PNNL-11000
UC-810
Project Technical Information

RECEIVED
MAR 2 2 1996
OSTI

Methods for Estimation of Covariance Matrices and Covariance Components for the Hanford Waste Vitrification Plant Process

M.F. Bryan
G.F. Piepel
D.B. Simpson

March 1996

Prepared for the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest National Laboratory
Operated for the U.S. Department of Energy
by Battelle Memorial Institute

Battelle

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED 83

PNNL-11000
UC-810
Project Technical Information

Methods for Estimation of Covariance Matrices and Covariance Components for the Hanford Waste Vitrification Plant Process

M. F. Bryan
G. F. Piepel
D. B. Simpson

March 1996

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest National Laboratory
Richland, Washington 99352

Reprint of historical document PHTD-C93-05.01P Rev. 0, dated June 1994. Data, formatting, and other conventions reflect standards at the original date of printing. Technical peer reviews and editorial reviews may not have been performed.

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 Battelle Memorial Institute, 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, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

Printed in the United States of America

**Available to DOE and DOE contractors from the
Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831;
prices available from (615) 576-8401.**

**Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161**

The document was printed on recycled paper.

SUMMARY

The high-level waste (HLW) vitrification plant at the Hanford Site was being designed to immobilize transuranic and high-level radioactive waste in borosilicate glass. Each batch of plant feed material must meet certain requirements related to plant performance, and the resulting glass must meet requirements imposed by the Waste Acceptance Product Specifications (WAPS; DOE, 1993). Properties of a process batch and the resulting glass are largely determined by the composition of the feed material. Empirical models are being developed to estimate some property values from data on feed composition.

Methods for checking and documenting compliance with feed and glass requirements must account for various types of uncertainties. This document focuses on the estimation, manipulation, and consequences of composition uncertainty, i.e., the uncertainty inherent in estimates of feed or glass composition. Three components of composition uncertainty will play a role in estimating and checking feed and glass properties: batch-to-batch variability, within-batch uncertainty, and analytical uncertainty. In this document, composition uncertainty and its components are treated in terms of variances and variance components for univariate situations, covariance matrices and covariance components for multivariate situations. The importance of variance and covariance components stems from their crucial role in properly estimating uncertainty in values calculated from a set of observations on a process batch.

Two general types of methods for estimating uncertainty are discussed: 1) methods based on data, and 2) methods based on knowledge, assumptions, and opinions about the vitrification process. Data-based methods for estimating variances and covariance matrices are well known. Several types of data-based methods exist for estimation of variance components; those based on the statistical method *analysis of variance* are discussed, as are the strengths and weaknesses of this approach. Alternative approaches are mentioned briefly. Methods for estimating covariance components are based on methods for estimating variance components.

Estimating uncertainty from process knowledge may be necessary when data are scarce or lacking entirely. A Monte Carlo procedure for this type of uncertainty estimation is illustrated for a hypothetical analytical process; a hypothetical analytical covariance matrix is developed for use in later studies. Measures of strength of belief in simulated uncertainty estimates are required to update or combine these estimates with incoming information; an approach to assigning such measures is developed.

Proper estimation of the uncertainties required for testing process/product specifications requires the combination of various components of uncertainty. Satterthwaite's method for combining components of uncertainty and for assessing the precision of the overall uncertainty estimate is recommended.

Several batch and glass properties will be estimated from empirical models based on feed composition. Overall uncertainty in estimated property values derives both from uncertainty in estimated feed composition and from uncertainty in model coefficients. A general method for collapsing multivariate composition and model uncertainties into univariate uncertainties for property values is described. Applying this method with the hypothetical analytical covariance matrix indicates that analytical uncertainty can be expected to produce a relative standard deviation between 3.5% and 8% in estimated property values. An alternative method for assessing the contribution of composition uncertainty to overall uncertainty is discussed.

A method for calculating the sample size required for estimation of univariate uncertainty (with specified precision and confidence) is developed. The problem of simultaneous inference arises in estimation of multivariate uncertainties. Simulation studies indicate that large sample sizes are required to produce good estimates of covariance matrices, but that covariance matrices estimated from relatively small samples, when propagated through property models, produce reasonably stable estimates of univariate uncertainty in modelled properties. It is recommended that estimates of composition covariance matrices be based on at least 20 observations.

Data on composition uncertainty will accumulate during vitrification operations. All available information (from preceding batches, as well as from the current batch) should be taken into account, so a method for combining previous estimates with current data is desirable. Several alternatives are discussed, and a univariate Bayesian approach to updating estimates of uncertainty is described in detail.

The compositional nature of the data involved in the HLW vitrification process creates difficulties in development and testing of statistical algorithms. Several related topics are addressed. Finally, applications of the work described in this document and suggestions for future work are presented.

GLOSSARY

Acceptable--A batch or composition for which all applicable requirements will be met (with some degree of statistical confidence, as discussed in the body of the document).

Analytical uncertainty--Uncertainty among analytical results from the same sample. This is a composite form of uncertainty, made up of *variability* induced during sample preparation and the inherent *error* of the measurement process itself.

Batch--A discrete quantity of material (waste, frit, recycle, or a combination of the three) to be processed by the Hanford high-level waste (HLW) vitrification plant.

Batch-to-batch variability--Heterogeneity between *batches* made from the same *waste type*.

Bias--Consistent departures of measured or estimated quantities from the true value (see also *error*).

