ORISE--97052302

A STUDY OF THE EFFECT OF NONRANDOM SAMPLING ON THE
ESTIMATION OF EXPOSURE VARIABILITY USING EXPOSURE-TIME
SERIES FROM THE OSHA IMIS HEALTH INSPECTION DATABASE

A Research Project

By
Lee Tanner Martinez

Submitted in partial fulfillment of the requirements
for the degree of Master of Science of Public Health in the
School of Public Health, The University of Alabama at Birmingham

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

**Portions of this document may be illegible
in electronic image products. Images are
produced from the best available original
document.**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ABSTRACT

The effect of nonrandom sampling on the estimation of exposure variability, namely the geometric standard deviation (GSD), was studied using time series of personal exposures selected from the Occupational Health and Safety Administration's (OSHA) IMIS Database. Seven data sets, representing seven air contaminants, were initially selected from the IMIS database using Paradox for Windows. These data sets contained records of 8-hour time-weighted average exposures to the specific contaminant, the job description of the exposed worker, the date of sampling, and the company name where the sampling occurred. Time series of exposures were then selected from each data set according to specific criteria, using a series of FoxPro programs. A GSD was then calculated for each time series and the GSDs were analyzed for each contaminant as a function of the period over which monitoring occurred and as a function of the number of measurements within an exposure-time series. Analyses of the GSDs as a function of sample size (n) demonstrate that as n for a given contaminant increases, the median GSD also tends to increase. When the data sets for a particular contaminant were considered as a function of the time interval over

which exposure measurements were collected, the observation periods of greater than one week consistently produced a higher median GSD than the within-week series. A possible explanation for the lower within-week GSDs may be the presence of autocorrelation and/or spatial correlation, which have the effect of reducing the variability of the exposure distribution. Since the nonrandom nature of many of the time series in the study allow for the possibility of such correlation, it is recommended that industrial hygienists employ either a random sampling strategy to prevent the effects of spatial correlation and autocorrelation, or employ a statistical procedure that accurately evaluates the presence of such correlation.

ACKNOWLEDGEMENTS

I would like to express my thanks to Mr. Robert Matthews for his work in converting the OSHA IMIS database to a PC compatible format. Sincerest gratitude to Mr. Terrell Herzig for his tireless efforts to develop and fine tune the FoxPro programs needed for completion of this project. Finally, I gratefully acknowledge the support of my committee members, Dr. Vernon E. Rose, Dr. J. Michael Hardin, and Dr. Jimmy L. Perkins. A special thanks to Dr. Perkins, the committee chairman, for his guidance, support and patience.

The research was performed under appointment to the U.S. Department of Energy, Industrial Hygiene Graduate Fellowship Program, administered by Oak Ridge Institute for Science and Education.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT.....	ii
ACKNOWLEDGEMENTS.....	iii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
INTRODUCTION.....	1
AUTOCORRELATION.....	4
SPATIAL CORRELATION.....	22
OSHA IMIS DATABASE.....	24
OBJECTIVE.....	29
MATERIAL AND METHODS.....	30
INITIAL SELECTION OF DATA SETS.....	30
CRITERIA FOR STUDY DATA.....	31
USE OF FOXPRO PROGRAMS FOR DATA SELECTION.....	33
Quality Control.....	37
GSD Calculation.....	37
RESULTS AND DISCUSSION.....	39
CONCLUSIONS.....	54
REFERENCES.....	57
APPENDIX I.....	60
APPENDIX II.....	64
APPENDIX III.....	69

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Summary of the FoxPro Programs' Operations.....	33
II	Summary of Substitute LOD Values.....	36
III	Summary of the Number of Records and Time Series Per Contaminant.....	44
IV	Summary of GSDs for Asbestos Data Sets of Different Sizes and Observation Periods.....	45
V	Summary of GSDs for Coke Oven Emissions Data Sets of Different Sizes and Observation Periods.....	45
VI	Summary of GSDs for Total Dust Data Sets of Different Sizes and Observation Periods.....	46
VII	Summary of GSDs for Formaldehyde Data Sets of Different Sizes and Observation Periods.....	46
VIII	Summary of GSDs for Iron Oxide Data Sets of Different Sizes and Observation Periods.....	47
IX	Summary of GSDs for Styrene Data Sets of Different Sizes and Observation Periods.....	47
X	Summary of GSDs for Welding Fume Data Sets of Different Sizes and Observation Periods.....	48

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Diagram of IMIS database structure.....	26
2A	Asbestos data sets > 5 days.....	40
2B	Asbestos data sets < 5 days.....	40
3A	Coke oven emissions data sets > 5 days.....	40
3B	Coke oven emissions data sets < 5 days.....	40
4A	Total dust data sets > 5 days.....	41
4B	Total dust data sets < 5 days.....	41
5A	Formaldehyde data sets > 5 days.....	41
5B	Formaldehyde data sets < 5 days.....	41
6A	Iron oxide fume data sets > 5 days.....	42
6B	Iron oxide fume data sets < 5 days.....	42
7A	Styrene data sets > 5 days.....	42
7B	Styrene data sets < 5 days.....	42
8A	Welding fume data sets > 5 days.....	43
8B	Welding fume data sets < 5 days.....	43
9	Plot of Median GSD Trends by Sample Size.....	49

INTRODUCTION

The practice of industrial hygiene involves the recognition, evaluation, and control of environmental hazards in the workplace. As part of the evaluation process, the industrial hygienist often utilizes monitoring strategies to measure the exposures of a population of workers. If statistically-grounded, the sampling strategies should provide a reasonable estimate of the mean and variance of the underlying exposure distribution. These two statistics can then be manipulated through the use of various statistical tools, such as confidence intervals, to assess the degree of exposure of the population over time.

It is now recognized that occupational exposures to air contaminants fluctuate from day to day. This variation is largely due to several variables including emissions from the contaminant source, the degree of ventilation, and the mobility and work practices of the employee.¹ Due to this variation, it has been proposed as a general rule that occupational exposures over time are adequately described by the lognormal distribution.² This distribution has the property that the logarithms of the measured air

concentrations are normally distributed, which allows for the use of statistical tools in making decisions concerning the exposure of workers.

The two-parameter lognormal distribution is most often used to characterize occupational exposures.³ This distribution can be completely described by the geometric mean (GM) and the geometric standard deviation (GSD). These two parameters can then be used to calculate a confidence interval around the mean exposure in order to evaluate the exposure with respect to some occupational exposure limit (OEL); however, in order to use the GM and GSD for such statistical purposes, the exposure data must be collected in such a manner as to make the assumption of independence reasonable.

Exposure measurements are often collected according to a sampling strategy that does not ensure independent data. One such scheme is worst-case sampling, in which workers assumed to have the greatest exposures are selected for monitoring. This subjective selection of workers for worst-case assessment is not random sampling and thus likely produces biased results. These results then, cannot be used to determine the mean and variance of the workers' exposure distribution.

Compliance sampling, sometimes referred to as campaign sampling, is also used as a means of assessing worker exposure. This strategy involves collecting a minimal number of exposure measurements during brief campaigns of at most a few days.⁴ The largest measured result is then compared directly to an occupational exposure limit as a means of determining compliance. Non-compliance is declared if the OEL is exceeded.

There are distinct problems associated with this strategy. First, it has been shown that due to the large variability of worker exposure, any population of exposures can be declared out of compliance given a sufficient sample size.⁵ In other words, the probability that any measured exposure will exceed the OEL increases with the number of measurements collected.¹ It follows that compliance can be maximized by minimizing the number of measurements.

Secondly, since the small number of measurements are obtained consecutively, they may not be independent. Thus, the common practice of using these measurements to estimate the population parameters (i.e. mean and variance) of the exposure distribution may lead to biased estimates, which could jeopardize workers' health.¹ These problems place the industrial hygienist in the uncomfortable position of choosing between a sampling strategy that adequately

describes the exposure distribution (requires relatively many measurements) and one that maximizes the probability of compliance through measurement minimization.

One sampling strategy that attempts to provide representative and random measurements is the exposure zone concept. This strategy groups workers based on process, work task, and contaminant source similarities.⁶ A random subgroup from each exposure zone is then selected for monitoring. Since workers within each zone are assumed to experience the same exposure levels, the information obtained from the subgroup is deemed representative of the group as a whole. It has been stated that if the variability of the underlying exposure distribution exceeds a geometric standard deviation of about 3, then the exposure zone needs to be reassessed.⁷ Regardless of the strategy used, it must ensure that the collected data are independent in order to provide an unbiased estimate of the mean and variance of the exposure distribution.

Autocorrelation

Statistical methods are often used by the industrial hygienist to estimate the mean and variance of a worker's exposure distribution to various occupational hazards. These methods generally assume that the individual measurements are independent. Yet, particular sequences or

time series of exposures are often not independent but rather are correlated over time.⁸ Thus, a third parameter, the autocorrelation coefficient, is needed to estimate the degree of correlation.

The term autocorrelation refers to the correlation of measurements of a single variable (eg. a worker's exposure) measured at different times.⁸ In other words, the present measured exposure is predicted to some extent by the past measurements.⁹ The degree of autocorrelation between observations is determined by the sample autocorrelation coefficients, which are used to describe the probability model of the generated data.¹⁰ Autocorrelation can lead to biased estimates of the mean and variance, if appropriate adjustments to the estimates are not made.

The practical importance of autocorrelation is that it can lead to a reduction in information known about parameters of a particular distribution, such as the sample mean and variance. This could then lead to the erroneous estimation of confidence intervals about the mean that are either too narrow (most likely) or too broad. The obvious implication in the first case is that the hygienist could declare an exposure level to be below its respective OEL. This could jeopardize workers' health, as well as place the employer at risk of being cited by OSHA. The latter case

could result in the unnecessary allocation of resources to reduce perceived excessive exposures. Regardless of the outcome, autocorrelation must be considered in industrial hygiene surveys.

