Adjusting for the Effect of Environmental Variability in Outdoor Engineering Control Studies; Detailed Derivations

September 2003

Stanley A. Shulman, R. Leroy Mickelsen, Kenneth R. Mead

Keywords: Factor Analysis, Measurement Error, Randomized Blocks

Abstract

In outdoor engineering control studies, measured analyte concentration levels change as environmental conditions change. Since the effectiveness of an engineering control varies with environmental conditions, comparisons of measurements taken with engineering controls operating versus those taken in an uncontrolled environment should adjust for these changes. However, environmental parameters are difficult to estimate. In this work, models based on factor analysis (Fuller [1]) are used to account for the effect of environmental variables. These models also describe the phenomenon that greater control efficiency tends to occur at the highest levels of the uncontrolled environment. (Shulman, Mead, and Mickelsen[2]). The approach is combined with the randomized pair (uncontrolled environment determination, engineering control determination) approach that is often used in these studies. Also investigated are the benefits of taking samples at different locations and of different analytes. Results of the factor analysis models are compared with those from regressions of the log ratio (controlled/uncontrolled) on the uncontrolled determination. Implications for statistical design are also discussed. Results from the example data set indicate that the factor analytic approach can identify a common factor, and thereby, provide evidence that the common factor is due to environment. However, the simpler regression approach provides estimated reductions that are also statistically unbiased. Therefore, the factor analytic approach may be most useful in early stages of a study, to assess common environmental effects, after which the simpler regression approach may be used to obtain estimates of control effectiveness.

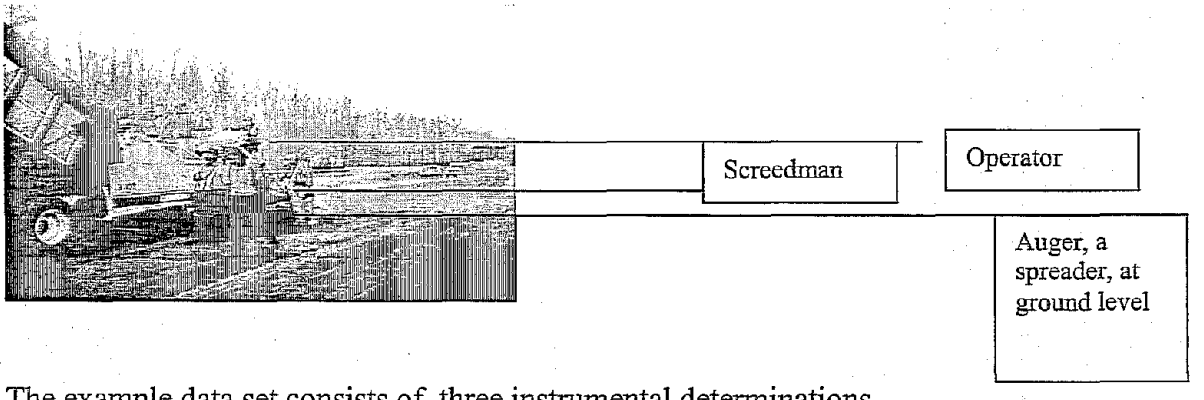
1) Introduction

Engineering control studies evaluate the effectiveness of the controls, estimation of which is affected by both environment and background levels. In indoor studies, where environmental variation can often be minimized, researchers collect data on the background analyte concentration levels, which may tend to increase over a day's work. If background determinations are available, then the effect of the increase in background can often be removed by subtracting the current background from the current analyte concentration. In outdoor studies, the possibly varying background is difficult to estimate. (Note: The physical boundaries that define the background may also vary with time.) In addition, the environment interacts with the ventilation system and causes high variability in measurements. In this presentation the aim will be to construct models that take environmental variation into account. The models suggested here use multiple simultaneous determinations to separate the environmental effect from the engineering

control effect. This approach will be compared to a regression method in which the natural log of the ratio (controlled determination/uncontrolled determination) is regressed on the uncontrolled (Shulman, et. al.[2]). In the rest of this paper, the uncontrolled determination will usually be referred to as "control-off 'or "off" and the controlled determination as "control-on" or "on." Suggestions for statistical design of outdoor studies are also included.

The work presented here continues that in Shulman, et. al.[2] The factor analytic approach presented here relies mainly on that given in Fuller [1]. The use of structural equation models and the presentation of material has benefited from that given in Krieg, et. al.[3].

2) Example Data



The example data set consists of three instrumental determinations taken simultaneously during an asphalt paving study. These include the following:

a) Measurements of airborne organics at the auger (org_aug) (Figure 1)
b) Measurements of airborne organics away from the auger, near the paver operator or screed workers (org_naug) (Figure 2)
c) Measurements of airborne particulate at the auger (part_aug) (Figure 3)

All three figures have the same vertical axis scale. The two organics measurements have the same horizontal axis scale. Outliers were identified by the criterion that the standardized residual in the regression of Ln(on/off) versus Ln(off), done individually for each of the three instruments (a,b,c), exceed 3. This statistical criterion was used in removing one sample set.

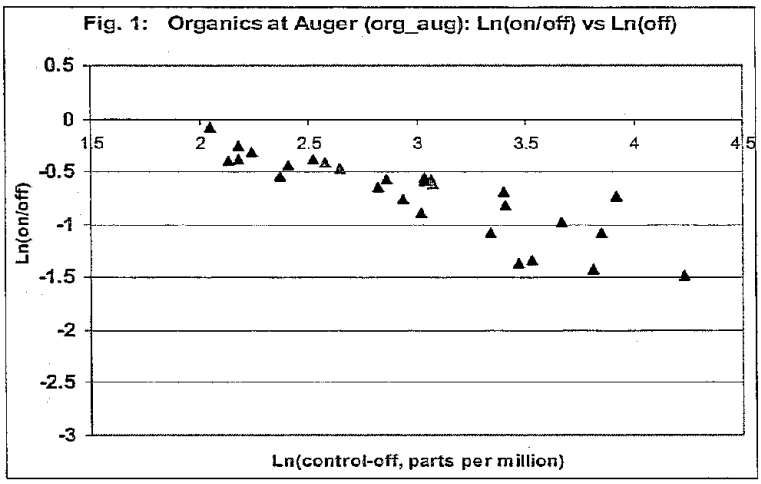


Fig. 1: Organics at Auger (org_aug): Ln(on/off) vs Ln(off)
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
1.5
2
2.5
3
3.5
4
4.5
Ln(on/off)
Ln(control-off, parts per million)

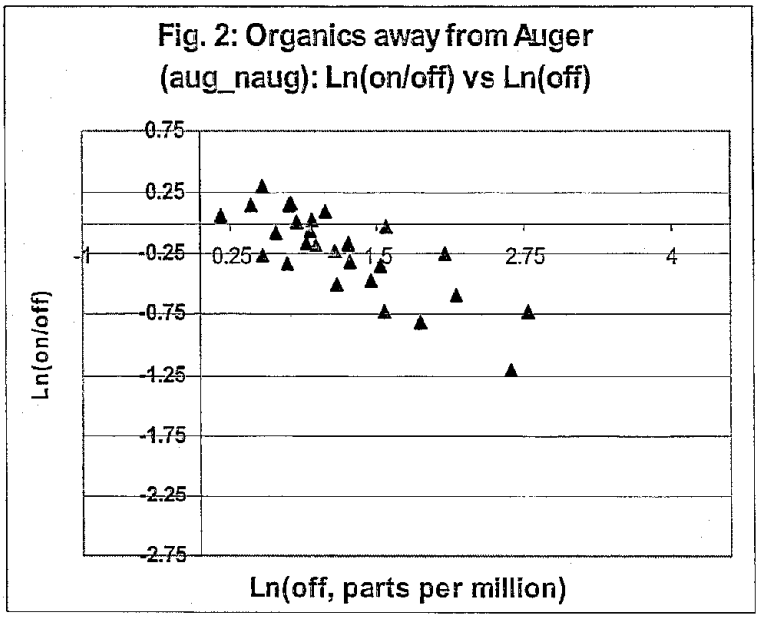


Fig. 2: Organics away from Auger (aug_naug): Ln(on/off) vs Ln(off)
0.75
0.25
-0.25
-0.75
-1.25
-1.75
-2.25
-2.75
-1
0.25
1.5
2.75
4
Ln(on/off)
Ln(off, parts per million)

