series is striking, and process engineers felt it was worth more detailed study.

To show how the SII chart can discredit an apparent association found in a process covariation chart, consider Figure 5. This shows the relationship between total bubbler flow (variable 42) and top drip (quality measurement 63). The first black square in the lower lefthand block suggested that large flow values were associated with good top drip levels. Although the boxplots generally support this view, the time series overlay shows that the control variable was held constant over much of the period, until the process fell out of control and the engineers began to chase it. Thus virtually all of the best 10% of top drip values occurred in a period of stability, and the control variable values didn’t change until after the process started to produce poor quality. This strongly reduces the chance of a useful causal relation between 42 and 63.

(Insert Figure 5 about here.)
For the datasets provided by PPG Industries and Alcoa, the result of the preanalysis stage was a dataset that can be used by any standard computer package. Also, the preanalysis discovered apparent relationships among the control and quality variables that suggest strategies for process improvement. Alcoa, PPG and the SCQI continue to examine the data, and all parties feel that useful process understanding has emerged. Although much of the work is specific and proprietary, the pattern of the analysis should apply to many continuous manufacturing industries.

Acknowledgement

This research was sponsored and partially supported by the Statistical Center for Quality Improvement. The authors thank Gene Klaber and Roy Crissman at PPG Industries for many discussions on current glass technology, Suraj Rao for programming assistance, and two referees for valuable suggestions.
Bibliography


Table 1: PPG Fresno Plant Control Variables.

Caption: For each control variable, this table indicates the identification number used in this paper, a short technical description supplied by PPG glass engineers, and the estimated lag between a change in the variable's value and the influence on the values of the quality variables listed in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Date in Julian Days (32756 = 11/4/89)</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Shift (1=12-8, 2=8-4, 3=4-12)</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>#1 Crown located between #1 &amp; #2 ports</td>
<td>1 shift</td>
</tr>
<tr>
<td>4.</td>
<td>#2 Crown located between #2 &amp; #3 ports</td>
<td>1 shift</td>
</tr>
<tr>
<td>5.</td>
<td>#3 Crown located between #3 &amp; #4 ports</td>
<td>1 shift</td>
</tr>
<tr>
<td>6.</td>
<td>#4 Crown located between #4 &amp; #5 ports</td>
<td>1 shift</td>
</tr>
<tr>
<td>7.</td>
<td>#5 Crown located between #5 &amp; #6 ports</td>
<td>1 shift</td>
</tr>
<tr>
<td>8.</td>
<td>#6 Crown located between #6 port and #1 Dog House</td>
<td>1 shift</td>
</tr>
<tr>
<td>9.</td>
<td>#7 Crown located at centerline of #1 Dog House</td>
<td>1 shift</td>
</tr>
<tr>
<td>10.</td>
<td>#8 Crown located in uptank Refiner</td>
<td>1 shift</td>
</tr>
<tr>
<td>11.</td>
<td>#9 Crown located in middle of Refiner</td>
<td>1 shift</td>
</tr>
<tr>
<td>12.</td>
<td>#10 Crown located in downtank Refiner</td>
<td>1 shift</td>
</tr>
<tr>
<td>13.</td>
<td>#1 Bottom located between End Wall and #1 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>14.</td>
<td>#2 Bottom located at centerline of #2 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>15.</td>
<td>#3 Bottom located at centerline of #4 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>16.</td>
<td>#4 Bottom located at centerline of #6 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>17.</td>
<td>#5 Bottom located at centerline of #1 Dog House</td>
<td>3 shifts</td>
</tr>
<tr>
<td>18.</td>
<td>#6 Bottom located in uptank Refiner</td>
<td>3 shifts</td>
</tr>
<tr>
<td>19.</td>
<td>#7 Bottom located in middle of Refiner</td>
<td>3 shifts</td>
</tr>
<tr>
<td>20.</td>
<td>#8 Bottom located in downtank Refiner</td>
<td>3 shifts</td>
</tr>
<tr>
<td>21.</td>
<td>#1 Bottom located between End Wall and #1 port</td>
<td>6 shifts</td>
</tr>
<tr>
<td>22.</td>
<td>#2 Bottom located at centerline of #2 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>23.</td>
<td>#3 Bottom located at centerline of #4 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>24.</td>
<td>#4 Bottom located at centerline of #6 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>25.</td>
<td>#5 Bottom located at centerline of #1 Dog House</td>
<td>6 shifts</td>
</tr>
<tr>
<td>26.</td>
<td>#6 Bottom located in uptank Refiner</td>
<td>6 shifts</td>
</tr>
<tr>
<td>27.</td>
<td>#7 Bottom located in middle of Refiner</td>
<td>6 shifts</td>
</tr>
<tr>
<td>28.</td>
<td>#8 Bottom located in downtank Refiner</td>
<td>6 shifts</td>
</tr>
<tr>
<td>29.</td>
<td>Left Bent Tube</td>
<td>1 shift</td>
</tr>
<tr>
<td>30.</td>
<td>Right Bent Tube</td>
<td>1 shift</td>
</tr>
<tr>
<td>31.</td>
<td>Left Radimatic at end of Refiner</td>
<td>0 shifts = out of refiner</td>
</tr>
<tr>
<td>32.</td>
<td>Middle Radimatic at end of Refiner</td>
<td>0 shifts</td>
</tr>
</tbody>
</table>
Table 1, continued.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.</td>
<td>Right Radimatic at end of Refiner</td>
<td>0 shifts</td>
</tr>
<tr>
<td>34.</td>
<td>Left FEF air in inches of water</td>
<td>2 shifts</td>
</tr>
<tr>
<td>35.</td>
<td>Left FEF gas in inches of water</td>
<td>2 shifts</td>
</tr>
<tr>
<td>36.</td>
<td>Right FEF air in inches of water</td>
<td>2 shifts</td>
</tr>
<tr>
<td>37.</td>
<td>Right FEF gas in inches of water</td>
<td>2 shifts</td>
</tr>
<tr>
<td>38.</td>
<td>Left combustion air flow in MCFM</td>
<td>2 shifts</td>
</tr>
<tr>
<td>39.</td>
<td>Right combustion air flow in MCFM</td>
<td>2 shifts</td>
</tr>
<tr>
<td>40.</td>
<td>Left air/gas ratio</td>
<td>2 shifts</td>
</tr>
<tr>
<td>41.</td>
<td>Right air/gas ratio</td>
<td>2 shifts</td>
</tr>
<tr>
<td>42.</td>
<td>Total bubbler flow in SCFM</td>
<td>2 shifts</td>
</tr>
<tr>
<td>43.</td>
<td>Submerged Cooler location in glass</td>
<td>0 shifts</td>
</tr>
<tr>
<td>44.</td>
<td>Percent of total gas on #1 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>45.</td>
<td>Percent of total gas on #2 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>46.</td>
<td>Percent of total gas on #3 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>47.</td>
<td>Percent of total gas on #4 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>48.</td>
<td>Percent of total gas on #5 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>49.</td>
<td>Percent of total gas on #6 port</td>
<td>3 shifts</td>
</tr>
<tr>
<td>50.</td>
<td>Total gas flow in SCFM</td>
<td>3 shifts</td>
</tr>
<tr>
<td>51.</td>
<td>Redox Ratio</td>
<td>0 shifts</td>
</tr>
<tr>
<td>52.</td>
<td>Thickness of glass</td>
<td>0 shifts</td>
</tr>
<tr>
<td>53.</td>
<td>Batch Percentage</td>
<td>2 shifts</td>
</tr>
<tr>
<td>54.</td>
<td>Salt Cake per 1000 Sand</td>
<td>2 shifts</td>
</tr>
</tbody>
</table>
Table 2: PPG Fresno Plant Quality Variables.