According to Box and Jenkins, the probability models for time series of most physical processes are referred to as stochastic processes.¹¹ A stochastic process is "a statistical phenomenon that evolves in time according to probabilistic laws." An important class of stochastic processes are those which are stationary. Stationarity implies that the mean, variance, and autocorrelation structure of the underlying exposure distribution do not change over the time period of interest.⁴

The correlation between observations some k time units apart, termed the time lag k , is called the autocorrelation function $\rho(k)$ of the time series.¹⁰ The autocorrelation function $\rho(k)$ is defined as the sample autocorrelation divided by the sample variance.⁹ The autocovariance, which defines the joint probability distribution of $P(X_t, X_{t+k})$, measures the covariance between X_t and X_{t+k} and is dependent only on the lag time k between the two.¹¹ An estimate of $\rho(k)$ is given by

$$r_k = \frac{\sum_{t=1}^n (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad (1)$$

where the numerator is equal to the autocovariance and $-1 < r_k < 1$ depending on the degree of correlation between observations. An autocorrelation coefficient (r_k) of zero indicates completely independent observations.⁹ Box and Jenkins have stated that a reasonable estimate of the autocorrelation function at a maximum time lag of $n/4$, where n is the number of consecutive measurements, would require at least fifty measurements.¹¹ Since the variance of the autocorrelation function for independent and identically distributed random variables is equal to $1/n$, 95% confidence intervals can be constructed around the coefficient by $r_k \pm 2/(n)^{0.5}$.¹⁰ The coefficient is considered significant at the 5% level if zero is not included in the confidence intervals.

A stochastic model commonly used for analyzing time series is the autoregressive model (AR).⁹ This is simply a regression model in which the variable is regressed on past values of itself. The model is expressed as

$$x_t = \alpha_1 x_{t-1} + \dots + \alpha_p x_{t-p} + z_t \quad (2)$$

where X_t is expressed as a composite of weighted past values plus a random shock (Z_t).¹⁰

According to Francis et al the first-order autoregressive model, (AR1), is most often used to describe time series of environmental data.⁸ This model expresses the current observation as a weighted function of the immediate past observation plus a random shock (Z_t). The process has the form

$$X_t = \alpha X_{t-1} + Z_t \quad (3)$$

where X_t and X_{t-1} are sequential measurements; Z_t is a series of uncorrelated, normally distributed random variables with mean = 0 and standard deviation = $\sigma(1-\alpha^2)^{0.5}$; and α is the weighting value for the past concentration, which varies from -1 to 1. The autocorrelation function for a first-order autoregressive process is related to the autocorrelation parameter by $\rho(k) = \alpha^k$, and thus displays exponential decay with increasing lag.¹¹

An important consideration for industrial hygienists is the role of autocorrelation in sample size estimations. Many such estimations are based on the assumption of independent data. For instance, Gilbert provides an equation for calculating the sample size, n , necessary to estimate the sample mean with a prespecified level of

accuracy and precision when data are uncorrelated.³ This estimator is expressed as

$$n = (Z_{1-\alpha/2} \sigma / d)^2 \quad (4)$$

where $Z_{1-\alpha/2}$ is the $(100-\alpha/2)$ confidence level, σ is the standard deviation of the population, and d is the absolute margin of error that can be tolerated. Since σ is not known in most instances, but rather is estimated by the sample standard deviation, s , then the Z-score is replaced by $t_{1-\alpha/2}$.

The problem with sample size determinations that assume independent data, is that if autocorrelation is present then the sample variance is likely to be underestimated.⁹ Subsequent estimations of n will then also be underestimated. Thus, the number of random measurements needed to adequately estimate the sample mean when measurements are not independent is given by

$$n = D(1 + 2 \sum_{i=1}^{n-1} \rho_i) \quad (5)$$

where $D = (Z_{1-\alpha/2} \sigma / d)^2$ and ρ_i is the true correlation between measurements at the i -th lag.³

Evidence documenting the presence of autocorrelation in environmental data has historically been derived from air pollution studies. Hirtzel and Quon studied the degree and persistence of autocorrelation in time series of hourly and mean 8-hour carbon monoxide (CO) measurements.¹² The data consisted of hourly concentrations of CO measured continuously over a three and one-half year period. The 42 month record of hourly CO concentrations was broken into 30 distinct stationary periods of 1 to 3 months duration using a t-test for differences between adjacent monthly means at the 95% confidence level. 12 of these stationary periods were analyzed to determine the statistical dependence of hourly and mean 8-hour carbon monoxide concentrations. The maximum autocorrelation coefficients for hourly CO measurements ranged from 0.86 to 0.42 for lags of 1 hour and 36 hours respectively; while maximum coefficients for 8-hour CO measurements ranged from 0.98 to 0.40 for the same respective time lags. Significant correlation among the CO data was found to persist up to a time lag of 360 hours, indicating that there are relatively few independent measurements in a continuous record of air quality.

This study also estimated the equivalent number of independent carbon dioxide measurements within a specified time period using a ratio of variance test.¹² This test is given as σ_1^2 / σ_2^2 , where σ_1 and σ_2 denote the respective

variance of hourly and daily average CO concentrations for the sample period. For the entire 42 month record, only 3 independent hours per day could be determined, while the number per year was found to be on the order of 100. An alternative estimate, using the autocorrelation coefficients, ρ_{ij} , gave a similar result. The theoretical parameter, given by

$$\sigma_2^2 = \frac{\sigma_1^2}{n^2} \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} \quad (6)$$

where ρ_{ij} is the autocorrelation between the i-th and j-th observations, assumes that the autocorrelation is negligible after two months. An estimate of the parameter is determined by taking estimates of the variables in the equation. The most optimistic estimate using this method found only approximately 1000 independent hours per year. These results indicated that the assumption of independent and random air quality data is not met. Thus statistical estimates of the variance of the sample mean would be underestimated, with a subsequent underestimation of the confidence intervals about the sample mean. The authors showed that confidence intervals would be underestimated by an order of magnitude if the autocorrelation was not taken into account.

Hirtzel and Quon have also studied the degree of autocorrelation in continuous records of hourly ozone (O_3) and daily maximum hourly ozone concentrations taken from five monitoring sites in the Chicago metropolitan area.¹³ Autocorrelation coefficients for daily maximum hourly O_3 concentrations ranged from 0.38 to 0.67 at a time lag of 1 day, and decayed exponentially to ± 0.2 at time lags of 20 days or more. Hourly ozone measurements had autocorrelation coefficients of 0.71 to 0.92 at a time lag of 1 hour for summer months, and coefficients from 0.72 to 0.85 for non-summer months. Autocorrelation coefficients were found to be significant up to time lags of 360 hours.

Estimates of the equivalent number of independent measurements in a continuous ozone record were also determined in this study using a ratio of variance test.¹³ Results indicated that only 20-79% of the total possible daily maximum O_3 measurements and 1-4% of the hourly O_3 measurements could be considered statistically independent for the five stations. These results illustrate once again that statistical estimates based on the assumption of independent air quality data will result in significant error if the correlation among the data is not considered.

Cheng et al conducted a time series analysis of coal data from preparation plants and found that failure to

account for the correlation in the data resulted in gross underestimation of the sulfur dioxide emission violations of actual coal.¹⁴ The geological formation of a coal seam has been found to create a spatial structure continuity in the coal mine which is somewhat preserved during mining, transport, and handling of the coal. This spacial structure is transformed into a temporal continuity in the coal feed when entering the preparation plants. The authors theorized that any sample value would be correlated in some fashion with neighboring points in time.

Using a first-order autoregressive model, the authors found the expected number of sulfur dioxide exceedences per year to be 5 times higher for raw coal and 20 times higher for clean coal when assuming autocorrelation, as compared to the expected number when assuming independence.¹⁴ The marked difference in expected numbers for the two types of coal is due to the fact that clean coal is more highly correlated than raw coal. The authors further demonstrated that the more correlated the data, the greater the assumption of independence underestimated the number of sulfur dioxide exceedences.

As noted earlier, the effect of autocorrelated data is to reduce the precision with which statistical descriptors, such as the sample mean and variance, can be estimated. If

autocorrelated measurements are assumed to be independent, then estimates of the sample mean and variance will be imprecise. Hirtzel et al derived equations to account for the dependence of data when estimating the variance of the sample mean.¹⁵ For independent observations, the variance of the sample mean is given by $\sigma_{\bar{x}}^2 = \sigma^2/n$. However, when a series of data is autocorrelated the variance is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2} \sum_{i=1}^{n-1} (n-i) \rho(i) \quad (7)$$

where i is the time interval between successive measurements. Equation (7) is simplified to

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n} \left[1 + \frac{2}{n} \sum_{i=1}^{n-1} (n-i) \rho_i \right] \quad (8)$$

for a first-order autoregressive process.³ The variance can be estimated from these equations by using estimates of the different variables. If $\rho_1 > 0$, the expression in the brackets is greater than one; therefore, the calculated variance of the sample mean, assuming autocorrelation, will be greater than the variance assuming independence.

These studies on the statistical dependence of air quality measurements have indicated the importance of autocorrelation in environmental data sets. Yet, until recently there has been little research on the

autocorrelation structure of occupational exposure distributions. Roach first studied the serial correlation between successive measurements of airborne contaminant concentrations.¹⁶ His conclusion that the serial correlation coefficient between concentrations is dependent upon the number of air changes per hour in the workplace is consistent with a first-order autoregressive process.

Rappaport and Spear used a first-order autoregressive model to describe the statistical dependence of short-term occupational exposures to airborne contaminants.¹⁷ The authors derived that significant autocorrelation should exist in occupational exposures when the averaging time for measured air concentrations is 15 minutes or less. Furthermore, they concluded that the correlation was dependent more upon the decay-rate constant of the air concentration in the worker's breathing zone, than the air-exchange rate in the entire room. This decay-rate constant was determined to be a function of the contaminant source, ventilation and mixing, and worker activities.

In addition, Spear, et al have discussed that when the autocorrelation structure of a particular exposure distribution can be determined, estimation of the parameters of the distribution at one averaging time (ex. 15-min) will allow estimation of the parameters for any other averaging

time (ex. 8-hour).¹⁸ This has obvious application for those contaminants that have both a 15-minute short-term exposure limit (STEL) and an 8-hour TWA. Thus, knowledge of the degree of autocorrelation in occupational exposures becomes an important issue for the industrial hygienist trying to assess workers' exposures.