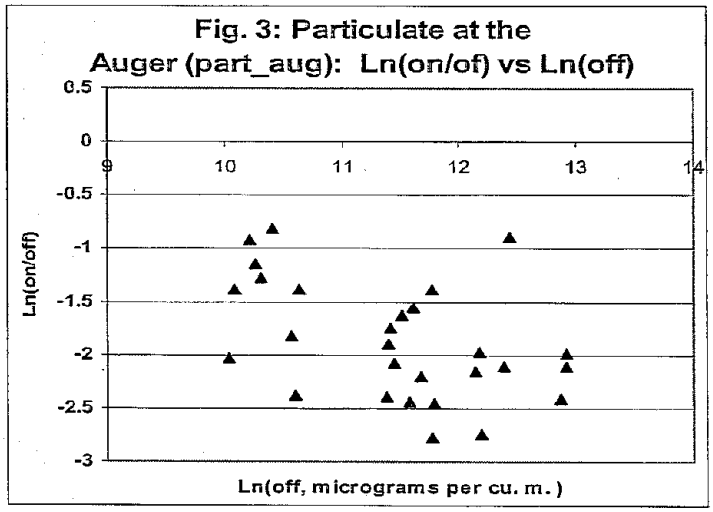


Fig. 3: Particulate at the Auger (part_aug): Ln(on/of) vs Ln(off)
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
9
10
11
12
13
14
Ln(on/off)
Ln(off, micrograms per cu. m.)

For each of these three sets of real-time data, there are measurements with control-on (taken simultaneously: org_aug_on, org_naug_on, part_aug_on) and with control-off (taken simultaneously: org_aug_off, org_naug_off, part_aug_off). Thus, the full set consists of six series of measurements. Although these were all real-time determinations, they were not identical. Due to limitations in the instrument sampling frequency, the airborne organics measurements were taken every four seconds and the airborne particulate measurements at the auger every six seconds. Data were collected in randomized pairs (control-on, control-off) over five days of sampling. Data resulting from one run at a fixed control condition are called a "trial," each of which lasted at least 1 ½ minutes. Thus, each of the pairs consisted of two trials. Twenty-eight trials were included in this analysis. Medians were computed for each trial for measurements a, b, and c because medians are much less correlated than individual readings. Deletion of data collected near transitions in paving status also reduced correlation. Also the use of medians minimized the presence of possible outliers. The pairs themselves were treated as a random sample of pairs from that day. (This is a simplification, since pairs were often taken in groups, depending on when paving was being done. Thus, several pairs were often taken in succession, followed by no sampling for a while because paving stopped.) The data may be viewed as coming from a split plot, where the whole plot is the control setting, and the subplots are the instrument type or location.

3) Background and Environmental Effects

Figures 1, 2, and 3 indicate that the ratio (on/off) decreases with increasing control-off determinations (because the increase in control-on is not as great as that in control-off.) Thus, the fraction reduction [1-(on/off)] increases with increasing control-off. This may relate to environmental variability or to varying background.

Background is the level of the analytes present when there is none produced by the specific work process under evaluation. (Note that adjacent work processes can contribute to background levels at the engineering control.) If on/off is the ratio of the control-on to the uncontrolled environment, and if there is a background of level $B > 0$ and if on<off, then $(on-B)/(off-B) < on/off$. Thus, adjustment for the background B would identify larger reductions due to control. However, to make a substantial difference, the value of B must be quite close to the minimum value of the control-on concentrations. Estimation is difficult and will not be attempted here.

If the uncontrolled environment is thought to represent the environmental variables, then the interpretation of Figure 1 is that the reduction due to control is highest when the environmental control is least and lowest when the environmental control is greatest. When contaminant levels in the uncontrolled environment are high, the environment provides little control of contaminants, and engineering control effectiveness might be higher than when the environment itself is a control.

4) Statistical Models

The models discussed here all generally assume normally distributed data.

Data for the above examples were collected in randomized pairs (control, uncontrolled) over a five day period. The simplest kind of model (Shulman, et. al. [2]) to use is a regression of the

ln (control-on)-ln(uncontrolled) determination versus the ln(uncontrolled):

$$\mathrm{Ln}(y_{p,on,s} / \ y_{p,of,s}) = \alpha + \beta \ \mathrm{Ln}(y_{p,of,s}) + e_{p,s} \qquad (1)$$

for pair p, control setting c ("on" =control-on, "of"= control-off) and sample location and/or type s. Fixed effects are denoted by Greek letters, and random effects by Latin letters. Variance $(e_{p,s})=\sigma_{ps}^2$ and $E(e_{p,s})=0$. The model (1) can be used to predict the ratio of controlled to uncontrolled environment by using least squares estimates of model parameters. It leaves unclear the degree to which the uncontrolled determination can be called the environmental factor. Since data from each instrument are treated separately, model (1) does not provide an estimate for the effect of a common environmental factor. Model (1) applies either under bivariate normality of (on, off), or under conditional normality of the Ln(on/off) for given off values.

An appropriate split-plot model (Cochran and Cox[4]) for the example data is:

$$\mathrm{Ln}(y_{p,c,s}) = \mu + b_p + \alpha_c + \ b\alpha_{p,c} + \gamma_s + \alpha\gamma_{c,s} + \ e_{p,c,s}, \qquad (2)$$

$\mathrm{Variance}(b_p)=\sigma_p^2$, $\mathrm{Variance}(b\alpha_{p,c})=\sigma_{pc}^2$, and $\mathrm{Variance}(e_{p,c,s})=\sigma_{cs}^2$

Each of the random components is assumed to be normally distributed with expectation 0 and to be a random sample from an infinite population. Although this kind of model allows for estimates of the variability across pairs, it does not quantify how the ratio will vary with the level of the uncontrolled environment. For an environmental variable x_{env},

$$\mathrm{Ln}(y_{p,c,s}) = \lambda_{c,s,0} + \lambda_{c,s,1} \ \mathrm{Ln}(x_{env,p}) + \ e_{p,c,s} \ , \qquad (3)$$

where the random components of (2) have been replaced by variables that indicate the linear dependence on a common random factor $x_{env.}$ $E(e_{p,c,s})=0$. If there were a second environmental factor, this could be included in (3) as an addend $\lambda_{c,s,2} \ \mathrm{Ln}(x_{env2,p})$. If there were k environmental factors these would appear as additional additive terms in (3) with slopes $\lambda_{c,s,j}$, j=1,2,..,k. This model is similar to a model that is useful in comparison of measurement methods (Krieg, et. al.[3]). For the example data, (3) consists of six equations for the six variables: org_aug_off, org_aug_on, org_naug_off, org_naug_on, part_aug_off, part_aug_on. Differencing the contolled and uncontrolled results for the same sample type s yields:

$$\mathrm{Ln}(y_{p,on,s})-\mathrm{Ln}(y_{p,off,s}) = (\lambda_{on,s,0} - \lambda_{off,s,0}) + (\lambda_{on,s,1} - \lambda_{off,s,1}) \ \mathrm{Ln}(x_{env,p}) + \ (e_{p,on,s} - e_{p,off,s} \ ). \qquad (4)$$

Thus, the difference between two levels of control would not be constant but would depend on the value(s) of the environmental variable, as long as the multiplier of the environmental variable is not zero. This will be studied in the example data.