Components of covariance--*Covariance matrices* representing hierarchical levels of uncertainty for multivariate data.

Components of variance--*Variances* representing hierarchical levels of uncertainty in univariate data.

Composition--The proportions of each chemical species in a batch of material to be processed by the HLW vitrification plant; usually expressed as mass fractions of nine major oxides (SiO_2, B_2O_3, Na_2O, Li_2O, CaO, MgO, Fe_2O_3, Al_2O_3, ZrO_2) and a catchall tenth category, "Others." In some cases, individual species normally included in Others may be segregated.

Composition uncertainty--Uncertainty in measured or estimated quantities stemming from *variability* in material and/or sampling and analytical *error*.

Compositional data--A type of multivariate data in which the numerical values in each datum are the proportions (or percentages) of the individual components of the material or characteristic being represented by the datum. From their nature as proportions (percentages), these numerical values must lie between 0 and 1 (0 and 100%), inclusive, and they must sum to 1 (100%).

Confidence--A measure of the long-run performance of a statistical procedure, expressed as the probability that the procedure produces the advertised result. For example, the procedure for producing a 95% confidence interval for the mean of a population has a 95% chance of producing an interval that traps the mean. Note that *confidence* pertains to the procedure and not to any particular result.

Confidence interval--A type of statistical interval designed to trap, with specified *confidence*, a single fixed true value, such as the mean of a random variable.

Correlation--A standardized *covariance* which must lie between -1 and 1, *correlation* is computed by dividing the covariance between two random variables by the product of the standard deviations of the two variables.

Correlation matrix--A standardized representation of the interrelationships between individual quantities that make up a multivariate datum, the *correlation matrix* is a symmetric matrix with 1's on the diagonal and the pairwise *correlations* in the off-diagonal positions.

Covariance--A measure of the tendency of two random quantities to vary together, *covariance* is defined as the *expected value* of the product of the deviations of the two random quantities from their respective means, i.e., Covariance(X,Y) = $E(X - \mu_X)(Y - \mu_Y)$. Positive covariance indicates that the two quantities tend to increase or decrease together. Negative covariance indicates that one quantity tends to increase while the other decreases (or vice versa). Covariance can be estimated from a sample of n pairs (X_i,Y_i), i = 1, ..., n, with the formula

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

Covariance components--See *components of covariance.*

Covariance matrix--A representation of the uncertainties and interrelationships between individual quantities that make up a multivariate datum, the *covariance matrix* is a symmetric matrix with the variances of the individual quantities on the diagonal and the pairwise *covariances* in the off-diagonal positions.

$E(\cdot)$--See *expected value.*

Error--The random deviation of a measured or estimated quantity from the true value, related to the imperfection of the sampling or analytical procedure.

Expectation--See *expected value.*

Expected value--The average value of a random quantity; in general, given a function, h(X), of a random variable X, the *expected value* (or *expectation*) of h(X) is defined as

$$E(h(X)) \equiv \int_{-\infty}^{\infty} h(x)\, dF(x) \ .$$

Feed--Though technically referring to material after processing in the Slurry Mix Evaporator, *feed* or *feed material* will here be used as a generic term to refer to any material being processed in the HLW vitrification plant, upstream of the melter itself (see also *melt*).

Long-term variability--Heterogeneity in material over waste types.

Mean--A statistical measure of the average or central tendency of a random quantity; the *mean*, μ, of a random variable X is simply the *expected value* of X, i.e., μ = *E*(X). The mean can be estimated from a sample, X_i, i = 1, ..., n, with the formula

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Melt--Material being processed by the HLW vitrification plant in the melter or before it has cooled and solidified into glass. Before reaching the melter, this material will be referred to as *feed.*

Model uncertainty--Uncertainty in an estimated property value stemming from imperfection of the model used to relate feed composition to the property.

Modelled properties--Properties of feed, melt, or glass for which statistical models are being developed to relate feed composition to the property values.

Moments--The *expected values* of powers of a random variable, X. The first moment, *E*(X), is the *mean*, μ. *Central moments* are expected values of powers of the difference between X and its mean; the second central moment, $E(X\text{-}\mu)^2$, is the *variance*.

Multiple-batch requirement or constraint--A requirement or constraint imposed over a set of batches to be processed by the HLW vitrification plant; e.g., a property for which the requirement is imposed on an entire *waste type*, rather than on the individual batches constituting the waste type. See also *single-batch requirement or constraint.*

Relative standard deviation--The ratio of the standard deviation to the mean; estimated by $S/\overline{X}$.

S--See *standard deviation.*

S^2--See *variance.*

Sampling uncertainty--See *within-batch uncertainty.*

Single-batch requirement or constraint--A requirement or constraint imposed on each individual batch to be processed by the HLW vitrification plant, with no reference to the characteristics of preceding or succeeding batches. See also *multiple-batch requirement or constraint.*

Standard deviation--Defined as the square root of the *variance*, the *standard deviation* is a measure of uncertainty on the same scale as the original quantity. Roughly, the standard deviation is the average distance of an observed value from the mean.

Uncertainty--A general term used to refer to any of several measures of the random behavior of some quantity; for example, see *composition uncertainty*, *model uncertainty*, *variability*, and *error*.

Variability--Uncertainty related to heterogeneity in material under examination; for example, see *batch-to-batch variability*.