Francis, et al have studied the effect of autocorrelation on the estimation of workers' daily 8-hour time-weighted average exposures.⁸ Three computer-simulated exposure time series were derived from a normal distribution with $\mu = 20\text{mg}/\text{m}^3$ and $\sigma = 10\text{mg}/\text{m}^3$. The three series, which followed a first-order autoregressive process, were assigned different autocorrelation coefficients at a lag of one. Series A was purely random with $\rho_1 = 0.0$; Series B was moderately correlated with $\rho_1 = 0.4$; and Series C was highly correlated with $\rho_1 = 0.8$. Variances estimated from Series B and C were found to be lower than that of the uncorrelated series. The authors also determined that for Series B and C there was almost a 16% and 31% chance, respectively, that the estimated sample mean would not be included in the 95% confidence intervals when based upon five consecutive measurements, a common strategy of industrial hygienists. Similar results were obtained when variances were estimated from three lognormal time series.

Francis, et al also examined three industrial hygiene data sets from actual workplaces in order to assess the degree of autocorrelation present in occupational exposure time series.⁸ The data sets consisted of successive 8-hour TWA exposures to mercury in a chloralkali plant, to lead in an alkyl lead manufacturing facility, and to an organic vapor in a production plant. Relatively little day-to-day correlation was found, with only a few of the chloralkali workers and two short (2 hours) series of organic vapor exposures having statistically significant ($p < 0.05$) first-lag correlation coefficients. The lack of dependence indicated that campaign sampling may provide adequate estimates of statistical parameters when collecting 8-hour TWA measurements.

A study by Carter found no evidence of autocorrelation in the 8-hour TWA styrene exposures of five workers in a cultured marble fixture facility.¹⁹ Exposure measurements were collected consecutively for 32 days. Correlation coefficients for lags up to $N/4$ were calculated with only one value differing significantly from zero at the 95% confidence level.

Hopson determined the serial correlation in consecutive 8-hour TWA occupational noise exposures for four workers in a cultured marble fixture facility.²⁰ One worker was found

to have significant correlation at lags up to 3 days, while another worker had significant correlation at a lag of 1.

A study by Kumagai, et al found evidence of autocorrelation in a series of short-term (7.5-minute, 15-minute, 30-minute, and 60-minute) time-weighted average exposure concentrations of workers exposed to solvents in various operations.²¹ Autocorrelation coefficients at time lag of 1 (r_1) were calculated for the series using exposure data from 16 worker-chemical combinations. Numerous significant values of r_1 were exhibited by the 7.5-minute, 15-minute, and 30-minute TWA's indicating that consecutive measurements with averaging times of 30 minutes or less tended to be autocorrelated. R_1 coefficients for the 60-minute TWA's showed little correlation. The autocorrelation functions of the individual workers were found to display various patterns that depended on the exposure time series. This study also examined the degree of dependence among 8-hour TWA concentrations in 10 worker-chemical combinations, but found little correlation between consecutive measurements. These results indicate that for shorter averaging times (<30 minutes) random sampling is needed to prevent the accumulation of redundant information, while for averaging times >60-minutes, campaign sampling may allow an adequate estimation of statistical parameters.

The few studies that have examined the presence of autocorrelation in time series of occupational exposures showed little, if any, statistically significant correlation in 8-hour time-weighted average measurements.^{8,19,20,21} This, however, does not disprove the existence of autocorrelation in such exposures, simply because most industrial hygiene data sets lack the substantial sequences of measurements necessary to provide an adequate estimate of the autocorrelation function. Thus, the question of whether an alternative approach may be used to determine the presence of autocorrelation in 8-hour TWA exposures is raised.

Such an alternative approach was used by Buringh and Lanting. This study focused on the assessment of exposure variability by examining simulated exposure-time series and 420 data sets of personal exposures from actual workplaces.²² Analysis of the exposure-time series, which were randomly drawn from a lognormal distribution, determined that the small numbers of measurements often taken in practice by industrial hygienists will lead to large errors in the estimates of the GSD, i.e. the variance is underestimated. The authors determined that roughly 30 random measurements were required to adequately estimate an individual GSD.

The workplace data sets were examined to determine the extent to which nonrandom sampling affected estimates of within-worker variability (i.e. GSD).²² The 420 time series represented 25 industries, with sample sizes varying from 3 to 13 measurements. 249 of the series consisted of 3 to 6 measurements taken within 1 week, with most consisting of consecutive samples. The remaining 171 series had interruptions varying from 1 week to 6 months, indicating a more random sampling approach. Goodness-of-fit tests revealed that most of the series were approximated by the lognormal distribution function.

The GSD for each data set was then calculated. The authors analyzed the GSDs of the data sets as a function of the number of measurements in a time-series and found that the mean GSD increased with increasing N.²² The data sets were also broken into 2 groups: series with measurements made consecutively over periods of less than 1 week and those where measurements were interrupted for periods of more than 1 week. When the mean GSDs of these groups were compared, the latter showed much more variability (1.74 v. 2.65). These results indicated that estimates of within-worker variability will be grossly underestimated if based on observation periods of less than 1 week and on sample sizes of less than 5.

Buringh and Lanting theorized that autocorrelation was the probable reason for this underestimation; however, because the data sets were too small to allow analysis of autocorrelation, the authors had to derive circumstantial evidence.²² Computer simulation was used to develop an autocorrelated time series according to a first-order autoregressive process in which the current concentration was, on average, 80% of the immediate past concentration and 20% of the new concentration. The distribution of GSDs for both the simulated series and the workplace data sets were then compared and found to be quite similar. These results were not proof of autocorrelation; however, when weighed in conjunction with other evidence such as the increase of GSDs for exposure-time series with interruptions of 1 week or more, the authors concluded that consecutive measurements collected over short observation periods are likely to underestimate the actual within-worker variability.

While the previous studies have shown that autocorrelation, or correlation over time, can be a significant problem in industrial hygiene monitoring, another potential problem is correlation over space, or spatial correlation.

Spatial Correlation

Spatial correlation is the correlation between data measurements that are taken at some distance apart. It differs from temporal autocorrelation in that the dependence in a spatial process extends in all directions, while the variate in the time series is influenced by time or past values.²³

Several possible applications of spatial correlation have been described for use in the analysis of air pollution data.²⁴ For instance, with knowledge of the spatial correlation of some air pollutant, the concentration at some given location can be determined from the measured concentration at a reference site. Also, spatial correlation can be used to determine the spacing or density of the monitoring station network required to measure the pollution concentration of a given area at a desired degree of accuracy. This will optimize the monitoring network with a subsequent reduction in capital and operating costs.

The practical implication for industrial hygienists occurs when occupational exposure measurements are collected according to a monitoring strategy that utilizes exposure zones. The resulting data for workers who are performing similar tasks may be correlated over space. Thus, statistical estimates will likely be biased if the spatially

correlated data is assumed to be independent. Gilbert provides an equation for calculating the spatial correlation coefficient, ρ_c , which is determined by averaging the cross-correlation coefficients between workers or stations (n_s).³ The correlation between two workers (x_1, x_2) is given by

$$\rho_{12} = \frac{\sum_{j=1}^n (x_{1j} - \bar{x}_1) (x_{2j} - \bar{x}_2)}{[\sum_{j=1}^n (x_{1j} - \bar{x}_1)^2 \sum_{j=1}^n (x_{2j} - \bar{x}_2)^2]^{1/2}} \quad (9)$$

where n is the number of measurements for each worker.

Once ρ_c is determined, then the number of measurements, n , needed to calculate the sample mean with a prespecified level of accuracy and precision can be determined by

$$nn_s = (Z_{1-\alpha/2} \sigma / d)^2 [1 + \rho_c (n_s - 1)] \quad (10)$$

where $Z_{1-\alpha/2}$ is the level of confidence, σ is the standard deviation, and d is the precision level.³ Gilbert also provides an equation for the variance of the sample mean for correlated data, which is given by

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{nn_s} [1 + \rho_c (n_s - 1)] \quad (11)$$

Little research has been conducted on the presence of spatial correlation in occupational exposures. A study by Carter found evidence of spatial correlation in the 8-hour TWA styrene exposures of workers in a cultured marble fixture facility.¹⁹ The spatial correlation coefficients for two groups of workers were found to be 0.87 and 0.77 respectively. The author determined that the variance of the sample mean would have been underestimated by as much as 50% if the data had been assumed to be independent.

In a study by Hopson, spatial correlation coefficients were found not to be statistically significant from zero in the 8-hour TWA noise exposures of workers performing similar tasks in a cultured marble fixture facility.²⁰ A study by McBride also could not find spatial correlation among five centrifugally-spun pipe casting operators' dust exposures.²⁵ The lack of spatial correlation was attributed to background sources of dust which were directional, physical isolation of the workers by safety shields, and low overall exposure variability within the exposure zone.

OSHA IMIS Database

The Occupational Safety and Health Act of 1970 was passed by the U.S. Congress in order to assure as far as possible that every working person in the United States had a safe and healthful work environment.²⁶ In 1971, the

Occupational Safety and Health Administration (OSHA) was created to discharge the responsibilities set forth by the Act. Part of those responsibilities was to conduct health and safety inspections of workplaces in order to assess compliance with OSHA regulations and health standards. In 1971 OSHA started conducting such inspections. Since 1979 the data that has been generated for these health inspections has been stored ,with the exception of several states for which storage began in 1981, in the Integrated Management Information System (IMIS) Inspection Derived File.

This IMIS database, which was purchased from OSHA in the form of 3480 Tape Cartridges, consists of well over one million records. It is essentially broken into five files, one main file and four subfiles (Appendix 1). The main file consists of records containing the demographic information specific to each worksite visited. For instance, it contains the name, region and address of the company, the number of people employed by the company, the reason for the inspection (i.e. planned, complaint, referral, etc...), the number of people covered by the inspection, and the number of violations found. The main file also contains information pertaining to any monitoring that was performed at the site. For instance, it contains the sample date and type, the work area or job title of the person sampled, the

number of workers in the HEG, and the number of hazards found.

The four subfiles, respectively, contain additional information for records specific to each worksite/worker (see Figure 1). The Noscreen Subfile (NS) contains sub-records of sites where an in-depth survey took place;

Figure 1. Diagram of IMIS database structure

the Screen Subfile (S) contains sub-records where only a screening survey occurred; the Hazardous Substance Subfile (HS) contains sub-records pertinent to the specific substance surveyed; and the Violation Subfile (V) consists of sub-records containing violations information. Each subfile can contain multiple sub-records for each main site/employee combination.

The data were stored on the cartridges in Extended Binary Coded Decimal Interchange Code (EBCDIC), which is essentially the mainframe version of ASCII code. Since this mainframe code is unreadable for use on microcomputers, the data had to be reformatted into a usable form.