The difficulty is that in many situations it is not possible to obtain good measures of the environmental variables or to even identify all of them. There is no reason to think that the environmental controls can be summarized in one variable. However, there is a statistical technique, factor analysis, which enables the estimation of the parameters of the models (3), and which allows assessment of the adequacy of a single factor (or of multiple factors). Assumptions for this factor model are:

1) Normal distribution of the data
2) Results from different pairs must be statistically independent
3) Responses are linear in the factors

For the example data the model (3) corresponds to six equations since there were three sample type-location combinations. The covariance matrix for any set of six measurements ($Ln(y_{p,c,s})$) in the same pair p has the form:

$$\Sigma_y = \Lambda \Lambda' + \Sigma_e, \qquad (5)$$

The components of Λ (the $\lambda_{c,s,j}$ of eq. (3)) are called the factor loadings, and Σ_e is a diagonal matrix, the diagonal elements of which are called specificities (the variances of the $e_{p,c,s}$ of eq. (3), discussed in Morrison[5]). Λ is not uniquely determined, since for any orthogonal matrix $\mathbf{P}$, $\mathbf{\Lambda P P' \Lambda' = \Lambda \Lambda'}$. Corresponding to the model (3), Λ may have just one column but the adequacy of just one factor can be tested. The components of Λ and Σ_e can be estimated by maximum likelihood (Fuller[6]), and these estimates can be shown to converge asymptotically to the true values. The factors are treated here as random with mean 0 and variance 1. (Alternatively, the factors can be treated as fixed, the f_is are assumed to be scaled so that the [1/(n-1)]x (corrected sum of squares) is 1, and the sample average is 0. However, the same maximum likelihood estimates can be used in either random or fixed factor model, as discussed in Anderson[7].) Besides estimating Λ, the predicted values can also be estimated. The value of this method for data of the form in the example is that the model permits the evaluation of the adequacy of common factors in describing the data. This kind of evaluation is not possible for model (1), even though that model does permit the prediction of the ratio on/off given the measurement of the uncontrolled environment.

What meaning is to be associated with these factors? Suppose there is just one factor. The factor is not unique, and can be expressed as a linear combination of the observations. However, the predicted values under the model do not change, even if a different linear combination is used for the factor.

A useful way to compare the models is to determine the variances of the differences between the predictions and the true Ln(on/off) ratios under either models (1) or (4). For model (1), the predictions are unbiased, if the linear relationship is true. Although usually the variances associated with model (1) are provided for given values of the explanatory variable, what is of interest here is the expectation of the squared difference between the prediction and the true natural log of the ratio of control-on to control-off (Appendix 1, Part A):

$$E\{[\alpha + \beta\ Ln(y_{p,of,s})] - [Ln(y_{p,on,s,tr}) - Ln(y_{p,off,s,tr})]\}^2 =$$
$$Var[Ln(y_{p,on,s,tr})] + \beta^2 Var[Ln(y_{p,off,s,tr})] + (\beta-1)^2 Var[Ln(y_{p,off,s}) - Ln(y_{p,off,s,tr})]$$
$$-2\beta Cov[Ln(y_{p,off,s}), Ln(y_{p,on,s})], \qquad (6)$$

where β is the slope of eq(1), "Var" is the variance of the designated variable, and "Cov" is the covariance of the designated variables, and "tr" designates the value of the subscripted variable excluding the error $e_{p,s}$ of eq(1) .

For the factor model, as in Fuller[8] , the variance can be shown to be (Appendix 1, Part A)

$$Var(\ Ln(y_{pred,\ fact}) - Ln(y_{true})) = \mathbf{\Lambda}\ (\mathbf{I}_k + \mathbf{\Lambda}'\ \Sigma_e^{-1}\ \mathbf{\Lambda}\ )^{-1}\ \mathbf{\Lambda}', \qquad (7)$$

where y_{true} is the random environmental determination excluding measurement error; $\mathbf{\Lambda}$ is given in equation (5); k is the number of factors in the model and $\mathbf{I}_k$ is the k-dimensional identity matrix.

As in (4), interest here is in the comparison of the difference of predictions in the factor model. Thus, we require the variance of the difference between the predictions and the difference of the true values. In our example, **c** could be a six column row vector, all elements 0, except for 1 and -1 in the first two positions for a comparison of org_aug_on and org_aug_off. Pre and post-multiplying (7) by **c** and **c'**, respectively, would yield the variance of the difference between the difference of two predicted values and the difference of the corresponding true values. In symbols,

$$Var\{\ [Ln(y_{pred,on,\ fact}) - Ln(y_{pred,off,\ fact})] - [Ln(y_{on,true\ /\ }y_{off,true})]\} = \mathbf{c}\ \mathbf{\Lambda}\ (\mathbf{I}_k + \mathbf{\Lambda}`\ \Sigma_e^{-1}\ \mathbf{\Lambda}\ )^{-1}\ \mathbf{\Lambda}`\mathbf{c}` \quad (8)$$

The factor model used here treats the pairs as a random sample, and thus the $y_{on,\ true}$ and $y_{off,\ true}$ values are random variables, as are the factors of (5). As was mentioned above, it is possible to treat factors as fixed, rather than random, but random seems more appropriate in this sampling situation. Because of the multivariate normality, the variances in (7) and (8) are constant, regardless of the observed concentration levels. Normality for the factor model is an important assumption.

For models (6), (7), and (8) sample estimates will be used in place of parameter values to obtain the estimated variances.

Although the above explanation of the factor analytic model is somewhat complicated, the aim of the model can be described simply as follows: identification of statistically independent factors which explain a substantial part of the variability in the data. The observed variables are expressible as linear combinations of these factors. The factors can be estimated, and predictions can be based on the factors. The value of this approach for the example data is that predictions for the ratio (control-on/control-off) can be expressed as functions of these factors. This dependence can aid in the understanding of why the ratio is not constant.

5) Aims

The aim here is the comparison of the factor model (3) results with those of the simple regression model (1). Issues to consider in this comparison are:

a) The effect of measurement error on the slope estimates from these models
b) Comparison of results from the example data
c) Comparison of standard errors from these models

The factor model was applied to the six data sequences of a), b), and c) discussed in section 2. Our main aim was to better understand the trend of the data shown in Figure 1. The particulate at the auger samples were taken next to the organic at the auger samples. Thus, even though they were different analytes, they should have been influenced by similar environmental variables. (This statement may be true for wind but not for temperature, which has more effect on organics, or for location, since, for instance, an adjacent farm field would likely have greater impact upon particulates than organics.) On the other hand, although the organic away from the auger samples were not adjacent to the organic at the auger samples, they were taken by the same kind of instrument. They could be sensitive to exactly the same analytes as the organic at the auger samples.

6) Measurement Error Issues

In equation (1) , $Ln(y_{p,of,s}) = x_{p,nc} + u_{p,nc}$, where $x_{p,nc}$ is an unknown (ln scale) random value for the uncontrolled environment, and $u_{p,nc}$ is the measurement error, with variances σ_x^2 and σ_u^2, respectively. The expected value of the regression slope estimate in (1) is not β, but $\beta\,\sigma_x^2/(\sigma_x^2+\sigma_u^2)$ (Fuller[9]). Thus, if the measurement error variance is 10% of the environmental variance, the estimate from (1) underestimates that from the regression on $x_{p,nc}$ by about 9%. The desirable property of the factor method, is that the estimates (of factor slopes and individual variances) can be shown mathematically to converge to the true value parameter values, provided that the specified model is correct.

However, there is a contrary point of view. Both models (1) and (4) provide unbiased estimates of the reduction due to the control, in spite of the difference in slope estimates. Also, the estimate from (1) is much simpler to produce and understand than that from (4).

There is another consideration. Because the factor model seeks to identify factors that are common to a set of analytes, it can provide an understanding of the effect of a common environmental factor. The simpler regression models cannot identify common factors.

7) Applying the Factor Models to the Example Data

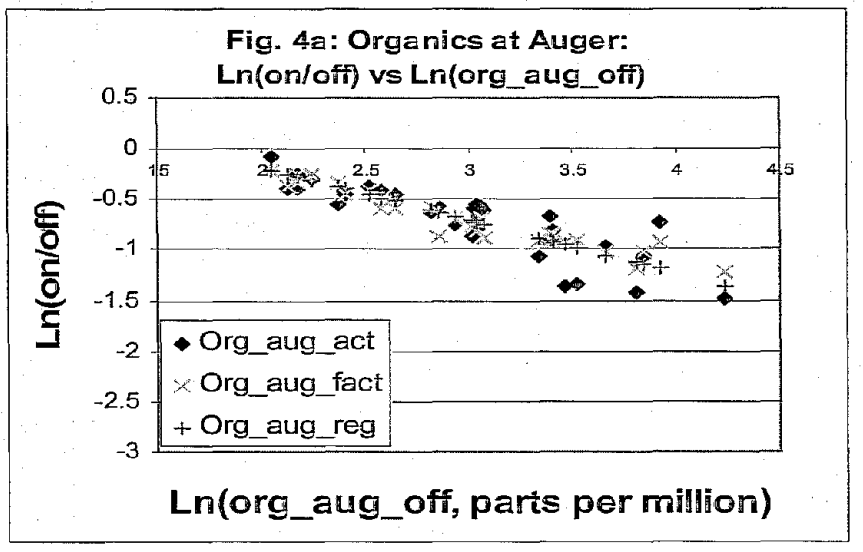
The first issue in applying the factor model to the example data is identifiability. Recall that the factors are parameterized to have mean 0 and variance 1. The 6x6 covariance matrix has 6 variances and 15 covariances, for a total of 21 distinct values. In the one-factor model, there are six factor loadings and six specificities for a total of 12 parameters. These can be shown to be identifiable by expressing the factor loadings and