Variance--A statistical measure of the random behavior of some quantity, *variance* is defined as the *expected value* of the squared deviation of a random variable, X, from its mean, μ, i.e., Variance(X) = $E(X - \mu)^2$. Variance can be estimated from a sample, X_i, i = 1, ..., n, with the formula

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Variance components--See *components of variance*.

Variance-covariance matrix--See *covariance matrix*.

WAPS properties and requirements--Properties of and requirements on glass produced by the HLW vitrification plant, as detailed in the *Waste Acceptance Product Specifications* (WAPS; DOE, 1993). These properties and requirements are related to the performance of the glass in the repository.

Waste type--A relatively homogeneous stream of waste to be processed by the HLW vitrification plant. Several to many *batches* will be made from a single waste stream.

Within-batch uncertainty--Uncertainty among samples from the same process batch; this is a composite form of uncertainty, made up of *variability* (heterogeneity) in the process batch and the inherent *error* of the sampling process itself.

$\overline{X}$--See *mean*.

ACRONYMS

ANOVA--Analysis of Variance

CVS--Composition Variability Study

DWPF--Defense Waste Processing Facility

HLW--High-Level Waste

IID--Independent and identically distributed

MEM--Measurement Error Model

PCC--The algorithms to be used by the HLW vitrification plant for product composition control

PCT--Product Consistency Test

PVTD--Pacific Northwest Laboratory (PNL) Vitrification Technology Development

PPMD--Process/Product Model Development

RSD--Relative standard deviation

WAPS--Waste Acceptance Product Specifications

CONTENTS

1.0 INTRODUCTION

The high-level waste (HLW) vitrification plant at the Hanford Site was being designed to immobilize transuranic and high-level radioactive waste in borosilicate glass. Each batch of plant feed material must meet certain requirements related to plant performance, and the resulting glass must meet requirements imposed by the Waste Acceptance Product Specifications (WAPS; DOE, 1993). Attributes[(a)] of a process batch and the resulting glass are largely determined by the composition of the feed material. Accordingly, methods for controlling feed composition and for checking and documenting product quality must be developed.

Similar vitrification operations will be performed in the Defense Waste Processing Facility (DWPF) at the Savannah River Site. DWPF has developed a Product Composition Control System for controlling feed composition and for checking and documenting product quality (Postles and Brown, 1991; WSRC, 1993). The *HWVP Project Waste Form Qualification Program Plan* (Randklev, 1993) calls for the development of a product composition control-type system to perform these functions for the Hanford HLW vitrification plant. No name for the Hanford product composition control system has yet been generally agreed upon. PCC (from product composition control) is used here to refer to the system under development for the Hanford HLW vitrification plant.

Control of HLW vitrification operations and product quality will be achieved using a series of mathematical/statistical algorithms. A major objective of the Process/Product Model Development (PPMD) cost account of the Pacific Northwest Laboratory Vitrification Technology Development (PVTD) Project is the development of algorithms for a PCC system. These algorithms are discussed in more detail by Bryan and Piepel (1993). For each

[(a)] Established usage reserves the word *property* for characteristics of the melt and glass (which will usually be estimated via models based on feed composition), but requirements and constraints will also be imposed on feed composition (oxide mass fractions and functions thereof). To avoid confusion, the word attribute will be used to refer to any characteristic upon which a requirement or constraint is imposed and for which, therefore, an uncertainty estimate is required.

process batch, the algorithms will: 1) choose a target feed composition, 2) estimate the actual feed composition by reconciling various process measurements, 3) use the estimated feed composition to estimate, check, and document various batch and product characteristics, and 4) recommend remediation strategies for process batches that do not meet requirements.

Remediation options are limited once material reaches the melter. Since feed composition largely determines batch and glass properties, these relationships will be exploited to ensure acceptable batch and glass properties and to perform any required remediation *before* material enters the melter. Development of empirical models relating feed composition to important properties is one objective of the ongoing Composition Variability Study (CVS; Hrma, Piepel, et al., 1992, 1994). The PCC algorithms will use these models to estimate batch and glass properties as functions of feed composition.

Composition uncertainty[a] must be taken into account when estimating and checking any batch or glass attribute. Composition uncertainty is the uncertainty inherent in estimates of feed composition. This type of uncertainty may stem from heterogeneity in material, imperfection of measurement processes, or both. The various categories and sources of composition uncertainty are discussed in detail by Bryan and Piepel (1994); highlights of that discussion are presented here. Three components of composition uncertainty will play a role in estimating and checking batch and glass attributes:

- Batch-to-batch variability -- Heterogeneity between process batches made from the same waste type and frit batch. This type of heterogeneity might also be called between-batch variability or within-waste type variability.

- Within-batch uncertainty -- A combination of heterogeneity within a single process batch and any imperfections in the sampling process. This type of uncertainty might also be called sampling uncertainty.

[a] Composition uncertainty might also be called *data uncertainty*, since it exists to some degree in virtually any process used to collect data. However, the main type of data to be used in HLW vitrification process/product control will be compositional data, so the more specific term is used here. The methods discussed in this document for composition uncertainty will also be used to account for uncertainties in other measured quantities. For example, in at least one of the algorithms, tank level measurements, which are unrelated to composition, will be used.

- Analytical uncertainty -- A combination of heterogeneity within a sample, variability induced during sample preparation, and any imperfections in the analytical process.