There were several problems with this procedure. First, the data was stored in a non linear fashion, such that distinct records for a subfile were essentially strung together in one continuous line of code. Consider the problem of reading two records shown in Figure 1, A to B and A to C. Each involves the same employee and would be coded on the tape as one continuous code. The "Inspection 1" data would be the same in each record; however, at the first node the information would change. The problem is how to tell the computer to read the data after each node. This is done with 4 NR (number) fields which relay to the computer how many of each subfile-type records follow. Each of the NR fields relates to S, NS, HS, or V respectively. For A to B and A to C, the Hazard Substance NR would be 3. For AB, the Screen NR would be 1 and Violations 3, but Noscreen would be 0. Therefore, in order to write a format statement for reading the data, the NR fields have to be read, which are then essentially put into the format statement.

Secondly, some of the data was "packed" so that each byte of memory contained more than one character. Normally each byte contains only one character. Since a computer cannot recognize that these characters are separate, a special command had to be written in order to "unpack" this data into a readable form.

The basic procedure for converting the data into a form usable on a microcomputer is described as follows. The first step was to create five blank files (1 mainfile and 4 subfiles) using Job Control Language (JCL), which is part of the Multiple Virtual Storage (MVS) system of IBM mainframe computers. This system is analogous to the Disk Operating System (DOS) of most microcomputers. Appendix IIA describes the sequence of JCL code used to create the five blank files.

The next step was to redefine these mainfile and subfiles using SAS format statements. See Appendix IIB. Next the variables were input. The mainfile input was straightforward; however, the subfiles require "Do loops" or repetitive sequences of steps that are repeated based on the value of the NR variable. These are seen at the end of Appendix IIC.

The final step was to convert the EBCDIC format into ASCII format, which is readable by microcomputers. This step was accomplished using a translation program. Appendix IID contains the sequence of code used to complete this step.

Objective

The objective of this study is to determine the effect nonrandom sampling over time may have on the estimation of variability, namely the geometric standard deviation (GSD), using time series of personal exposure data selected from the Occupational Safety and Health Administration's (OSHA) IMIS database. In essence, the Buringh and Lanting study will be duplicated for U.S. data.

MATERIALS AND METHODS

Initial Selection of Data Sets

The OSHA IMIS database was searched using Paradox for Windows, a database management program, in order to select data sets of personal exposures to various air contaminants. The choice of contaminants was based solely on their common occurrence in industry, and thus the likelihood that an extensive data set would be found for each within the database. Initial searches yielded seven data sets, representing seven air contaminants, containing records of 8-hour time-weighted average exposures to the specific contaminant, the job description of the exposed worker, the date of sampling, and the company name where the sampling occurred.

The records within each data set were then sorted, using Paradox for Windows, according to company name and job description. This placed the records in alphabetical order by company name; and then within a particular company, the job descriptions were ordered alphabetically with their corresponding exposure level and sample date.

Upon examination of the data sets following the sort, it was noted that the spellings of various job descriptions and company names were often not standardized (ex. John Doe Company v. John Doe Co.), or spellings were incorrect. It was determined that this would present a problem at a later stage of the study, thus the data sets were manually scanned to standardize the spelling of company names and job descriptions, and to correct any misspellings.

Criteria for Study Data

Once these initial data sets had been completely sorted and standardized, then specific criteria were set in order to select data to be used in the study. The criteria were as follows. 1) Exposure-time series created within specific job descriptions within a given company only. 2) For each time series at least three exposure measurements were necessary. 3) At least two of the three exposure measurements must have occurred within a consecutive five day period. 4) Only 8-hour TWA measurements were used. 5) Within a given exposure-time series, no more than 25% of the total number of measurements could have exposure levels of 0.0 (or < limit of detection, LOD). This last criterion was set to prevent subsequent geometric standard deviation (GSD) calculations from being disproportionately skewed by an excessive number of such records.

The rationale for the other four criteria were as follows. Basing exposure-time series on specific job descriptions was necessary because the IMIS database did not include the names of individual workers that were monitored. This made it impossible to select time series for a particular worker. Thus the assumption was made that within a given company, the exposure measurements for a specific job description, whether they were for the same worker or not, were representative of the exposures of all workers with that job description and consequently were included in the same exposure-time series.

The time series were limited to single companies as it could not be assumed that job descriptions were consistent from company to company. A minimum of two data were needed per time series so a GSD could be calculated. These two data needed to occur within a five day period so that the effects of autocorrelation, if any, could be seen. A comparison GSD needed to be calculated, that was presumed not to suffer from autocorrelation. Thus data taken over greater than five days were assumed to have less effect of autocorrelation. This second group (> 5 days) could be composed of the first two data points plus at least one more; hence, the need for time series of $n \geq 3$. Finally, it would not be reasonable to combine 8-hour TWA data with 15-minute STEL data, so only 8-hour measurements were used.

Use of FoxPro Programs for Data Selection

The actual selection of data meeting the study criteria was accomplished by a series of programs written in FoxPro. The following text describes the actions of each program, in the order in which each program was used. Table 1 summarizes the functions of each program while Appendix III shows the actual programs.

The Chgdate program (see Chgdate.prg in Appendix III) converted the numeric sample date field, as recorded in the OSHA database, to a legitimate field. The format of the sample date field was in the form yym,mdd (ex. 921,012 would be October 12, 1992). This format is illegitimate for use in FoxPro so it was converted to the form of mm/dd/yy. Once this was accomplished, then the new format replaced the existing format in each record.

TABLE I. Summary of the FoxPro Programs' Operations

Step	Program name	Function
1	Chgdate.prg	Converted sample date from a numeric form (ex. yym,mdd) to a legitimate date field (ie. yy/mm/dd).
2	EX1_A.prg	Selected those exposure-time series for a specific job description within a particular company that had at least 3 measurements. Subgrouped the series by sample dates that occurred within 5 consecutive days
3	Siftfrps.prg	Removed series for which exposure level values of 0.0 exceeded 25% of the total number of measurements in the series.
4	Rplzero.prg	Replaced remaining 0.0 exposure levels with the substitute value for each respective contaminant.
5	Complog.prg	Calculated the natural logarithms of the exposure level values.
6	Bldg.prg	Calculated S_i for each within-week group and each total time series. Created a table containing all summary data (called Masterlog). Masterlog was then split into corresponding Sub and Master data sets.

The EX1_A program (see EX1_A.prg in Appendix III) then selected exposure-time series of measurements meeting the study criteria. This was accomplished by reading all unique job descriptions into an array, called FNames, and saving them in sequential order by the sample date into temporary files. These temporary files were given unique numerical names under the field name of MASTER. Then each company/job description combination was examined to determine if there were at least three exposure measurements per combination. If not, then the records were deleted.

After all company/job description combinations with less than 3 records were eliminated, then the exposure-time series for each remaining company/job description was examined to determine if two or more measurements were taken within a consecutive five day period. If so, this subgroup of records was placed into a unique numerical file under the field name of CHILD, which could be related back to the respective MASTER file. If the time series did not contain at least two records occurring within a five day period, then the series was eliminated. This created two data sets for each contaminant: the MASTER set containing all complete exposure-time series for a particular contaminant and the CHILD subgroup containing only those subgroups of records that occurred within consecutive five day periods.

The Siftgrps program (see SIFTGRPS.PRG in Appendix III) then eliminated those exposure-time series containing records, with 0.0 exposure levels, whose numbers exceeded 25% of the total number of records in the time series. This action was carried out by examining each MASTER file and eliminating those that exceeded the criteria. Related CHILD files were then subsequently removed.

The Rplzero program (see Rplzero.prg in Appendix III) then replaced any remaining records with exposure levels of 0.0 with a substitute value derived for each respective contaminant. The substitute values were determined from the mass LOD and the recommended volumetric flowrate of the respective OSHA analytical method for each contaminant. These values were obtained either directly from the OSHA method itself or from OSHA Directorate of Technical Support personnel at the Salt Lake Technical Center in Salt Lake City, Utah. A concentration LOD for each contaminant was calculated and this value was used to replace the 0.0 values. It has been shown that $LOD/\sqrt{2}$ perhaps would be a better replacement value, producing GSD estimates that are less biased.²⁷ However, the error should be small and consistent for comparing various groups of GSDs. See Table II for the replacement values derived for each contaminant. The program first replaced the values in each MASTER file and then in each CHILD file.

TABLE II. Summary of Substitute LOD Values

Contaminant	Mass LOD	Recommended Volume (L)	Concentration LOD
Asbestos	2400 fibers	960	0.002 fibers/cc
Coke oven emissions	0.006 mg	960	0.006 mg/m ³
Formaldehyde	0.48 µg	24	20.0 µg/m ³
Iron oxide fumes	0.013 mg	960	0.013 mg/m ³
Styrene	4.7 µg	10	0.47 mg/m ³
Total dust	0.2 mg	960	0.2 mg/m ³
Welding fumes	2.0 µg	960	0.002 mg/m ³

The rationale for replacing the 0.0 values is as follows. If the 0.0 values for a given time series were used to calculate a GSD, then the resulting GSD would likely be too high, reflecting the more variable nature of the time series. Thus, by replacing 0.0 values with the respective LOD value, the estimated GSD would be more accurate with respect to the true GSD. In the event that a time series consisted of up to 25% of the LOD values, then the resulting GSD would likely be too low, reflecting the less variable time series. However, this possibility is fairly remote since manual scanning of the data sets determined that few time series consisted of more than approximately 5% LOD values.

The next step was to calculate the natural logarithm of the exposure level values. This was accomplished with the Complog program (see Complog.prg in Appendix III) which computed a natural log for each record and placed it under the field name of Log. This was done first on the MASTER files and then on the CHILD files.

The last step was to calculate the standard deviation of the log values (S_L) for the exposure-time series under MASTER and the subgroups under CHILD, and then to construct a data set containing all of the selected time series. This was done by the Bldlog program (see Bldlog.prg in Appendix III).

The Bldlog program used the data files created above to calculate the variance, the average and the standard deviation of the log values for each time series and subgroup. These fields were added to the existing files.

Quality Control

A quality control procedure was employed in order to determine the effectiveness of the FoxPro programs in selecting the required data. This procedure involved manually viewing the original data sets and selecting time series according to the study criteria. These time series were then compared to those in the selected data sets. For each data set 5 - 10 time series were compared. This procedure determined that the FoxPro programs did accomplish their designed purpose.