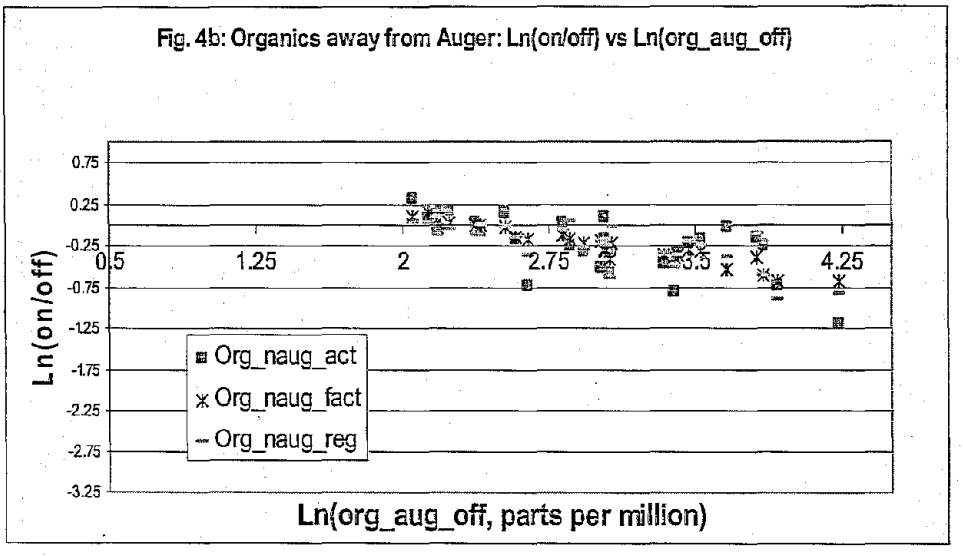
specificities in terms of the variances and covariances. For instance, suppose that the elements of Λ are written as λ_{i1} and the specificities as σ_i^2 , for i=1,2,…,6. (Thus, the notation of (3) is being changed so that (c,s) has six possible pairs, which are denoted by the index i.) Thus, $Var(y_i)= \lambda_{i1}^2 + \sigma_i^2$. Also, $Cov(y_1, y_j)= \lambda_{11} \lambda_{j1}$, j>1. From the covariance expressions, all λ_{j1}, j>1, can be expressed as functions of λ_{11}, and , from the variance expressions, so can the σ_i^2. Since $Cov(y_1, y_3)= \lambda_{11} \lambda_{31}$, λ_{11} is determined. In the model (3) , each $\lambda_{i0}=E(y_i)$, and is identified. Since there are additional covariances, not used in the above, the model is overidentified, and can, therefore, be treated as identified (Bollen[10]).

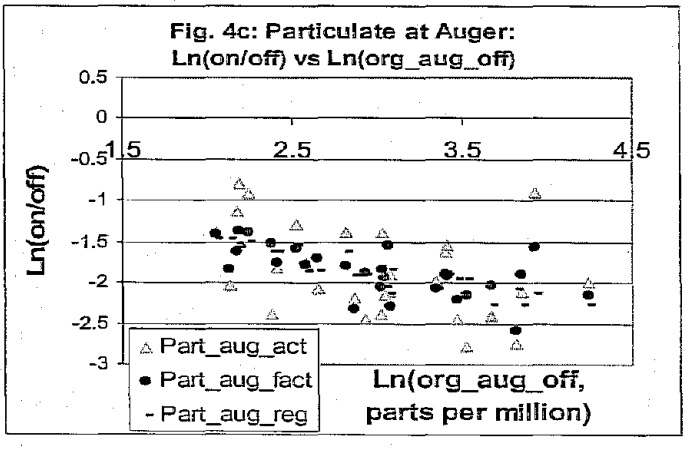
In the two-factor model it is possible to take $\lambda_{12} =0$, by a suitable orthogonal rotation applied to Λ, as was discussed below equation (5). Therefore, for j >2, $Cov(y_1, y_j)= \lambda_{11} \lambda_{j1}$, so that λ_{j1} are functions of λ_{11}. Since $Cov(y_2, y_j)= \lambda_{21} \lambda_{j1} + \lambda_{22} \lambda_{j2}$ for j>2, each λ_{j2} can be expressed in terms of λ_{11} (through λ_{21}) and λ_{22}. Since the equations $\{Cov(y_3, y_j)= \lambda_{31} \lambda_{j1} + \lambda_{32} \lambda_{j2}, j>3\}$ are functions of of λ_{11} and λ_{22} , all elements of Λ are determined. (From $Cov(y_3, y_4)$ λ_{22} can be expressed as a function of λ_{11}^2, whose value can be determined from the relation for $Cov(y_3, y_5)$.) Since $Var(y_j)= \lambda_{j1}^2 + \lambda_{j2}^2 + \sigma_j^2$, the σ_j^2 are also identified. As in the one-factor case, $\lambda_{i0}=E(y_i)$, and all λ_{i0} are identified. Thus, the model is identified.

The example data were analyzed using both Proc Mixed and Proc Calis in SAS [11]. (See Appendix 2 for example SAS code.) The ability of the factor model (4) to explain the variation in the data, relative to the split-plot model (2), can be assessed by statistics such as BIC and AIC. Both the one-factor and two-factor models have smaller values for these statistics than do the split-plot models, suggesting they describe the data better. However, the χ^2 used to test adequacy of fit has p- value of about 0.035. Fitting the three-factor model leads to a statistically significant difference from the two-factor model, though the full three-factor model appears not to be identified, since some specificities have zero estimates under Proc Mixed. Thus, there is some lack of fit of the two-factor model, but, since based on the BIC and AIC statistics it is an improvement over the split plot models the two factor model will be used in the following discussions. By some alternative criteria (comparative fit index and non-normed fit index both exceed 0.95) the two-factor model appears to be acceptable (Hatcher[12]). From residual plots, statistical assumptions of homogeneous variance appear to be met in the two-factor model.

In Figures 4a,b,c, the various Ln (on/off) values are plotted versus Ln (org_aug_off). In the legends, the predicted values from the factor model have "fact" attached to their names. Predicted values from the regression models have "reg" attached to their names, and those with the "act" suffix denote the data. The figures make clear the common dependence on the "off" factor. Because the data are plotted versus the organics at the auger, only the regression results for organics at the auger fall on a straight line. Also, those results are all closest to that line.