When estimating and checking modelled properties, *model uncertainty* also must be taken into account. Model uncertainty is the uncertainty that derives from the use of empirical models (i.e., models fitted to data). Estimating this type of uncertainty is another objective of the CVS and therefore is not discussed in this document. The role and use of model uncertainty in comparing property values to requirements is discussed in some detail by Bryan and Piepel (1994).

The original scope of the work reported in this document included estimating various uncertainties based on data obtained from DWPF. However, these data were not available in time to be incorporated here. Therefore, this work concentrated on identifying, implementing, and testing methods for estimating uncertainties, on manipulating uncertainty estimates, and on the effects of sample size on precision of estimation. In order to address all these issues, some uncertainty estimates were needed, so a method for using process knowledge to estimate uncertainties was developed and implemented, and data from previous studies at the Pacific Northwest Laboratory were used. The major topics covered by this document are

- estimating univariate and multivariate composition uncertainties and components thereof (Sections 3 and 4);
- constructing estimates of composition uncertainty from knowledge of the sampling and analytical process (Section 5);
- combining components of uncertainty and measuring the quality of the resulting estimate (Section 6);
- methods for and results from propagating composition uncertainty through empirical property models to yield estimates of the contribution of composition uncertainty (and components thereof) to uncertainty in estimated property values (Section 7);
- sample sizes required for estimating composition uncertainty and its components (Section 8); and
- updating estimates of uncertainty (Section 9).

Section 2 presents statistical concepts and notation that are required in the rest of the document. Miscellaneous topics are covered in Section 10. Applications and suggestions for future work are presented in Section 11.

2.0 STATISTICAL PRELIMINARIES

For precision and brevity in much of what follows, it is necessary to employ some statistical terminology and notation. This section introduces the required terminology and notation; however, a full exposition and explanation of this material is beyond the scope of this document. Fuller coverage of this material is available in most texts on probability and mathematical statistics (e.g., Lindgren, 1976). This document also uses the concepts and notation of linear algebra, vectors, and matrices. Some of these concepts and notation are defined below; fuller coverage of this material can be found in books on linear algebra (e.g, Searle, 1982).

Statistics is the art and science of making decisions in the face of uncertainty. Accordingly, a major task of statistics is the modelling and characterization of uncertainty. The most common statistical method of modelling uncertainty employs the concept of a *random variable*. Intuitively, a random variable is a quantity that cannot be measured exactly (either because its value is not fixed or because the measurement process is imperfect). Therefore, the behavior of a random variable is described in terms of the probability that the true value of the random variable exists in some set of possible values. Random variables are often denoted by capital letters, e.g., X, while individual values or realizations of a random variable are often denoted by lower case letters, with a subscript to indicate which observation is being represented. For example, n observations of the random variable X might be denoted x_1, x_2, ..., x_n, or, equivalently, x_i, i = 1, ..., n. A group of n observations may also be represented by a vector, $\underline{x}$.

Two basic types of random variables exist. A *discrete random variable* is one for which the number of possible values is finite or countably infinite. In many cases, discrete random variables are counts of the number of occurrences of certain events. For example, the number of defective items produced by a manufacturing process can range from zero to the number of items produced. A *continuous random variable* is one for which the number of possible values is uncountably infinite. In many cases, continuous random variables take on values in an interval of possible values. For example, the value of many measured characteristics (length, weight, concentration, viscosity) must lie between some more or less well known lower and upper bounds, but, at least theoretically, the individual measurements may take on any value in the interval. Although many of the concepts discussed below apply to both discrete and continuous random variables, most of the quantities involved in HLW vitrification process/product control are best modelled by continuous random variables; therefore, this presentation focuses on continuous random variables.

2.1 DISTRIBUTION AND DENSITY FUNCTIONS

Two mathematical functions are useful in describing the behavior of a (continuous) random variable: the *distribution* (or distribution function), and the *density* (or density function). To each random variable X, there corresponds a distribution function, $F(x) \equiv \Pr\{X \leq x\}$[a], the probability that the random variable X is less than or equal to the fixed value x. As a function, $F(\cdot)$ is monotonic and nondecreasing. Since for each fixed x, F(x) is a probability, F(x) must lie in the interval [0,1].

The density function, f(x), exists for most of the common statistical distributions. When it exists, the density function is simply the first derivative of the distribution function, i.e., $f(x) = F'(x)$. The density function characterizes the local behavior of the random variable. By its nature, $f(x) \geq 0$ for all x, and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 .$$

In order to achieve this unit integral, a density function incorporates a normalizing constant (usually a function of the parameters of the distribution, which are discussed below).

Many families of random variables (and the corresponding distributions and densities) have been found useful in statistical applications. For example, the most commonly encountered family of statistical distributions is the family of *normal* (or Gaussian) distributions. The density function for a normally-distributed random variable X is

[a] The symbol "≡" should be read as "is *defined* to be equal to."

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

(μ and σ^2 are the parameters of the normal distribution and are discussed further below).

Another important family of random variables is the gamma family. The density function for a random variable X that follows a gamma distribution is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} I_{(0,\infty)}(x) \ .$$

where α and β are the parameters of the gamma distribution (discussed below) and $I_{(0,\infty)}(x)$ is zero if $x \leq 0$ and one otherwise (indicating that a gamma random variable takes on only positive values).