GSD Calculation

Once the Master and Sub data sets were created, they were imported back into Paradox where they were subgrouped

by contaminant. This produced both a Sub and a Master data set for each contaminant. These data sets were then imported into Solo, a statistical analysis software package. Solo's Transformations Panel was used to convert the S_L values to geometric standard deviations by the following equation:

$$GSD = \exp(S_L) \quad (12)$$

Solo's Descriptive Statistics Panel was then used to analyze the GSDs for each contaminant as a function of the period over which monitoring occurred (i.e. within 5 days v. greater than 5 days) and as a function of the number of measurements within an exposure-time series. PSIPILOT, a software plotting program, was then used to construct histograms of the distributions of GSDs as a function of the period over which monitoring occurred.

RESULTS AND DISCUSSION

The distribution of GSDs for each contaminant, as a function of both the sample size and observation period, was tested for it's approximation to the normal distribution using the Solo Statistical Package. This was important to know in terms of summarizing the GSD data. The distribution of GSDs for all combinations of sample size/contaminant and observation period/contaminant were not normally distributed, as determined by the I-test.²⁸ Most of the distributions were skewed to the right, although they were not lognormally distributed either.

Figures 2 - 8 show the distributions of GSDs by observation period for each contaminant, most of which appear to be fairly exponential in nature. The histograms show that for a given contaminant the exposure measurements collected over time periods of greater than five days tended to have greater variability than those 5-day series. This is determined by looking at the length and weight of the tails. The 5-day series also have greater numbers of low GSDs than the series collected over longer time periods.

FIGURE 2A. Asbestos data sets >5 days

FIGURE 2B. Asbestos data sets <5 days

FIGURE 3A. Coke oven emissions data sets >5 days

FIGURE 3B. Coke oven emissions data sets <5 days

FIGURE 4A. Total dust data sets >5 days

FIGURE 4B. Total dust data sets <5 days

FIGURE 5A. Formaldehyde data sets >5 days

FIGURE 5B. Formaldehyde data sets <5 days

FIGURE 6A. Iron oxide data sets >5 days

FIGURE 6B. Iron oxide data sets <5 days

FIGURE 7A. Styrene data sets >5 days

FIGURE 7B. Styrene data sets <5 days

FIGURE 8A. Welding fume data sets >5 days

FIGURE 8B. Welding fume data sets <5 days

Table III shows the total number of records obtained for each contaminant and the number of time series per sample size (n) and observation period (>5 days v. <5 days). For a given contaminant, each >5 day time series may have consisted of 2 or more <5 day series. Thus, the total number of <5 day series for a given contaminant is greater than the number of complete (i.e. >5 days) time series, or greater than the number of series summed up over all n. Summaries of the Solo analyses of GSDs as a function of the number of measurements (n) within an exposure-time series and as a function of the time period over which monitoring occurred are presented by contaminant in Tables IV - X.

The variable nature of the distributions can be seen in the ranges of GSDs, as shown in Tables IV - X. Within a given range, the GSDs tended to be grouped towards the low end, with just a few individual high GSDs. This disparity produced the skewed, although not lognormal, distribution.

TABLE III. Summary of the Number of Records and Time Series Per Contaminant

Contaminant	No. of Records	Number of Time Series					
		>5 days	≤5 days	n= 3	4	5	≥6
Asbestos	611	90	174	16	18	13	39
Coke oven emissions	157	28	52	6	7	4	12
Formaldehyde	324	49	89	6	13	7	23
Iron oxide fumes	453	59	120	8	15	4	31
Styrene	503	72	129	10	14	8	40
Total dust	810	144	238	36	39	27	46
Welding fumes	294	43	68	11	14	5	13

As a result, the mean GSDs, as presented in Tables IV - X, may not be suitable values of comparison since they do not reflect central tendency. Thus, the median GSDs will be considered as the values of interest, since within a given subgroup 50% of the GSDs lie above and 50% lie below the median GSD.

As shown in Tables IV - X, as the sample size for a given contaminant increases, the median GSD also tends to increase. This pattern is fairly consistent for all of the data sets, with sample sizes of six or greater yielding the highest median GSD estimates.

**TABLE IV. Summary of GSDs for Asbestos Data
Sets of Different Sizes and Observation Periods**

Number of Measurements	<u>GSD</u>	Median GSD	Range of GSDs
3	2.92	2.53	1.3 - 3.9 ^A , 4.8, 6.7
4	3.68	2.56	1.5 - 5.9 ^B , 8.1, 9.8
5	4.15	3.25	1.7 - 8.75
>6	4.22	3.35	1.4 - 10.1
Data sets collected within 1 week	2.33	1.78	1.0 - 9.9
Data sets with interruptions of > 1 week	4.36	3.14	1.3 - 8.4 ^C , 8.4-14.5 ^D

A - Includes 14 measurements

B - Includes 16 measurements

C - Includes 80 measurements

D - Includes 10 measurements

**TABLE V. Summary of GSDs for Coke Oven Emissions Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	<u>GSD</u>	Median GSD	Range of GSDs
3	3.91	3.76	1.6 - 5.8
4	2.52	2.18	1.5 - 2.7 ^A , 5.2
5	2.04	1.64	1.4 - 3.5
>6	2.96	2.77	1.4 - 4.4
Data sets collected within 1 week	2.17	1.64	1.0 - 4.5 ^B , 5.6, 6.1, 6.7
Data sets with interruptions of > 1 week	2.80	2.30	1.4 - 5.8

A - Includes 6 measurements

B - Includes 49 measurements

**TABLE VI. Summary of GSDs for Total Dust Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	<u>GSD</u>	Median GSD	Range of GSDs
3	2.50	2.16	1.1 - 3.5 ^A , 4.2, 5.0, 5.4, 7.0
4	2.31	1.87	1.2 - 5.1
5	2.13	1.67	1.2 - 3.2 ^B , 4.2, 5.2, 5.6
>6	2.98	2.34	1.4 - 5.6 ^C , 5.9, 6.4, 8.3
Data sets collected within 1 week	1.88	1.46	1.0 - 4.0
Data sets with interruptions of > 1 week	2.51	2.01	1.1 - 5.7 ^D , 5.9, 6.3, 6.8, 8.3

- A - Includes 33 measurements
- B - Includes 24 measurements
- C - Includes 43 measurements
- D - Includes 140 measurements

**TABLE VII. Summary of GSDs for Formaldehyde Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	<u>GSD</u>	Median GSD	Range of GSDs
3	2.84	3.14	1.3 - 4.1
4	3.12	2.01	1.2 - 7.5
5	3.30	2.30	1.2 - 7.5
>6	2.93	2.34	1.4 - 5.6 ^A , 8.9
Data sets collected within 1 week	2.04	1.44	1.0 - 5.1 ^B , 5.1, 5.5, 6.6
Data sets with interruptions of > 1 week	3.02	2.45	1.2 - 6.5 ^C , 7.1, 7.7, 8.9

- A - Includes 22 measurements
- B - Includes 86 measurements
- C - Includes 46 measurements

**TABLE VIII. Summary of GSDs for Iron Oxide Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	$\overline{\text{GSD}}$	Median GSD	Range of GSDs
3	1.89	1.72	1.2 - 2.8
4	2.15	1.77	1.2 - 3.5 ^A , 4.1, 4.7
5	2.06	1.78	1.6 - 3.1
>6	3.06	2.80	1.5 - 5.3 ^B , 6.0, 6.8
Data sets collected within 1 week	2.01	1.70	1.0 - 4.7 ^C , 5.3, 5.7, 6.3
Data sets with interruptions of > 1 week	2.70	2.18	1.2 - 5.6 ^D , 6.8, 7.3, 9.0

- A - Includes 13 measurements
- B - Includes 29 measurements
- C - Includes 117 measurements
- D - Includes 56 measurements

**TABLE IX. Summary of GSDs for Styrene Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	$\overline{\text{GSD}}$	Median GSD	Range of GSDs
3	1.68	1.62	1.1 - 2.3
4	2.32	1.62	1.2 - 2.9 ^A , 9.1
5	2.36	2.12	1.1 - 5.2
>6	2.50	2.18	1.2 - 5.4 ^B , 8.2, 9.6
Data sets collected within 1 week	1.68	1.46	1.0 - 4.0 ^C , 6.1, 7.8
Data sets with interruptions of > 1 week	2.33	1.95	1.0 - 5.3 ^D , 7.7, 8.9, 9.6

- A - Includes 13 measurements
- B - Includes 38 measurements
- C - Includes 127 measurements
- D - Includes 69 measurements

**TABLE X. Summary of GSDs for Welding Fume Data
Sets of Different Sizes and Observation Periods.**

Number of Measurements	$\overline{\text{GSD}}$	Median GSD	Range of GSDs
3	2.03	1.86	1.2 - 5.0
4	2.09	1.80	1.3 - 3.0 ^A , 5.2
5	2.19	2.12	1.7 - 2.6
>6	2.36	2.12	1.6 - 3.3 ^B , 5.0
Data sets collected within 1 week	1.70	1.48	1.0 - 3.1 ^C , 4.9, 5.2
Data sets with interruptions of > 1 week	2.17	1.97	1.2 - 3.0 ^D , 4.8, 4.9, 5.1

A - Includes 13 measurements

B - Includes 12 measurements

C - Includes 66 measurements

D - Includes 40 measurements

These larger sample sizes are considered to contain more independent measurements relative to those of five or fewer observations, since it is assumed that as n increases the average time in between measurements is likely to increase. Therefore, autocorrelation decreases and the true variance is not underestimated. This pattern is also reflected in similar sample sizes observed from the Buringh and Lanting study.²²

The exceptions to the above pattern are the data sets for formaldehyde and coke oven emissions where the sample size of three yielded the highest median GSDs (See Figure 9). However, in both cases only six time series were available for $n = 3$. These small sample sizes could have

led to spurious values by chance alone, since the overall trend across all contaminants is that GSD increases with n.