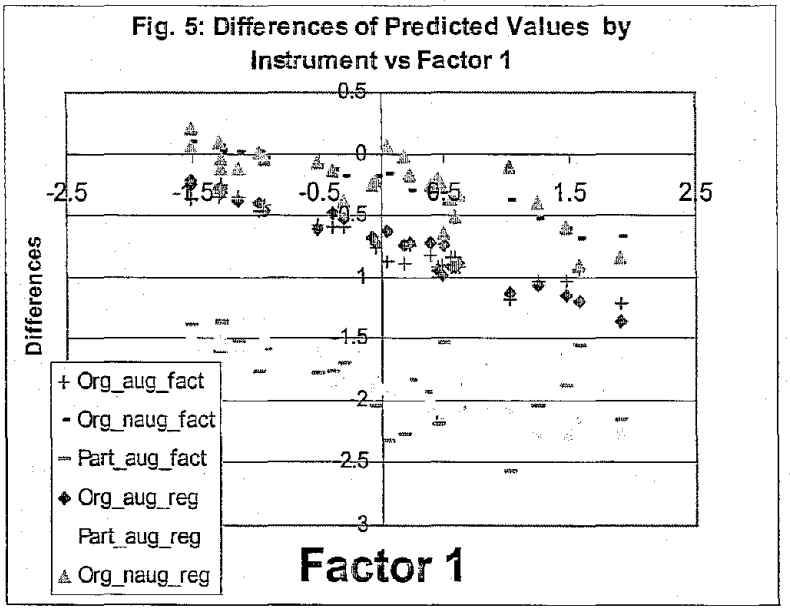


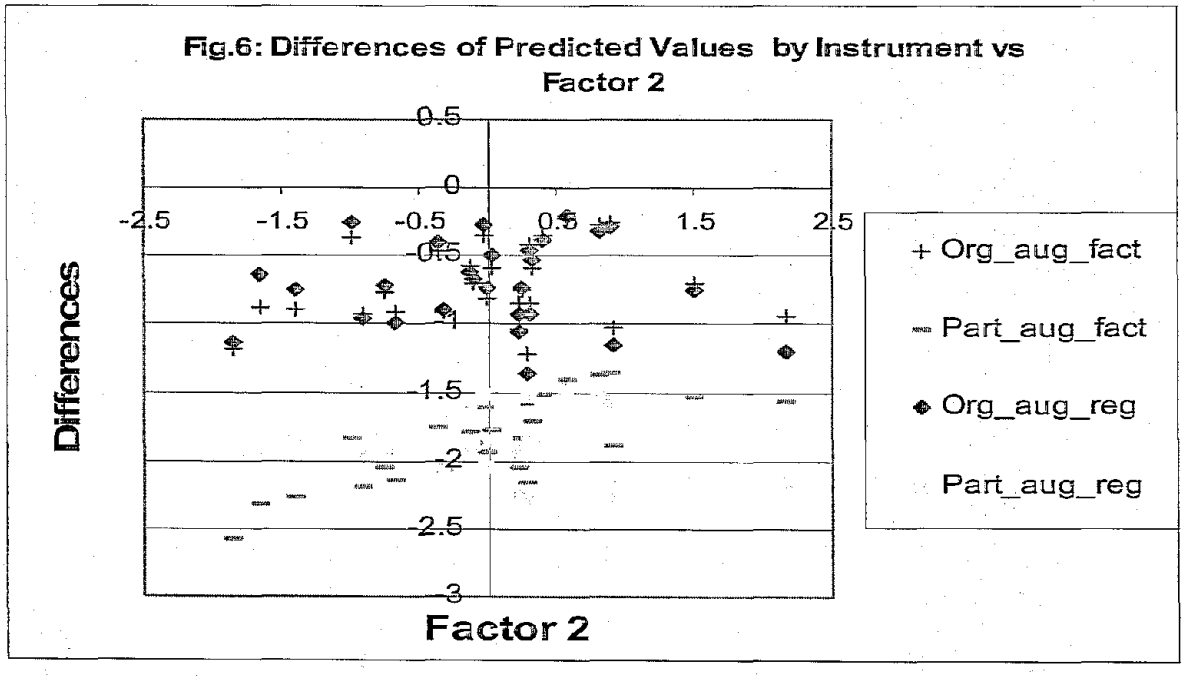


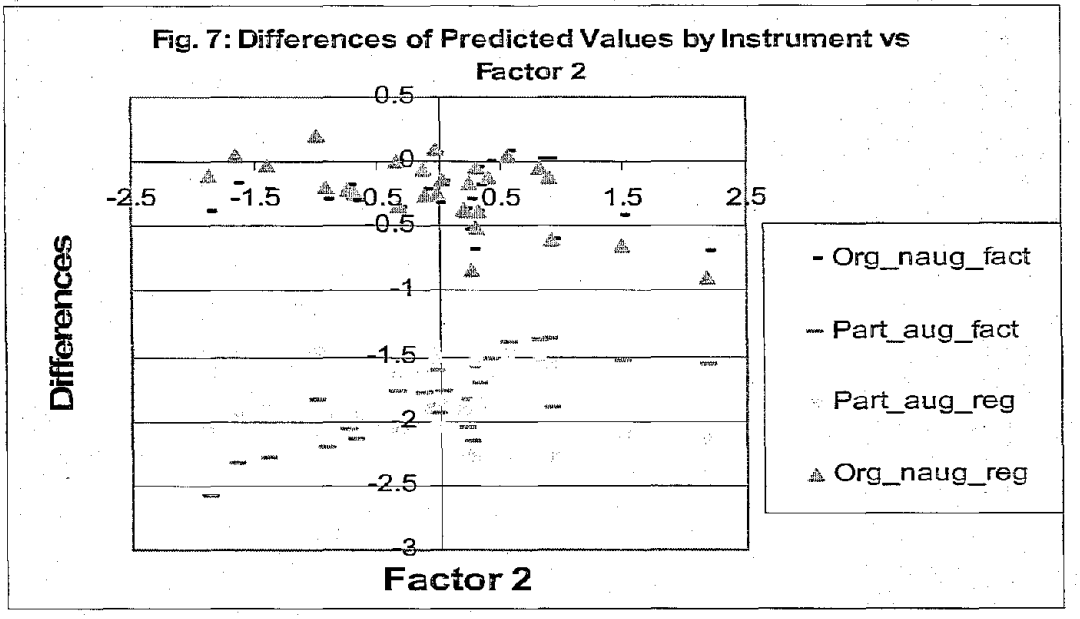


In Figures 5, 6, and 7 the predicted values of the differences (on-off) from the factor analytic model (4) and regression model (1) are plotted versus factors 1 and 2. These factors are linear combinations of the observed variables. Two plots are provided for

factor 2, since the two sets of organics results overlap. It is not possible to say what aspect of the common environment they describe, but the fact that they do provide some adequacy of fit to the data indicates that they allow for some common relationships among the three different measurements. With regard to factor 1, all three instruments have increasingly negative differences (smaller ratio of control-on determinations to uncontrolled determinations) with increasing values of factor 1. Whereas the particulate differences increase as factor 2 increases, the two organics instruments indicate smaller changes.







Another conclusion from Figures 4c and 5 is that the particulate data have somewhat uncertain slope partly because there is considerable variability in the measurements.

An idea of the relative contribution of the two factors to each of the six observed variables is provided by the slope estimates (the λ_{cs1} and λ_{cs2} of model (3)) shown in Table 1. Except for the org_naug_on, the slope estimates for the factor 1 are much larger than for factor 2. Whereas the org_aug data are almost described by factor 1, the other two instrumental determinations require input from factor 2.

Table 1
Slope Estimates (standard errors)* for Factors 1 and 2

	Factor 1	Factor 2
Organics at auger,control off	0.585(0.0887)	0
Organics at auger,control on	0.300(0.0535)	0.0981(0.0419)
Organics away from auger,control off	0.502(0.106)	0.328(0.0808)
Organics away from auger,control on	0.273(0.0756)	0.283(0.0589)
Particulate at auger, control off	0.837(0.132)	0.065(0.110)
Particulate at auger, control on	0.640(0.132)	0.343(0.105)

* Estimated slopes and standard errors from Proc Calis.

It is useful to estimate the fraction of the total variance explained by the two factors, and the fraction not explained. These estimates appear in Table 2. Except for the aug_naug_on results, factor 1 accounts for at least 60% of the total variability in each variable. For no variable is the proportion of variance unexplained greater than 0.20.

Table 2
Proportion of Variance Explained by the 2 Factors and Proportion Unexplained :

	Factor 1	Factor 2	Unexplained
Organics_auger_off	0.919	0	0.081
Organics_auger_on	0.755	0.081	0.165
Organics_nonauger_off	0.617	0.264	0.119
Organics_nonauger_on	0.416	0.446	0.138
Particulate_auger_off	0.880	0.0053	0.115
Particulate_auger_on	0.629	0.181	0.190

When differences are taken, as in equation (4), the results are:

Table 3
Slope Estimates (standard errors) for Model (4)* and Model (1)

	Factor 1	Factor 2	Slopes from model (1)
Organics at auger	-0.285(0.065)	0.0981(0.0419)	-0.321(0.037)
Organics away from auger	-0.229(0.060)	-0.045(0.0691)	-0.273(0.042)
Particulate at auger	-0.197(0.108)	0.278(0.105)	-0.253(0.096)

* Standard errors obtained from asymptotic variance matrix given in Proc Mixed

The estimates in Table 3 make clearer the reasons for the shapes in the figures above. With regard to factor 1, which we have called the common environmental variable, all three instruments yield smaller ratios (on/off) as that variable increases. Lines appear to be approximately parallel. With regard to factor 2, the slopes are much smaller in absolute value for the two organics determinations than the slopes for factor 1. For the particulate the slope for factor 2 is larger than that for factor 1, and indicates an opposite trend. To some extent the two factors can cancel each other's effects for the particulate data.