The members of a family of random variables are distinguished by the values of the associated *parameters.* The parameters of a random variable appear in the density function and are often denoted by lower-case Greek letters. For example, the parameters of the normal density given above are μ and σ^2, while the parameters of the gamma density given above are α and β. Often, the dependence of the behavior of a random variable on the associated parameters is shown by a slight modification of notation: for example, the density of a random variable following a normal distribution with parameters μ and σ^2 may be denoted $f(x|\mu,\sigma^2)$, and the density of a gamma distribution with parameters α and β may be denoted $f(x|\alpha,\beta)$.

A common statistical shorthand for the phrase "the random variable X follows a normal distribution with parameters μ and σ^2" is "$X \sim N(\mu,\sigma^2)$." The shorthand for "the random variable X follows a gamma distribution with parameters α and β" is "$X \sim \Gamma(\alpha,\beta)$."

An important special case of the gamma distribution is the chi-square distribution. This distribution has a single parameter, f, known as the *degrees of freedom.* A chi-square distribution with f degrees of freedom [$\chi^2(f)$] is simply a gamma distribution with parameters f/2 and 2, i.e., the $\Gamma(f/2,2)$ distribution.

2.2 MEAN AND VARIANCE

The *expectation* of a function, h(X), of the random variable X is defined as:

$$E(h(X)) \equiv \int_{-\infty}^{\infty} h(x)\, dF(x) = \int_{-\infty}^{\infty} h(x)\, f(x)\, dx$$

(the last expression makes sense only if the density function exists). Several such functions are important enough to warrant specific names. The *mean* of a random variable X is defined as:

$$\mu_X \equiv E(X) \equiv \int_{-\infty}^{\infty} x \; dF(x) \; .$$

The mean of a random variable is a measure of the central value (or central tendency) of the random variable. The most common measures of dispersion about this central value are the *variance*:

$$\sigma_X^2 \equiv E(X-\mu_X)^2 \equiv \int_{-\infty}^{\infty} (x-\mu_X)^2 \; dF(x)$$

and the closely related *standard deviation*:

$$\sigma_X \equiv \sqrt{\sigma_X^2} \quad .$$

(When the meaning is clear from context, the subscripts on μ_X, σ_X^2, and σ_x may be omitted.) Due to the simple relationship between variance and standard deviation, much of the discussion (though not, of course, the equations) in this document could be framed in terms of either quantity, and shifts between variance and standard deviation go unremarked henceforth.

The mean and variance are examples of *moments* of a distribution. Moments are simply expectations of powers of the random variable (often centered by subtracting the mean). The moments of a distribution convey information on the location and shape of the distribution and hence on the behavior of the random variable. The first moment of a distribution is the mean and, as mentioned above, is a measure of the central value (location) of the distribution. The second (central) moment is the variance and hence is a measure of the spread (scale) of the distribution. The third moment measures the skewness of the distribution, and the fourth moment measures kurtosis (how "heavy-tailed" and peaked the distribution is).

The moments of a distribution are not usually the parameters of the distribution. The exception is the normal distribution, for which the parameters μ and σ^2 are indeed the mean and variance, respectively. The mean and variance of many distributions are simple functions of the parameters. For example, the mean and variance of a $\Gamma(\alpha,\beta)$ distribution are $\alpha\beta$ and $\alpha\beta^2$, respectively; the mean and variance of a chi-square distribution with f degrees of freedom are f and 2f, respectively.

In some cases, it is useful to specify only the mean and variance of a random variable, without ascribing to it a distributional form (such as normal or gamma).. In this case, an adaptation of the shorthand above is employed -- "$X \sim (\mu,\sigma^2)$" means that X is a random variable with mean μ and variance σ^2.

2.3 MULTIVARIATE DATA, COVARIANCE, AND CORRELATION

The discussion of random variables above concentrated on the *univariate* situation, i.e., the modelling of a single quantity (even though many measurements or observations of that quantity may be available). However, in many situations (including much of the material discussed in this document), the *simultaneous* behavior of several different quantities is of interest. This is the *multivariate* situation. The obvious example here is the composition of a vitrification process batch. For use in melt/glass property models, batch composition is usually expressed as mass fractions (proportions or percentages) of nine individual oxides (SiO_2, B_2O_3, Na_2O, Li_2O, CaO, MgO, Fe_2O_3, Al_2O_3, ZrO_2) and a catchall tenth category, "Others." Since these mass fractions must sum to one, they are obviously not independent of one another; hence their simultaneous behavior is of interest.

In multivariate statistics, subscripts are used to distinguish between different random variables. For example, the 10 components of a vitrification process batch can be denoted by

X_1, X_2, ..., X_{10}. Individual observations of a single random variable are usually indicated by a second subscript; for example, x_{ij} is the j-th observation of the i-th random variable.

Most of the standard univariate distributions and densities have multivariate generalizations. When modelling several random variables simultaneously, *joint distributions* and *joint densities*, which are functions that model the simultaneous probabilistic behavior of the variables, must be considered. In addition, when examining the effects of one variable on another, *marginal distributions* and *marginal densities*, which model the probabilistic behavior of one or more variables given the values of other variables, become important. The notation can get quite complex, so, rather than attempting a general treatment, notation is introduced below only as necessary.