The following variables represent quality measurements obtained by PPG upon different features of plate glass produced in a plant near Fresno, California. The table indicates the technical name and the identification number used in this paper.

<table>
<thead>
<tr>
<th>No.</th>
<th>Quality Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Transmittance at 380 nm corrected to .250</td>
</tr>
<tr>
<td>56</td>
<td>Transmittance at 400 nm corrected to .250</td>
</tr>
<tr>
<td>57</td>
<td>Transmittance at 500 nm corrected to .250</td>
</tr>
<tr>
<td>58</td>
<td>Transmittance at 999 nm corrected to .250</td>
</tr>
<tr>
<td>59</td>
<td>Small bubbles per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>60</td>
<td>Large Bubbles per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>61</td>
<td>Ream Knots per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>62</td>
<td>Stones per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>63</td>
<td>Top Drip per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>64</td>
<td>Tridymite per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>65</td>
<td>Total Defects per 100 square feet normalized to .250&quot;</td>
</tr>
<tr>
<td>66</td>
<td>Seeds per square foot normalized to .250&quot;</td>
</tr>
<tr>
<td>67</td>
<td>Actual Small Bubbles per 100 square feet</td>
</tr>
<tr>
<td>68</td>
<td>Actual Large Bubbles per 100 square feet</td>
</tr>
<tr>
<td>69</td>
<td>Actual Ream Knots per 100 square feet</td>
</tr>
<tr>
<td>70</td>
<td>Actual Stones per 100 square feet</td>
</tr>
<tr>
<td>71</td>
<td>Actual Top Drip per 100 square feet</td>
</tr>
<tr>
<td>72</td>
<td>Actual Tridymite per 100 square feet</td>
</tr>
<tr>
<td>73</td>
<td>Actual Total Defects per 100 square feet</td>
</tr>
<tr>
<td>74</td>
<td>Actual Seeds per square foot</td>
</tr>
</tbody>
</table>
Figure 0.1: Typical Missing Values Chart: A "." is a planned missing value, a "!" is a value missing for unknown reasons, a "|" is an unplanned missing value whose cause is known. Features shown in the figure include: (a) a plant shutdown, (b) a pattern of missing data, (c) a subsystem shutdown, and (d) a subsystem failure.
Figure 0.2: Extreme Values Chart for the Fresno PPG Glass Data: A "." indicates a very small value, a "|" indicates a very large value. Serious process deterioration arose near time point 32767, and engineers attempted several adjustments to restore control.
Figure 0.3: Process Covariation Chart for the Fresno PPG Glass Data
Figure 0.4: SIB Chart for the Fresno PPG Glass Data: Variables 38 and 61
Figure 0.5: SIB Chart for the Fresno PPG Glass Data: Variables 42 and 63