If the model (1) underestimates slopes in the regression of control-on versus uncontrolled setting, then the results will be more negative slopes when differences are taken. Standard errors tend to be smaller for model (1), perhaps because there is no variability associated with the independent variable in the regression approach. More important than these standard errors are the standard deviations associated with the squared difference of factor and regression model predictions from the true (on/off) values. These are given in Table 4.

Table 4
Estimated Standard Deviations of Differences between Model Predicted Values and True

Values

Prediction under:	Factor Model*	Model (1)**
Organics_auger	0.094	0.137
Organics_nonauger	0.039	0.138
Particulate_auger	0.154	0.315

* Based on equation (8)
** Based on equation (6)
Simulation results suggest that the factor model standard deviations based on substitution of factor model estimates in eq. (8) are underestimates, but the factor model standard deviations from the simulations are still smaller.

Both models demonstrate the tendency for (on/off) ratios to decrease with increasing levels of the uncontrolled environment (especially for organic determinations). However, the factor model makes clearer the presence of a common factor that we could call "environmental." If possible, it would be beneficial to identify and measure specific environmental variables and include these in the factor model. This is discussed further in the next section.

8) Benefits of Replication for Factor Models

Questions of interest include the following:

1) If several instruments of the same kind are used at a sampling location, how much smaller is the variance of the predictions?
2) If instruments for different analytes are used at the same location, or instruments of either the same or different kinds are used at different locations, this might require the inclusion of additional factors. Are variances of predictions increased if factors are added?

These questions only apply to the factor model, since the regression model deals with each instrument individually. Answers to the above questions are:

1) Suppose there are n replicates of the same kind of instrument at the same location. Then the components of Λ that correspond to these instruments should have components for the controlled environment that are approximately equal and components for the uncontolled environment that are approximately equal. If one factor is sufficient, then Λ is $2n \times 1$. It is shown in Appendix 1 that the larger n is and the larger the ratios of the squared factor loadings to the specificities, then the smaller the variance in equation (8). Thus, the more variability explained by the factor, and the greater the number of pairs, the smaller the variance (See Appendix 1).

2) The simplest way to partially answer this question is to consider the situation that for one instrument, the (on, off) determinations are approximately functions of just one

factor. (This seems to be approximately true for the org_aug determinations.). Suppose the other instruments have dependence on two factors, rather than just the single primary factor. Suppose that Λ_1 and Λ_2 are the loadings, respectively, on the first and second factors. In the example data most elements of Λ_1 exceed the corresponding elements of Λ_2 in absolute value. The degree to which the components of Λ_1 > components of Λ_2 (in absolute value) determines how much larger the variance will be for this situation than in the situation where the factor Λ_1 is sufficient. The present data do suggest that there is a dominant Λ_1 for these data. In this case, less is lost by making measurements with different kinds of instruments at different locations (See Appendix 1).

Design considerations that relate to the above discussion are:

1) There is no guarantee that Λ_1 will be dominant, but since this may depend on having a variety of environmental conditions, it is important to sample over several days, so as to have as much variety as possible.
2) When resources are limited, both in terms of the time available for sampling and the number of instruments available, it may be best to replicate the same instrument at the same location. Even if the data are limited to two instruments of the same kind at the same location, the adequacy of a single factor can be tested.
3) If there are adequate resources to attempt to sample potential explanatory variables, this would be helpful for interpretation of factors. For instance, wind measurements, both speed and direction, could be included as additional response variables in the model. If they correlate well with a dominant factor, the dominant factor could be interpreted, in part, as representing the effect of the wind.
4) Clearly the factor model is much more complicated to collect data for and analyze than the regression model. It probably is unreasonable to try to carry out a factor analysis for every outdoor control technology assessment. For multi-site studies, it could be useful to do this larger kind of evaluation to assess the dependency of engineering control effectiveness on the level of the uncontrolled environment. Once this is established, then the simpler regression methods can be used.

9) Discussion

There are somewhat different assumptions and interpretations for the two models considered here. The regression model assumes that Ln (on/off) values are normally distributed and a linear function of Ln (off), as appears to be true for the example data. Also, predictions are made either for randomly sampled (on,off) pairs (if bivariate normality is appropriate) or for the given Ln (off) values. The factor model identifies the random factors that best describe the model, under assumptions of normality and linearity, and statistical independence of factors, if there are more than one. The identification of a major common factor, as is done for the example data, is an appealing aspect of the factor model. Comparison of factor loadings by variable can provide better understanding of the data. The factor model gives unbiased estimates of the loadings on common factors and the regression model does not. It may be that comparison of factor loadings across different sites could be useful in understanding differences between sites.

It is helpful to see how the above remarks apply to the example data. From Table 3 it appears that there is little difference in the trends identified by the two organics instruments. For the particulate data, the factor model indicates little decrease in the (on/off) ratio, since the coefficients for factors 1 and 2 are close in magnitude and have opposite signs, and the factors have the same variances (=1).

From Figure 4, it is clear that the particulate ratios are smaller than for either organic determination. (Average ratios are about 0.15 for part_aug, about 0.47 for org_aug, and about 0.77 for org_naug.) The particulate data appear much more variable than the organic data. This may be because there are more sources of particulate than of organics during paving operations. The org_naug data levels are much lower than the org_aug levels. Experience has indicated that it is usually harder to see small (on/off) ratios for the org_naug data than for any auger data. Nevertheless, the trend is similar to that for the org_aug. Perhaps this supports the idea of having fewer sources for the organics. The smaller ratios for part_aug than for org_aug have been puzzling results that have occurred in other paving studies. Figures 4, 5 and 6 do suggest that differences in (on/off) between the two auger determinations decrease considerably with increasing factor 1, which we are calling the common environmental control variable.

10) Conclusions

Both the factor model and the regression model provide unbiased estimates of the reduction due to the control. Thus, the simplicity of the regression models is an advantage, both in fitting it, and in the assumptions required for it to be valid. An advantage of the factor method is that it can make clearer the presence of common factors, which may have an environmental interpretation. Identification of common environmental factors that explain much of the variability in the data is especially important in early stages of a study because factor methods can determine whether control effectiveness varies with the level of the environmental factor. If this relationship is established, simpler regression methods may be adequate for subsequent statistical analyses in the study.

11) Acknowledgments

The authors thank NIOSH reviewers Edward Krieg, Jr., Martin Petersen, and John Sheehy for helpful comments. Also, the material presented here is a more complete version of that in Shulman, et. al.[13]. Also thanks to Connie Jo Wilson for considerable help with formatting of the manuscript.

12) References

1) Fuller, W. Measurement Error Models. Wiley, NY, 1987.

2) Shulman, S.A., Mead, K.R., Mickelsen, R.L. "Modeling Performance of Engineering Controls when Reductions Are Largest at the Highest Environmental Concentrations of the Hazardous Contaminant," 2002 Proceedings of the American Statistical Association,

Section on Physical and Engineering Sciences [CD-ROM], Alexandria, VA: American Statistical Association, 2002

3) Krieg, Jr., E.F., Kesner, J.S., Knecht, E.A. "A Structural Equation Model for Method Comparison Studies." Unpublished manuscript.

4) Cochran, W.G., Cox, G.M. Experimental Designs, 2nd Edition. Wiley, NY, 1957, p. 293.

5) Morrison, D.F. Multivariate Statistical Methods, McGraw-Hill, NY, 1967, pp. 261-262.

6) Fuller, op. cit., p. 353.

7) Anderson, T.W. "Estimating Linear Statistical Relationships," Annals of Statistics, 1984, Vol.12, p.24.

8) Fuller, op. cit., p. 364.

9) Fuller, op. cit., p. 3.

10) Bollen, K. Structural Equations with Latent Variables. Wiley, NY, 1989, p. 89.

11) SAS/STAT User's Guide, Version 8. SAS Institute, Inc. Cary, NC, 1999. p. 435, p. 2085.