In multivariate statistics, the tendency of several quantities to vary together ("co-vary") is of interest. The statistical *covariance* between two random variables X_i and X_j is defined as:

$$\sigma_{ij} \equiv E(X_i - \mu_i)(X_j - \mu_j) \quad ,$$

where the expectation is taken with respect to the joint distribution of X_i and X_j (i.e., this is a double integral). Whereas the variance of a random variable must be nonnegative (by definition), the covariance between two random variables can be positive, negative, or zero. Positive covariance indicates that the two variables tend to vary together; i.e., if one is large (relative to its mean), the other tends also to be large, and if one is small, the other tends to be small. (The repetitive use of the word "tend" is necessitated by the probabilistic nature of the behavior of random variables.) Negative covariance indicates that the two variables tend to vary "in opposite directions;" i.e., if one is large (relative to its mean), the other tends to be small (relative to its mean), and vice versa. Zero covariance indicates that the behavior of one variable does not affect the behavior of the other.[a]

[a] This is not strictly true. Statistical covariance is actually a measure of *linear* covariance, so a strongly curved relation between two random variables is not necessarily reflected in the standard definition of covariance. It is in fact possible to construct two random variables with zero covariance, even though one is an exact function of the other.

Covariances are not scale-invariant, and their magnitudes are affected by the variances of the random variables involved. These characteristics complicate interpretation and comparison of covariances. Statistical *correlation* is essentially a standardized, unitless covariance. The correlation between X_i and X_j is defined as:

$$\rho_{ij} \equiv \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \ .$$

Correlations must lie in the interval [-1,1]. Interpretation of the sign of a correlation is similar to that for a covariance. In addition, the closer the correlation is to 1 (or -1), the nearer the relationship between the two variables is to perfect linearity. The correlation between two random variables is zero if and only the covariance between these two variables is zero. Two variables that have zero correlation (covariance) are said to be *uncorrelated*; if the correlation (covariance) is non-zero, the two variables are said to be *correlated*. Correlated observations are not independent.

Matrix notation is quite useful in multivariate statistics. In this document, matrices are denoted by upper case letters (e.g., Σ or S), and symbols for vectors are underlined (e.g., $\underline{\mu}$). The *random vector*, $\underline{X}$, is a vector of random variables, X_i, i = 1, ..., p. The associated *mean vector* (the vector of means of the individual random variables) is denoted by $\underline{\mu}$. A convenient method for summarizing the variances and pairwise covariances of the elements of the random vector $\underline{X}$ is the *variance-covariance matrix* (for brevity, called the *covariance matrix* below):

$$\Sigma \equiv \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_p^2 \end{bmatrix} .$$

The covariance matrix contains the variances of the individual random variables in the diagonal positions and the pairwise covariances in the off-diagonal positions. As a consequence of the definition of covariance, the covariance matrix is symmetric (i.e., $\sigma_{ij} = \sigma_{ji}$). If the underlying random vector has p elements, the covariance matrix has p rows and p columns; i.e., its dimension is p × p.

Correlations can also be represented in matrix form; the *correlation matrix* is defined as:

$$P \equiv \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{bmatrix}.$$

The diagonal elements of a correlation matrix are always one (since, by definition, the correlation of a random variable with itself is one); the pairwise correlations appear in the off-diagonal positions. Like the covariance matrix, the correlation matrix is symmetric ($\rho_{ij} = \rho_{ji}$).

Multivariate generalizations of many common statistical distributions exist. The notation used to specify the (joint) distribution associated with a random vector $\underline{X}$ parallels that used for a univariate random variable. For example, "$\underline{X} \sim MVN(\underline{\mu},\Sigma)$" indicates that the random vector $\underline{X}$ follows a multivariate normal distribution with parameters $\underline{\mu}$ and Σ. "$\underline{X} \sim (\underline{\mu},\Sigma)$" indicates that $\underline{X}$ follows a (multivariate) distribution with mean vector $\underline{\mu}$ and covariance matrix Σ.

2.4 ESTIMATING POPULATION PARAMETERS WITH SAMPLE STATISTICS

Up to this point, various statistical distributions, parameters, and other theoretical constructs used to model the behavior of random variables have been defined and discussed. In much of statistics, such models for some *population* (real or abstract) of items are postulated or hypothesized, and information is collected about a *sample* drawn from this population. The objectives of this activity include checking the models, estimating parameters, and drawing inferences about the population, based on the sample. Estimation often involves calculating sample analogues to population parameters, moments, and other characteristics. Some of these estimation procedures, and the associated notation, are discussed below.

The usual assumption about a sample is that it is drawn *at random* from the underlying population. The technical definition of a *random sample* is somewhat involved, but essentially a random sample is one in which each item in the population has an equal chance of being selected. A related concept is that of *independent and identically distributed (IID) observations.* Given a sample of size n, x_i, i = 1, ..., n, the assumption might be that each x_i is a realization of a single random variable X, or, equivalently, that the distribution of X_i is the same for all i. This is the concept of identically distributed observations. The concept of independence is essentially that the value of X_i is unaffected by the values of any of the other X_j's (j ≠ i). The statistical shorthand used to describe this situation is "X_i, i = 1, ..., n ~ IID D($\underline{p}$)," where D is the assumed distribution and $\underline{p}$ is the vector of parameters of D. One link between random sampling and IID observations is this: if D($\underline{p}$) is the statistical distribution for a given population, and X_i, i = 1, ..., n, is a random sample from the population, then X_i, i = 1, ..., n ~ IID D($\underline{p}$).