As shown in Tables IV -X, when the data sets for a particular contaminant were considered as a function of the time interval over which exposure measurements were collected (i.e. independent v. less independent samples), the observation periods of greater than one week consistently produced a higher median GSD than the 5-day series. This pattern, which was observed for all seven contaminants, is consistent with the pattern observed in Dutch industries in the Buringh and Lanting study.²²

Figure 9. Median GSD versus sample size

The 5-day series generally consisted of five or fewer exposure measurements, most of which were collected over consecutive days. This nonrandom approach to sampling is generally reflective of campaign sampling, whereby only a minimal number of samples are collected during a brief campaign of at most a few days.⁴ Since previous studies have shown that such samples may not be independent, and thus may produce biased statistical estimates, the question arises as to whether the 5-day results are reflecting correlation over time or space.^(8,19,20,21,22)

One possible explanation for the observance of lower median GSDs in the short-term series may be the autocorrelation of the exposure measurements within the series. Autocorrelation, or correlation over time, has been implicated as a potential problem when nonrandom sampling produces exposure measurements that are not independent.^(8,19,20,21,22) Since the 5-day series used in this study were likely collected in a nonrandom fashion, the possibility does exist that the measurements for these time series tended to be positively autocorrelated and thus that the GSD estimates were biased low. However, analysis of the autocorrelation structures of the 5-day series was not possible due to the limited sizes of the data sets.

A second possible explanation for the lower median GSD results observed for the short-term exposure-time series may be the presence of spatial correlation. Spatial

correlation, or the correlation among measurements that are taken some distance apart (ie. workers sampled on the same day), is a potential problem when monitoring the exposures of workers performing similar tasks.

Each exposure-time series included in this study consisted of the exposure measurements for a specific job description and not necessarily an individual worker. Additionally, some of the series used in this study also included multiple measurements from the same day. Thus, if a given short-term series consisted of exposure measurements that were collected on the same day for different workers with the same job description (i.e. an exposure zone), then the possibility exists that those measurements were not independent due to spatial correlation. This dependence would produce negatively biased GSD estimates, resulting in the lower median GSDs.

The spatial correlation structure was not estimated because of the inability to identify the exposure measurements for the different workers included in a given exposure series; however the influence of spatial correlation could possibly be estimated by determining the number of measurements within a time series that were collected on the same day(s). For instance, a high number of such measurements would indicate the likely presence of

spatial correlation, while a low number would indicate its absence.

Of course, the observed results could reasonably be due to a combined effect of autocorrelation and spatial correlation since the exposure-time series consisted of measurements taken over consecutive days and measurements taken on the same day for workers with the same job description. Again, this possibility could not be analyzed because of the limitations of the study data.

Finally, the results could possibly be due to the non-stationarity of the exposure distributions of different time series. The term stationarity implies that the mean, variance, and autocorrelation structure of the underlying distribution do not change over time.⁴ Since many of the large time series were likely collected over long time periods (ie. up to several years), then it is possible that the distributions for a given series changed due to process and other changes.

Thus, measurements collected early in the observation period (for example, during the first year) would represent an exposure distribution that may be different from the distribution for measurements taken at a later time (ie. several years). Since such time disparate distributions may

have different GSDs, any grand GSD calculations based on exposure measurements from both distributions would be inflated. In other words, the GSD for a > 5 day observation period could be reflecting the influence of two exposure distributions, resulting in a higher median GSD for the combined set than for either of the two distributions.

CONCLUSIONS

The analyses of the seven contaminant data sets, as a function of sample size and observation period, indicate that estimation of the GSD as a measure of exposure variability will yield lower results when based on small sample sizes (< 6) and short observation periods (< 1 week). Thus the use of nonrandom sampling strategies such as campaign sampling, where only a few exposure measurements are collected over several consecutive days, will likely produce results that will lead to biased statistical estimates such as the geometric standard deviation.

Although the limited sizes of the exposure-time series included in the data sets did not allow proof of day-to-day autocorrelation or worker-to-worker spatial correlation, it is clear that the nonrandom nature of many of the exposure-time series in this study allows the possibility for the correlation of exposure measurements over time and/or space. Thus, in order for the industrial hygienist to ensure that worker exposures are properly evaluated so that resulting statistical estimates are accurately estimated, two sampling approaches remain.

The first involves determining the correlational structure, whether time or spatial, in those situations where the hygienist believes that correlation may be significant. Evaluating the presence of such correlation will allow resulting statistical estimates, such as the variance and mean, to be more accurately made. Gilbert provides equations for estimating the variance and mean when the correlation coefficients are known for a normal distribution.³

There are a couple of downfalls to this approach. First, the equations provided by Gilbert assume that the underlying exposure distribution is normally distributed; therefore, they may not be applicable to environmental data which has been shown to be lognormally distributed. Thus, equations may have to be derived in order to estimate statistical parameters for occupational exposures when spatial and/or time correlation is present.

Secondly, the time and resources that must be spent to properly evaluate the correlation structure may be prohibitive. For instance, it has been recommended that a minimum of fifty consecutive measurements are needed to adequately determine the autocorrelation function.¹¹ Thus it may behoove the industrial hygienist to allocate these resources to a more reasonable approach.

The second, and more reasonable, approach is for the industrial hygienist to employ a random sampling strategy. Such a strategy should allow worker exposures to be evaluated in a manner that avoids biased statistical estimates while at the same time utilizing resources in a more efficient manner. This can be accomplished by distributing measurements over time and not sampling workers from the same exposure group on the same day.

REFERENCES

1. Rappaport, S.M. and S. Selvin: A Method for Evaluating the Mean Exposure from a Lognormal Distribution. *American Industrial Hygiene Association Journal*, 48(4):374-379 (1987).
2. Esmen, N. and Y. Hammad: Lognormality of Environmental Sampling Data. *Environmental Science and Health*, A12:29-41 (1977).
3. Gilbert, R.O.: *Statistical Methods for Environmental Pollution Monitoring*. New York: Van Nostrand Reinhold Company Inc., 1987.
4. Rappaport, S.M.: Exposure Assessment Strategies. Presented at the International Workshop on Exposure Assessment for Epidemiology and Hazard Control, Woods Hole, Massachusetts, 1988.
5. Rappaport, S.M.: The Rules of the Game-An Analysis of OSHA's Enforcement Strategy. *American Journal of Industrial Medicine*, 6:291-303 (1984).
6. Corn, M. and N. Esmen: Workplace Exposure Zones for Classification of Employee Exposures to Physical and Chemical Agents. *American Industrial Hygiene Association Journal*, 40:47-57 (1979).
7. Leidel, N.A. and K.A. Busch: Statistical Design and Data Analysis Requirements. In: Patty's Industrial Hygiene and Toxicology, 2nd ed. Vol. 3A, pp. 395-507. L.J. Cralley and L.V. Cralley, eds. John Wiley and Sons, New York, 1985.
8. Francis, M., S. Selvin, R. Spear, and S. Rappaport: The Effect of Autocorrelation on the Estimation of Worker's Daily Exposures. *American Industrial Hygiene Association Journal*, 50(1):37-43 (1989).
9. Gottman, J.M.: Time Series Analysis-A Comprehensive Introduction for Social Scientists. Cambridge: Cambridge University Press, 1981.
10. Chatfield, C.: The Analysis of Time Series-An Introduction. London: Chapman and Hall, 1984.
11. Box, G.E.P. and G.M. Jenkins: Time Series Analysis: Forecasting and Control, revised edition. San Francisco: Holden-Day, 1976.

12. Hirtzel, C.S. and J.E. Quon: Statistical Dependence of Hourly Carbon Monoxide Measurements. Journal of the Air Pollution Control Association, 29:161-163 (1979).
13. Hirtzel, C.S. and J. E. Quon: Statistical Analysis of Continuous Ozone Measurements. Atmospheric Environment, 15:1025-1034 (1981).
14. Cheng, B.H., B. Woodcock, D. Sargent, and A. Gleit: Time Series Analysis of Coal Data from Preparation Plants. Journal of the Air Pollution Control Association, 32: 1137-1141 (1982).
15. Hirtzel, C.S., J.E. Quon, and R.B. Corotis: Mean of Autocorrelated Air Quality Measurements. Journal of the Environmental Engineering Division, Proceedings of the American Society of Civil Engineers, 108:488-501 (1982).
16. Roach, S.A.: A Most Rational Basis for Air Sampling Programmes. Ann. Occup. Hyg, 20:65-84 (1977).
17. Rappaport, S.M. and R.C. Spear: Physiological Damping of Exposure Variability During Brief Periods. Ann. Occup. Hyg, 32:21-33 (1988).
18. Spear, R.C., S. Selvin, and M. Francis: The Influence of Averaging Time on the Distribution of Exposures. American Industrial Hygiene Association Journal, 47:365-368 (1986).
19. Carter, S.R.: An Estimation of Autocorrelation and Spatial Correlation from 8-Hour Time-Weighted Average Exposures. Masters Research Project, School of Public Health, The University of Alabama at Birmingham (1990).
20. Hopson, N.L.: A Study of Autocorrelation in a Population of Occupational Noise Exposures. Masters Research Project, School of Public Health, The University of Alabama at Birmingham (1990).
21. Kumagai, S., I. Matsunaga, and Y. Kusaka: Autocorrelation of Short-Term and Daily Average Exposure Levels in Workplaces. American Industrial Hygiene Association Journal, 54:341-349 (1993).
22. Buringh, E. and R. Lanting: Exposure Variability in the Workplace: Its Implications for the Assessment of Compliance. American Industrial Hygiene Association Journal, 52:6-13 (1991).
23. Ord, J.K. and A.D. Cliff: Spatial Autocorrelation: London Pion Limited, 1973.

24. Elsom, D.M.: Spatial Correlation Analysis of Air Pollution Data in an Urban Area. *Atmospheric Environment*, 12:1103-1107 (1978).
25. McBride, J.B.: A Study of Spacial Correlation in a Workplace Exposure Zone. Masters Research Project, School of Public Health, The University of Alabama at Birmingham (1991).
26. Plog, B.A., G.S. Benjamin, and M. A. Kerwin: Fundamentals of Industrial Hygiene, 3rd ed. National Safety Council, 1988.
27. Hornung, R.W. and L. D. Reed: Estimation of Average Concentration in the Presence of Nondetectable Values. *Applied Occupational and Environmental Hygiene*, 5:46-51 (1990).
28. Martinez, J. and B. Iglewicz: A Test for Departure from Normality Based on S Biweight Estimator of Scale. *Biometrika*, 68:331-333 (1981).