12) Hatcher, L. Step by Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. SAS Institute, Inc. Cary, NC, 1994, p. 197.

13) Shulman, S.A., Mickelsen, R.L., Mead, K.R. "Adjusting for the Effect of Environmental Variability in Outdoor Engineering Control Studies." To Be Published in 2003 Proceedings of the American Statistical Association, Section on Quality and Productivity [CD-ROM], Alexandria, VA: American Statistical Association.

14) Morrison, op. cit , p. 88.

15) Rao, C.R. Linear Statistical Inference and its Applications, 2nd Edition. Wiley, NY, 1973, p. 64.

16) Fuller, op. cit., p. 355.

13) Appendix 1

A) Variance of Predictions in the Factor and Regression Models

The following development is based on the presentation in (Fuller[1]), though the result

(15) below does not appear there.

The model (3) can be written in the form:

$$\mathbf{Z} = \mu_Z + \Lambda \mathbf{w} + \mathbf{e} = \mathbf{z} + \mathbf{e} \qquad (9)$$

where **Z** is a px1 vector, Λ is a p*k matrix, **w** is a kx1 vector, and the measurement error **e** is a px1 vector. **Z** is the observed variable and **w** is the unobserved factors. **E(w)=0.** μ_Z is the expectation of **Z.** **z** is the part of **Z** distinct from the measurement error (the $\mathrm{Ln}(y_{true})$ of (7)) that we wish to predict. The aim here is to obtain the variance of the difference between **z** and the predictor of **z** , given in (15) below. Note that $\mu_Z = \mu_z$ As in the text, it is assumed that **Var(e)=Σ_e** and **Var(w)** is a kxk identity matrix. (9) corresponds to the factor models (3) and (5).

Under a multivariate normal distribution for **w** and **Z,** the conditional expectation of **w** given **Z** is (Morrison[14])

$$\dot{w} = \mu_W + \Sigma_{wZ}\,\Sigma_Z^{-1}\,(\mathbf{Z} - \mu_Z) \text{ and } \Sigma_{\dot{w}} = \Sigma_{wZ}\,\Sigma_Z^{-1}\,\Sigma_{Zw}, \qquad (10)$$

where Σ_{wZ} is the kxp covariance matrix of **w** and **Z'**, and $\Sigma_{Zw} = (\Sigma_{wZ})'$
The conditional expectation has optimal properties as a predictor (Rao[15]), and under multivariate normality, the relationship is linear.

Under the model (9), $\Sigma_{wZ} = \Lambda'$ and $\Sigma_Z = \Lambda\Lambda' + \Sigma_e = \Sigma_z + \Sigma_e$

$$\Sigma_Z^{-1} = [\,\Sigma_e^{-1} - \Sigma_e^{-1}\,\Lambda\,(\mathbf{I} + \Lambda'\Sigma_e^{-1}\Lambda)^{-1}\Lambda'\Sigma_e^{-1}\,], \text{ (Fuller[16])}$$

where I is a kxk identity matrix, and Σ_e is a pxp diagonal matrix with positive diagonal elements.

$$\Sigma_{wZ}\,\Sigma_Z^{-1} = [\mathbf{I} - \Lambda'\Sigma_e^{-1}\,\Lambda\,(\mathbf{I} + \Lambda'\Sigma_e^{-1}\Lambda)^{-1}\,]\,\Lambda'\Sigma_e^{-1} = (\mathbf{I} + \Lambda'\Sigma_e^{-1}\Lambda)^{-1}\,\Lambda'\Sigma_e^{-1}$$

Let $\mathbf{V} = \Lambda'\Sigma_e^{-1}\Lambda$. Thus, by (10), the predictor of w is (11)

(10) may be rewritten to give $\dot{\mathbf{z}}$, the predictor of **z,** the conditional mean of **z** given $\dot{w}$, again assuming multivariate normality, where :

$$\dot{\mathbf{z}} = \mu_z + \Sigma_{z\dot{w}}\,\Sigma_{\dot{w}}^{-1}\,(\dot{w} - \mu_{\dot{w}}) \text{ and } \Sigma_{\dot{z}} = \Sigma_{z\dot{w}}\,\Sigma_{\dot{w}}^{-1}\,\Sigma_{\dot{w}z} \qquad (12)$$

$$\Sigma_{z\dot{w}} = E[\,(\mathbf{z}-\mu_z)(\dot{w} - \mu_{\dot{w}})'\,] = E[(\mathbf{z}-\mu_z)(\mathbf{Z}-\mu_z)'\,]\,\Sigma_e^{-1}\,\Lambda\,(\mathbf{I} + \mathbf{V})^{-1}\,] \qquad (13)$$
$$= \Lambda\Lambda'\Sigma_e^{-1}\Lambda(\mathbf{I} + \mathbf{V})^{-1} = \Lambda\mathbf{V}\,(\mathbf{I} + \mathbf{V})^{-1}$$

From (11), (12), and (13)

$$\dot{\mathbf{z}} = \mu_z + [\Lambda\mathbf{V}\,(\mathbf{I} + \mathbf{V})^{-1}]\,[\mathbf{V}^{-1}\,(\mathbf{I} + \mathbf{V})]\,[\,(\mathbf{I} + \mathbf{V})^{-1}\Lambda'\Sigma_e^{-1}(\mathbf{Z} - \mu_Z)]$$

The above expression can be simplified, since $\mathbf{V(I+V)^{-1}=(V^{-1}+I)^{-1}}$. Therefore,

$$\dot{\mathbf{z}} = \mu_z + \Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'\Sigma_e^{-1}(\mathbf{Z}-\mu_Z)$$
$$=\mu_z + \Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'\Sigma_e^{-1}(\mathbf{z}-\mu_Z+\mathbf{e}) \qquad (14)$$

The aim is to calculate $\mathbf{Var(\dot{z}-z)}$. From (14)

$$(\dot{\mathbf{z}}-\mathbf{z})=[\Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'\Sigma_e^{-1}-\mathbf{I}](\mathbf{z}-\mu_Z)+[\Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'\Sigma_e^{-1}]\mathbf{e}$$

Let $\mathbf{W}=\Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'\Sigma_e^{-1}$

$$\mathbf{Var(\dot{z}-z)}=(\mathbf{W-I})(\Lambda\Lambda')(\mathbf{W-I})' + \mathbf{W}\Sigma_e\mathbf{W}'.$$

Since $(\mathbf{W-I})\Lambda=\Lambda(\mathbf{I}+\mathbf{V})^{-1}\mathbf{V}-\Lambda=\Lambda[(\mathbf{I}+\mathbf{V})^{-1}\mathbf{V}-\mathbf{I}]=(-\Lambda)(\mathbf{I}+\mathbf{V})^{-1}$, therefore

$$\mathbf{Var(\dot{z}-z)}=\Lambda(\mathbf{I}+\mathbf{V})^{-2}\ \Lambda' + \Lambda(\mathbf{I}+\mathbf{V})^{-1}\mathbf{V}(\mathbf{I}+\mathbf{V})^{-1}\Lambda' \qquad (15)$$
$$=\Lambda(\mathbf{I}+\mathbf{V})^{-1}\ [\mathbf{I}+\mathbf{V}]\ (\mathbf{I}+\mathbf{V})^{-1}\Lambda'$$
$$=\Lambda(\mathbf{I}+\mathbf{V})^{-1}\ \Lambda'$$
$$=\Lambda(\mathbf{I}+\Lambda'\Sigma_e^{-1}\Lambda)^{-1}\ \Lambda'$$

The interest here is in the differences between components of **z**. For instance, let **c**= (1 -1 0 0 0 0). The variance of the difference between the prediction **cż** and **cz** is:

$$\mathbf{Var[c(\dot{z}-z)]}=\mathbf{c}\Lambda(\mathbf{I}+\Lambda'\Sigma_e^{-1}\Lambda)^{-1}\ \Lambda'\mathbf{c}' \qquad (16)$$

(16) is used to calculate the standard deviations for the factor model in Table 3.

The corresponding results for the regression model are derived as follows:
Assume bivariate normal variables x_i, i=1, 2.
Let $x_i = x_{i,tr} + u_{i,tr}$, where for each i=1,2, the two addends are statistically independent normal variables. $x_{i,tr}$ is a random variable with variance σ_{tri}^2 and $u_{i,tr}$ is independently distributed error (including measurement error) with variance σ_{ui}^2 and expectation 0, and $\sigma_i^2=\sigma_{tri}^2+\sigma_{ui}^2$. Also, the $u_{i,tr}$ s are independent.