Assume that a random sample of size n is available from a population with mean μ and variance σ^2; i.e., X_i, i = 1, ..., n, ~ IID (μ,σ^2). The sample-based estimate of the *population mean*, μ, is the *sample mean*:

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

The sample-based estimate of the *population variance*, σ^2, is the *sample variance*:

$$s_x^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2 \ . \tag{1}$$

The sample-based estimate of the *population standard deviation*, σ, is the *sample standard deviation*:

$$s_x \equiv \sqrt{s_x^2} \ .$$

The sample mean, $\overline{X}$, is a point estimator of the population mean, μ. In many situations, both a point estimate of the population mean and some idea of the quality of this estimate are required. To address this issue, it must be recognized that the sample mean is a random variable, since it is a function of the random variables X_i, i = 1, ..., n. Therefore, the sample mean has an associated mean and variance. It can be shown that the sample mean is unbiased, i.e., that $E(\overline{X}) = \mu$, so the question of the quality of the sample mean as an estimator of the population mean comes down to the uncertainty in the sample mean. This uncertainty is measured by the standard deviation (or the variance) of $\overline{X}$. In a wide range of cases, the standard deviation of $\overline{X}$ is well estimated by

$$s_{\overline{X}} = \frac{s_x}{\sqrt{n}} \ .$$

This quantity, also known the *standard error*[a] of the mean, is used to construct confidence intervals for the population mean.

The preceding discussion of the standard error of the mean is not, in and of itself, of great importance for the purposes of this document. However, it is included as a concrete illustration of the concept that statistical estimators, such as the sample mean, variance, and standard deviation, are random variables and thus have associated uncertainty. This uncertainty must be quantified in order to judge the quality of the estimators and to draw inferences about true (population) values. The PCC algorithms must deal with uncertainties in statistical estimators, as well as with uncertainties in data.

(a) The term "standard error" is often used to refer to the standard deviation *of an estimator*, as opposed to the standard deviation associated with individual observations.

In the multivariate case, each observation is a vector (rather than a single number). For example, if interest focuses on p characteristics of each item and n items are examined, the data comprise n vectors, each of length p. Denote the observed value for the j-th characteristic of the i-th item as x_{ij}, where j = 1, ..., p, and i = 1, ..., n, and assume that the observations are IID. The sample-based estimate of the *population covariance* between characteristics j and k, σ_{jk}, is the *sample covariance*:

$$\hat{\sigma}_{jk} \equiv s_{jk} \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right) .. \qquad (2)$$

where $\bar{x}_j$ and $\bar{x}_k$ are the sample means of the j-th and k-th characteristics, respectively. The sample-based estimate of the *population covariance matrix*, Σ, is the *sample covariance matrix*:

$$S \equiv \begin{bmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{bmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\underline{x}_i-\underline{\bar{x}}\right)\left(\underline{x}_i-\underline{\bar{x}}\right)^T , \qquad (3)$$

where $\underline{x}_i$ is the i-th observation (a column vector containing the observed values of the p characteristics for the i-th item), $\underline{\bar{x}}$ is the column vector containing the sample means for the p characteristics), and the superscript "T" indicates vector transpose. Since there are p characteristics, the sample covariance matrix is a p × p matrix, and, like the population covariance matrix, it is symmetric. The elements of the sample covariance matrix may be computed individually [using the formula for single sample covariances given in Equation (2)], or the whole matrix may be computed using the vector formula given in Equation (3) -- these methods are equivalent (unless there are missing data).

The sample-based estimate of the *population correlation* between characteristics i and j, ρ_{ij}, is the *sample correlation*:

$$\hat{\rho}_{ij} \equiv r_{ij} \equiv \frac{s_{ij}}{\sqrt{s_i^2 s_j^2}}$$

The sample-based estimate of the *population correlation matrix*, P, is the p × p symmetric *sample correlation matrix*:

$$R \equiv \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix} .$$

2.5 PROPERTY MODEL NOTATION

Finally, some knowledge of the property models under development by CVS is required in the discussions of estimating and propagating composition uncertainty. The property models being developed by CVS are *second-order mixture models*, the general form of which is

$$\phi_k = \sum_{i=1}^{10} b_{ik} x_i + \sum_{i=1}^{9} \sum_{j>i}^{10} b_{ijk} x_i x_j \ , \qquad (4)$$

where ϕ_k is the k-th melt/glass property (or, in some cases, a simple mathematical transformation thereof), the x_i and x_j are the mass fractions of the i-th and j-th oxides, and the b_{ik} and b_{ijk} are the coefficients of the relation between the oxide mass fractions and ϕ_k (to be estimated from the CVS database). The oxide mass fractions used in a mixture model must sum to 1, that is,

$$\sum_{i=1}^{10} x_i = 1 \ .$$

Several of the models developed by CVS are *first-order*, meaning that, for some properties (k), $b_{ijk} = 0$ for all i and j. The form of a first-order model is

$$\phi_k = \sum_{i=1}^{10} b_{ik} x_i \ . \tag{5}$$

Both the first-order model and the second-order model can be written in the form:

$$\phi_k = \underline{x}^T \underline{b}_k \ , \tag{6}$$

where $\underline{x}$ is the vector containing the oxide mass fractions (and cross-products thereof, if the model is second-order), and $\underline{b}_k$ is the vector of estimated coefficients relating these composition data to the k-th property. Such models are linear in the estimated coefficients, $\underline{b}_k$. First-order models are also linear in the data, $\underline{x}$.