APPENDIX I

Site Mainfile

Act_nr	N	Unique number that ids. the OSHA-1 and is the record identifier.
Region	N	Region code of submitting office (01-10).
Areacode	N	Code for submitting federal Area Office or State 18 (b) office.
Ofc_code	N	Code identifying a Federal District Office or State 18 (b) Sub-office (00-99).
Sic	N	Primary SIC code for establishment activity
St_flag	A1	Federal (b) or State (a) inspection
Csho_id	A5	Identification number of compliance officer who conducted the inspection.
Name	A30	Name of establishment of business inspected.
Addr	A30	Location of site inspected
State	A2	State where site located
Zip	N	Zip code of site
Citycode	N	City code
Cntycode	N	County code
Owner_tp	A1	Owner designation; a=private; c=state govt.; b=local govt.; d=Federal govt.
Owner_cd	N	Agency code for Federal Agency Inspection
Adv_notc	A1	Advance notice given; b=no; x=yes
Opendate	N	Date of opening conferences or entry or attempted entry into the site of the inspection (yymmdd)
Clscdate	N	Date of the last closing conference held before issuing citations or exit from the inspection site
S_h	A1	Safety/Health; S = safety; H = health
Insptype	A1	Type of inspection Unprogrammed: A=fat/cat E=variance B=complaint F=followup C=referral G=related D=monitoring J=other Programmed: H=planned K=other I=related L=non-inspection
Insp_scp	A1	Inspection coverage A=comprehensive C=records only B=partial D=no inspection
Nr_estab	N	Max. nos of people employed at site during previous 12 months.
Nr_cover	N	Nos. of employees affected by inspection
Nr_empr	N	Nos. of employees controlled by employer nationally
Walk	A1	Employee representative present during inspection b = no; x = yes
Emp_int	A1	Employee(s) interviewed during inspection (State 18 (b) only); b = no; x = yes
Union	A1	Employees represented by Union
Cas_clos	N	b = open; x = closed
Pen_date	N	Date penalties are due

Fta_date	N	Date failure-to-abate penalties are due
Dur_src	A1	Code indicating source of penalty due date b = no due date n = date set to 9/1/85 by Nat'l office d = date set by the Field-submitted on OSHA 167-I r = when R event code on OSHA 166-I, penalty due date = Final order date + 60 days j = when J event code on OSHA 166-I, penalty due date = Final order date + 90 days f = final order date entered on OSHA 166-I i = Citation issuance date + 20 days a = informal settlement date + 20 days
Pa_hours	N	Sum of all inspection activity hours reported
Contdate	N	Earliest date of contest received for contested inspection
Remitamt	N	Total penalty dollars remitted for the inspection
Fta_amt	N	Total FTA penalty dollars remitted for the inspection
Tot_pen	N	Total current penalties issued for inspection
Tot_fta	N	Total FTA penalties issued for the inspection
Tot_viol	N	Total violations issued for the inspection
Totsviol	N	Total serious violations (S,W,R) issued for inspection
Hformtyp	A1	Health form type, Air Sample (B,S) form 91- noise 92, 93-direct read, 98=screen(wipes, bulk)
Hsamp_nr	N	Health sample nos.
Hsamp_dt	N	Date of sample
Hsamp_cs	A5	Identification nos. of the CHSO
Hjobdesc	A20	Job title of employee- is employee driven
Hnr_exp	N	Number of workers in HEG
Hexps_fq	N	Narrative on frequency of exposure; may not be stored
Hashdate	N	Date of shipment
Harcdate	N	Date of receipt
Hne_dba	N	Noise levels found
Nr_haz	N	Number of hazardous substances/ employee; even if not sampled
Nr_viol	N	Number of violation blocks (0-220)
Nr_noscr	N	Health data (91,92,93)
Nr_scr	N	Health data screening (98 only)
Nr_totsg	N	
Key	N	

Noscreen Subfile

Nr_noscr	N	Comprehensive health data (91,92,93)
Key	N	Field common to all files
Subcode	A4	4 Digit code for the substance monitored

Anal_req	A1	Whether sample was analyzed or not
Samptype	A1	Type of sample taken (area, personal)
Exp3type	A1	TWA, STEL, Ceiling, peak, full shift (codes)
Explevel	N	Level of exposure to substance
Units	A1	Units of exposure
Pel	N	Pel
Adj_pel	A1	Unusual work shift
Severity	N	Indicates severity of exposure
Cita1	A1	Indicates the OSHA std which the employer
Cita2	A1	violated by over exposing the employee
Cita3	A1	to the particular substance or hazard
Cita4	A1	
Cita5	A1	
Cita6	A1	
Cita7	A1	
Cita8	A1	

Screen Subfile

Nr_scr	N	Health data screening (98 only)
Key	N	
Ssubcode	A4	Indicates the name and OSHA IMIS code of the substance sampled
Ssamptyp	A1	Tpe of sample taken; p=personal; a=area; b=bulk; w=wipe; s=screen; l=blood; u=urine
Ssamp_nr	N	Nos. of samples taken per type
Sexplvl	N	Level of exposure
Sunits	A1	Units of exposure
Sev_rnge	A1	Indicates the severity within range of the sampling by comparing the actual results to the OSHA Permissible Exposure Level (PEL) 0-.49: results between 0 and .49 times std .5-.99 1-1.59 1.6-1.99 2-3 3 ??? exposure data exists but is insufficient to compute the severity

Hazardous Substance Subfile

Nr_hz	N	Number of hazardous substances
Key	N	Identification key for citation substances apply to

Cit11_nr	N	Citation number
Item1_nr	N	Violation item number
Item1_gr	A2	Grouped violation flag
Hazsub1	A4	Substance code(s) for all substances directly
Hazsub2	A4	related to item cited. Up to 5 codes may be
Hazsub3	A4	entered if more than 5 substances are present;
Hazsub4	A4	the 5 with the highest exposure levels are
Hazsub5	A4	listed

Violation Subfile

Nr_viol	N	Number of violation blocks
Key	N	Violation id. key - unique for an inspection
Del_flag	A1	Violation deleted for reason other than error in submission of data.
Isu_date	N	Date citation issued by Area Director
Cit2_nr	N	Citation number
Item2_nr	N	Violation item number
Item2_gr	A2	Group grouped violation code bb = non-grouped violation Ab = primary item of grouped violation Bb = Zb = Members of grouped violation
Violtype	A1	Type of violation (reflects any changes or modifications); s=serious; w=willful; u=unclassified; r=repeat; o=other
Std	A22	22-digit standard cited
Nr_instances		Number of instances of violation of standard

APPENDIX IIA

```
//EPID011C JOB (X,CMAA,REG),MATTHEWS,
// CLASS=M,TIME=1,NOTIFY=EPID011,MSGCLASS=T
//*ROUTE PRINT RMT11
/*JOBPARM L=99
/*SETUP
*****
/*SETUP INPUT - 353505 (TAPE IS LABELED OSHA2)
/*SETUP
*****
// EXEC SAS
//FT11F001 DD SYSOUT=*
//FT12F001 DD SYSOUT=*
//IN DD DSN='EPID011.Q30434B.DRVDHLTH',UNIT=TAPE,VOL=SER=353505,
// LABEL=(1,SL),DISP=OLD
//OUT1 DD DSN='EPID011.LTAS.OSHA.SITE.SAS',DISP=OLD
//OUT1A DD DSN='EPID011.LTAS.OSHA.SITE.DATA',DISP=OLD
//*
//OUT2 DD DSN='EPID011.LTAS.OSHA.HAZSUB.SAS',DISP=(NEW,CATLG),
// UNIT=SYSDA,SPACE=(TRK,(20,5),RLSE)
//OUT2A DD DSN='EPID011.LTAS.OSHA.HAZSUB.DATA',DISP=(NEW,CATLG),
//
DCB=(LRECL=80,BLKSIZE=23440,RECFM=FB),SPACE=(TRK,(20,5),RLSE),
/ UNIT=SYSDA
//*
//OUT3 DD DSN='EPID011.LTAS.OSHA.VIOL.SAS',DISP=(NEW,CATLG),
// UNIT=SYSDA,SPACE=(TRK,(20,5),RLSE)
//OUT3A DD DSN='EPID011.LTAS.OSHA.VIOL.DATA',DISP=(NEW,CATLG),
//
DCB=(LRECL=80,BLKSIZE=23440,RECFM=FB),SPACE=(TRK,(20,5),RLSE),
/ UNIT=SYSDA
//*
//OUT4 DD DSN='EPID011.LTAS.OSHA.NOSCR.SAS',DISP=(NEW,CATLG),
// UNIT=SYSDA,SPACE=(TRK,(20,5),RLSE)
//OUT4A DD DSN='EPID011.LTAS.OSHA.NOSCR.DATA',DISP=(NEW,CATLG),
//
DCB=(LRECL=80,BLKSIZE=23440,RECFM=FB),SPACE=(TRK,(20,5),RLSE),
/ UNIT=SYSDA
//*
//OUT5 DD DSN='EPID011.LTAS.OSHA.SCR.SAS',DISP=(NEW,CATLG),
// UNIT=SYSDA,SPACE=(TRK,(20,5),RLSE)
//OUT5A DD DSN='EPID011.LTAS.OSHA.SCR.DATA',DISP=(NEW,CATLG),
//
DCB=(LRECL=80,BLKSIZE=23440,RECFM=FB),SPACE=(TRK,(20,5),RLSE),
/ UNIT=SYSDA
```

APPENDIX IIB

```
DATA OUT1.SITE(DROP=
    CIT1_NR ITEM1_NR ITEM1_GP HAZSUB1 HAZSUB2
    DEL_FLAG ISU_DATE CIT2_NR ITEM2_NR ITEM2_GP
    VIOLTYPE STD_HAZSUB3 HAZSUB4 HAZSUB5
    SUBCODE ANAL_REQ SAMPTYPE EXP3TYPE EXPLEVEL
    UNITS PEL ADJ_PEL SEVERITY CITA1 CITA2 CITA3 CITA4
    CITA5 CITA6 CITA7 CITA8
    SSUBCODE SSAMPTYP SSAMP_NR SEXPLVL SUNITS SEV_RNGE)

OUT2.HAZSUB(KEEP=OBS KEY NR_HAZ ITEM1_NR ITEM1_GP
    HAZSUB1-HAZSUB5 CIT1_NR)
OUT3.VIOL_REC(KEEP=OBS KEY DEL_FLAG ISU_DATE CIT2_NR
    ITEM2_NR ITEM2_GP VIOLTYPE STD NR_VIOLS)
OUT4.NO_SCR(KEEP=OBS KEY SUBCODE ANAL_REQ SAMPTYPE EXP3TYPE
    UNITS PEL ADJ_PEL SEVERITY CITA1 CITA2 CITA3 CITA4
    CITA5 CITA6 CITA7 CITA8 NR_NOSCR EXPLEVEL)
OUT5.SCREEN(KEEP=OBS KEY SSUBCODE SSAMPTYP SSAMP_NR SEXPLVL
    SUNITS SEV_RNGE NR_SCR);
```