The regression model under bivariate normality is (Morrison[14]):

$$x_2=\mu_2 + \rho(\sigma_2/\sigma_1)(x_1-\mu_1) + (1-\rho^2)^{0.5}\sigma_2 e_2,$$

where ρ=Correlation(x_1, x_2), μ_i=E(x_i), and $e_2 \sim N(0,1)$, where E indicates expectation.

Thus,

$$(x_2-x_1)=(\mu_2-\mu_1) + [\rho(\sigma_2/\sigma_1)-1](x_1-\mu_1) + (1-\rho^2)^{0.5}\sigma_2 e_2,$$

Let $Est(x_2 - x_1) = (\mu_2 - \mu_1) + [\rho(\sigma_2/\sigma_1) - 1](x_1 - \mu_1)$
$= (\mu_2 - \mu_1) + [\rho(\sigma_2/\sigma_1) - 1](x_{1tr} - \mu_1) + [\rho(\sigma_2/\sigma_1) - 1]\, u_{1tr}$
The aim is to estimate $E[Est(x_2 - x_1) - (x_{2tr} - x_{1tr})]^2$.

$Est(x_2-x_1)-(x_{2tr} - x_{1tr}) = Est(x_2 - x_1)-[x_{2tr} - (x_{1tr} - \mu_1)] + \mu_1$
$= (\mu_2-\mu_1)+[\rho(\sigma_2/\sigma_1)-1](x_{1tr}-\mu_1)+[\rho(\sigma_2/\sigma_1)-1]u_{1tr} -[x_{2tr} -(x_{1tr} - \mu_1)]+ \mu_1$
$= -(x_{2tr} - \mu_2) + [\rho(\sigma_2/\sigma_1)](x_{1tr} - \mu_1) + [\rho(\sigma_2/\sigma_1) - 1]\, u_{1tr}$

$E[Est(x_2 - x_1) - (x_{2tr} - x_{1tr})]^2$
$= \sigma_{tr2}^2 + [\rho(\sigma_2/\sigma_1)]^2 \sigma_{tr1}^2 + [\rho(\sigma_2/\sigma_1) - 1]^2 \sigma_{u1}^2 - 2[\rho(\sigma_2/\sigma_1)]\, \rho(\sigma_2\, \sigma_1),$

where $\rho(\sigma_2\, \sigma_1)$ in the last addend on the right is $Covariance(x_2, x_1) = Covariance(x_{2,tr}, x_{1,tr})$

B) Replication Issues

Suppose there are n replicates of the same kind of instrument at the same location. Then the components of Λ that correspond to these instruments should have components for controlled that are approximately equal and components for the uncontrolled environment that are approximately equal. The vector $\mathbf{c} = (1/n)(1\ -1\ 1\ -1\ 1\ -1\ \ldots)$, where there are n pairs of (1 -1)s. If one factor is sufficient, then Λ is 2n x 1. If the elements of Λ are λ_{i1}, then we would expect that $\lambda_{11} = \lambda_{31} = \lambda_{51}$, etc., and $\lambda_{21} = \lambda_{41} = \lambda_{62}$, etc. Likewise, for the specificities, the controlled determinations should approximately equal σ^2_1 and the uncontrolled should approximately equal σ^2_2. Thus,

$c\Lambda = (1/n)\ (n(\lambda_{11} - \lambda_{21})) = (\lambda_{11} - \lambda_{21})$, which is the same value if there had been just one instrument. For one factor, equation (16) has the simple form $(\lambda_{11} - \lambda_{21})^2 / (1 + \Lambda'\Sigma_e^{-1}\Lambda)$. Thus, the numerator does not depend on n, but the denominator does.

For this design, $(1 + \Lambda'\Sigma_e^{-1}\Lambda) = 1 + n(\lambda^2_{11}/\sigma^2_1 + \lambda^2_{21}/\sigma^2_2)$. Thus the larger n is and the larger the ratios $(\lambda^2_{11}/\sigma^2_1)$ and $(\lambda^2_{21}/\sigma^2_2)$ are, then the smaller is the variance in equation (16).

Another issue concerns the problem that whereas some determinations may be functions of just one factor, this need not be true of all determinations. If additional kinds of instruments are added to a study, and if these require additional factors, then how much greater will the variance in equation (16) be for the instrument that requires only one factor?

Let $\mathbf{c} = (1\ -1\ 0\ 0\ 0\ 0)$ and suppose that there are two factors, but the pair of (controlled, uncontrolled) determinations for the first instrument are adequately described by the first factor. Let the first and second columns of Λ be Λ_1 and Λ_2.

$$(\mathbf{I} + \Lambda'\Sigma_e^{-1}\Lambda) = \begin{vmatrix} (1 + \Lambda'_1\Sigma_e^{-1}\Lambda_1) & \Lambda'_1\Sigma_e^{-1}\Lambda_2 \\ \Lambda'_1\Sigma_e^{-1}\Lambda_2 & (1 + \Lambda'_2\Sigma_e^{-1}\Lambda_2) \end{vmatrix}$$

Thus, equation (16) may be written as

$| (\lambda_{11}- \lambda_{21}) \quad 0| \quad (\mathbf{I} + \Lambda'\Sigma_e^{-1} \Lambda)^{-1} \quad | (\lambda_{11}- \lambda_{21}) \quad 0| \; `=(\lambda_{11}- \lambda_{21})^2 \; a,$

where $a =(1 + \Lambda'_2 \Sigma_e^{-1}\Lambda_2)/ [(1 + \Lambda'_1 \Sigma_e^{-1}\Lambda_1)(1 + \Lambda'_2 \Sigma_e^{-1}\Lambda_2) -(\Lambda'_1 \Sigma_e^{-1}\Lambda_2)^2].$

Had there been just one factor, with exactly the same loadings as Λ_1, the multiplier a would be $1/ (1 + \Lambda'_1 \Sigma_e^{-1}\Lambda_1)$. Thus, the variance will be larger here, since

$a= 1/[(1 + \Lambda'_1 \Sigma_e^{-1}\Lambda_1)-(\Lambda'_1 \Sigma_e^{-1}\Lambda_2)^2 / (1 +\Lambda'_2 \Sigma_e^{-1}\Lambda_2)]$

The degree to which the components of $\Lambda_1 > \Lambda_2$ (in absolute value) determine how much larger the variance will be than the one factor situation. The present data do suggest that there is a dominant Λ_1 for these data. In this case less is lost by making measurements with different kinds of instruments at different locations. For the example data, $a=(1/22.6)$, and $1/[(1 + \Lambda'_1 \Sigma_e^{-1}\Lambda_1)=1/35.1$. Thus, the variance is about 50% bigger than if there were only one factor equal to Λ_1 and Σ_e was unchanged.

14) Appendix 2: SAS Programs

1) Proc Calis Code

```
proc calis ucov aug pshort residual pestim maxiter=2000 platcov outram=ram1
outstat=outs1;
var y1 y2 y3 y4 y5 y6 ;
lineqs
y1=a10 intercept + a11 f1 +          e1,
y2=a20 intercept + a21 f1 + a22 f2 + e2,
y3=a30 intercept + a31 f1 + a32 f2 + e3,
y4=a40 intercept + a41 f1 + a42 f2 + e4,
y5=a50 intercept + a51 f1 + a52 f2 + e5,
y6=a60 intercept + a61 f1 + a62 f2 + e6 ;
std e1-e6 = eps1-eps6, f1=1,f2=1;
run;
```

Note: y1-y6 are the natural logs of the six instrumental variable determinations, Org_aug_off, org_aug_on, org_naug_off, org_naug_on, part_aug_off, part_aug_on As explained in the text, a12 =0.

2) Proc Mixed Code

```
proc mixed method=ml asycov; classes loc gr date type;model y=loc;
repeated / subject=pair type=fa(2) ;
```

Note: y is the natural log of the six instrumental determinations; "loc" designates which instrument-control setting is being used (Org_aug_off, org_aug_on, org_naug_off,

org_naug_on, part_aug_off, part_aug_on); "pair" designates the set of six simultaneous determinations.