APPENDIX IIC

INFILE IN OBS=100 MISSEVER;

INPUT ACT NR	IB4.
REGION	2.
AREACODE	3.
OFC_CODE	2.
SIC	IB2.
ST_FLAG	\$1.
CSHO_ID	\$5.
NAME	\$30.
ADDR	\$30.
STATE	\$2.
ZIP	PD3.
CITYCODE	IB2.
CNTYCODE	PD2.
OWNER_TP	\$1.
OWNER_CD	IB2.
ADV_NOTC	\$1.
OPENDATE	PD4.
CLSEDATE	PD4.
S_H	\$1.
INSPTYPE	\$1.
INSP_SCP	\$1.
NR_ESTAB	PD3.
NR_COVER	PD3.
NR_EMPR	PD4.
WALK	\$1.
EMP_INT	\$1.
UNION	\$1.
CAS_CLOS	PD4.
PEN_DATE	PD4.
FTA_DATE	PD4.
DUR_SRC	\$1.
PA_HOURS	IB4.
CONTDATE	PD4.
REMITAMT	PD6.2
FTA_AMT	PD6.2
TOT_PEN	PD6.2
TOT_FTA	PD6.2
TOT_VIOL	PD3.
TOTSVIOL	PD3.
HFORMTYP	\$1.
HSAMP_NR	IB4.
HSAMP_DT	PD4.
HSAMP_CS	\$5.
HJOBDESC	\$20.
HNR_EXP	PD3.
HEXPS_FQ	\$54.
HASHDATE	PD4.
HARCDATE	PD4.

HNE_DBA PD3.
NR_HAZ IB2.
NR_VIOLS IB2.
NR_NOSCR IB2.
NR_SCR IB2.
NR_TOTSG IB2. @;

KEY = _N;
OUTPUT OUT1.SITE;

DO I=1 TO NR_HAZ;
INPUT CIT1_NR 2.
ITEM1_NR 3.
ITEM1_GP \$2.
(HAZSUB1-HAZSUB5) (\$4.) @;
OUTPUT OUT2.HAZSUB;
END;

DO I=1 TO NR_VIOLS;
INPUT DEL_FLAG \$1.
ISU_DATE PD4.
CIT2_NR 2.
ITEM2_NR 3.
ITEM2_GP \$2.
VIOLTYPE \$1.
STD \$22. @;
OUTPUT OUT3.VIOL_REC;
END;

DO I=1 TO NR_NOSCR;
INPUT SUBCODE \$4.
ANAL_REQ \$1.
SAMPTYPE \$1.
EXP3TYPE \$1.
EXPLEVEL PD7.7
UNITS \$1.
PEL PD7.7
ADJ_PEL \$1.
SEVERITY PD7.7
(CITA1-CITA8) (\$1.) @;
OUTPUT OUT4.NO_SCR;
END;

DO I=1 TO NR_SCR;
INPUT SSUBCODE \$4.
SSAMPTYP \$1.
SSAMP_NR 3.
SEXPLVL PD7.7
SUNITS \$1.
SEV_RNGE \$1. @;

OUTPUT OUT5.SCREEN;

END;

DROP I;

APPENDIX IID

```
PROC PRINT DATA=OUT1.SITE      ;TITLE 'SITE DATA';
PROC PRINT DATA=OUT2.HAZSUB    ;TITLE 'HAZSUB DATA';
PROC PRINT DATA=OUT3.VIOL_REC  ;TITLE 'VIOL_REC DATA';
PROC PRINT DATA=OUT4.NO_SCR    ;TITLE 'NO_SCR DATA';
PROC PRINT DATA=OUT5.SCREEN    ;TITLE 'SCREEN DATA';
```

```
PROC XCOPY INDD=OUT1 OUTDD=OUT1A EXPORT;
PROC XCOPY INDD=OUT2 OUTDD=OUT2A EXPORT;
PROC XCOPY INDD=OUT3 OUTDD=OUT3A EXPORT;
PROC XCOPY INDD=OUT4 OUTDD=OUT4A EXPORT;
PROC XCOPY INDD=OUT5 OUTDD=OUT5A EXPORT;
//
```

APPENDIX III

```
* Chgdate.prg
*
* PRogram to convert from numeric date to legal date
*
* Terrell Herzig
* Program stores new date to date var.
*
use temp
do while .not. eof()
  nd=alltrim(str(hsamp_Dt))
  yr=substr(nd,1,2)
  mm=substr(nd,3,2)
  dd=substr(nd,5,2)
  sdte=ctod(mm+"/"+dd+"/"+yr)
  replace sampdte with sdte
skip
enddo
return
```

* EX1_A.prg

* Program to group by consecutive days

* Read all unique hjobdesc codes into an array

public nacnt

use temp

index on name+hjobdesc to temp unique

copy to a fields name,hjobdesc for .t.

close all

* Array created and filled with unique hjobdesc

* create new files with unique file names

* and load with seq. dates. Remember

* these files by saving file handle in

* array FNAMES.

select 1

use a

index on name+hjobdesc to a

select 2

use temp

index on name+hjobdesc to temp

select 3

use fnames

select 1

do while .not. eof()

store name+hjobdesc to job

select 2

seek job

fhandle=sys(3)

copy to &fhandle for name+hjobdesc=job

select 3

append blank

replace parent with fhandle

replace child with "MASTER"

select 1

skip

enddo

Examine MASTER groups for less than 2 obs and delete.

close all

select 1

use fnames

do while .not. eof()

store parent to fpoint

select 2

use &fpoint

if reccount()<2

store alltrim(fpoint)+'.dbf' to fpoint

use

erase &fpoint

select 1

delete

endif

select 1

skip

enddo

back

close all

*

* We now have groups broken out by job codes

* Now we must test for consecutive days, and subgroup.

*

close all

select 3

use fname2

select 1

use fnames

do while .not. eof()

store parent to fpoint

select 2

use &fpoint

index on sampdte to dtemp

store sampdte to bp

skip

do while .not. eof()

if sampdte<=bp+5

else

* write recs between pointers.

do writeobs

endif

skip

enddo

* Write last rec

do lastrec

select 1

skip

enddo

return

procedure writeobs

store sampdte to ep

nf=sys(3)

copy to &nf for sampdte>=bp .and. sampdte<ep

select 3

append blank

replace child with nf

replace parent with fpoint

select 2

seek ep

bp=ep

return

procedure lastrec

nf=sys(3)

copy to &nf for sampdte>=bp

select 3

append blank

replace child with nf

replace parent with fpoint

select 2

return

SIFTGRPS.PRG

Terrell Herzig

Program examines subgroups and removes invalid groups

```
select 1
use fname2
do while .not. eof()
store child to fpoint
select 2
use &fpoint
if reccount()<2
select 1
delete
select 2
fpoint=alltrim(fpoint)+'.dbf'
use
erase &fpoint
else
count to zero for explevel=0.00
if zero/reccount()>.25
select 1
delete
select 2
fpoint=alltrim(fpoint)+'.dbf'
use
erase &fpoint
endif
endif
select 1
ip
ddo
e fname2
ck
```

GET RID OF MASTER WITH 0.00 > 25%

```
lect 1
e fnames
while .not. eof()
store parent to fpoint
select 2
use &fpoint
if reccount()<3
select 1
delete
select 2
fpoint=alltrim(fpoint)+'.dbf'
use
erase &fpoint
else
count to zero for explevel=0.00
if zero/reccount()>.25
select 1
delete
select 2
fpoint=alltrim(fpoint)+'.dbf'
use
erase &fpoint
```

```
endif  
endif  
select 1  
kip  
nddo  
se fnames  
ack  
eturn
```

Rplzero.prg

Terrell Herzig

This program replaces zero value exposures with
a nonzero value.

```
select 1
use fnames
do while .not. eof()
    store parent to fpoint
    select 2
    use &fpoint
    replace explevel with .006 for explevel=0
    select 1
```

skip
addo

repeat for each subtable in group

```
select 1
use fname2
do while .not. eof()
    store child to fpoint
    select 2
    use &fpoint
    replace explevel with .006 for explevel=0
    select 1
```

skip
addo
return

Complog.prg

Terrell Herzig

Computes natural log for each obs.

```
lect 1
e fnames
  while .not. eof()
store parent to fpoint
select 2
use &fpoint
replace log with log(explevel) for .t.
select 1
ip
ddo
```

Now do each subgroup

```
lect 1
e fname2
  while .not. eof()
store child to fpoint
select 2
use &fpoint
replace log with log(explevel) for .t.
select 1
ip
ddo
turn
```


Bldlog.prg

Terrell Herzig

```
store "COKEOVEN" to msub
select 5
use log
select 1
use fnames
do while .not. eof()
store parent to f1
select 2
use &f1
store name to mname
store hjobdesc to mjob
CALCULATE STD(log), VAR(log), AVG(log) TO mstd, mvar, mavg
store reccount() to mn
select 5
append blank
replace name with mname
replace hjobdesc with mjob
replace parent with f1
replace child with "MASTER"
replace n_log with mn
replace std_log with mstd
replace avg_log with mavg
replace var_log with mvar
replace substance with msub
select 1
skip
addo
```

Now do for subtables

```
select 5
use log
select 1
use fname2
do while .not. eof()
store child to f1
store parent to mpar
select 2
use &f1
store name to mname
store hjobdesc to mjob
CALCULATE STD(log), VAR(log), AVG(log) TO mstd, mvar, mavg
store reccount() to mn
select 5
append blank
replace name with mname
replace hjobdesc with mjob
replace parent with mpar
replace child with f1
replace n_log with mn
replace std_log with mstd
replace avg_log with mavg
replace var_log with mvar
replace substance with msub
select 1
```

tip ,
ddo
eturn