3.0 ESTIMATING UNIVARIATE UNCERTAINTY

As mentioned in Section 2.2, the variance (or standard deviation) of a random variable is a measure of the spread of the distribution of the random variable about its mean. The amount of information conveyed by observations of the random variable is inversely proportional to the variance; i.e., an observation of a random variable with a large variance is less informative than an observation of a random variable with smaller variance. This should be clear from the limiting case: if the variance is zero, the random variable is actually constant, and a single observation conveys all information about the random variable. Thus, variance is a measure of uncertainty. Throughout this document, uncertainty is operationally defined as the variance(s) [or standard deviation(s)] and, in some cases, the covariance(s), of one or more random variables. Thus, estimation and manipulation of uncertainty is discussed here in terms of estimation and manipulation of variances (standard deviations), covariances, and covariance matrices.

3.1 VARIANCES AND VARIANCE COMPONENTS

The simplest approach to estimating the variance of a single random variable X [where $X \sim (\mu,\sigma^2)$] is to obtain n IID observations of X (e.g., a random sample from the population) and to use s_X^2 [Equation (1) of Section 2.4] as an estimate of σ^2. The estimated variance can then be used for a variety of activities, including testing hypotheses and constructing confidence intervals for μ.

This simple case can be formulated differently without changing the essential nature of the problem:

$$Y_i = \mu + \varepsilon_i ,$$

where i = 1, ..., n, and $\varepsilon_i \sim (0,\sigma^2)$. In this case, $Y_i \sim$ IID (μ,σ^2), so estimation and testing proceeds just as in the previous formulation.

An observed or measured quantity may be subject to several separate sources of uncertainty. For example, consider the model[(a)]:

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk} , \tag{7}$$

where $\alpha_i \sim (0,\sigma_\alpha^2)$, $\beta_{ij} \sim (0,\sigma_\beta^2)$, $\varepsilon_{ijk} \sim (0,\sigma^2)$, and all the random variables are uncorrelated. The underlying quantities α_i, β_{ij}, and ε_{ijk} are often called *random effects*. Since the random effects are assumed to be uncorrelated, the uncertainty in each Y_{ijk} is simply

$$\sigma_Y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma^2 .$$

The quantities σ_α^2, σ_β^2, and σ^2 (and estimates of these quantities) are known as *variance components* (or *components of variance*). Observations that share α_i (or α_i and β_{ij}) are correlated and hence are not independent. (The form of the covariance is given below.)

Equation (7) is an example of a *two-way nested random model* (Searle et al., 1992). The term "nested" is applied because the random effects are hierarchical; for example, uncertainty exists among analyses within a single sample, among samples within a single batch, and among batches within a waste type. Each observed, measured, or estimated attribute value includes uncertainty introduced at each level of this hierarchy. The uncertainty at each level in this hierarchy can be represented by a variance. These hierarchical representations of uncertainty are the variance components.

Equation (7) is, in fact, an appropriate statistical model for observations of all attributes (oxide mass fractions, modelled properties, etc.) of material involved in the HLW

[(a)] The models discussed in this and following sections are examples of *linear models*. Linear models are important in a wide variety of statistical disciplines and methodologies, including the analysis of variance, regression, experimental design, variance components, and multivariate statistics. For an introduction to general linear models and their wide applicability, see Searle (1971) or Graybill (1976). In linear models, it is common to use some lower-case Greek letters (e.g., α, β, ε) to represent random variables (or *random effects*) that contribute to the variability in an observed value, while other lower-case Greek letters (e.g., μ, σ^2) are used to represent parameters of the distributions of these and other random variables.

vitrification process.[a] In this case, Y_{ijk} is an observation, measurement, or estimate based on the k-th analysis of the j-th sample from the i-th batch, σ_α^2 represents batch-to-batch variability, σ_β^2 represents within-batch uncertainty, and σ^2 represents analytical uncertainty. The covariance between analyses of the same sample from the same batch is $\sigma_\alpha^2 + \sigma_\beta^2$; the covariance between observations of different samples from the same batch is σ_α^2.

Functions of the observed Y_{ijk} (e.g., various means) will be used to estimate batch and glass attributes. To compare these calculated values to requirements, estimates of the uncertainty in the calculated values will be required. The importance of variance components stems from their crucial role in properly estimating uncertainty in values calculated from a set of observations, Y_{ijk}. Although the uncertainty in a single Y_{ijk} is simply the sum of the three variance components, the uncertainty in a function of several Y_{ijk} can involve the individual variance components in more complicated ways. The form of the proper uncertainty depends on several things, foremost of which is the form of the function of the Y_{ijk}, which is related to the inference that must be drawn.

Most of the requirements imposed on HLW operations and product are *single-batch requirements*, which apply to single process batches. In contrast are the *multiple-batch requirements*, which apply to groups of batches. (See Bryan and Piepel, 1993 and 1994, for definitions, discussion, and examples of single-batch and multiple-batch requirements). The inferences required, and therefore the estimated quantities and proper uncertainties, differ between these two types of requirements.

As an example of a single-batch requirement, consider viscosity at 1150°C, the true value of which should be between 2 and 10 Pascal-seconds for each process batch. Let Y represent viscosity at 1150°C, and assume that: 1) the i-th batch is under examination; 2) b samples are taken from this batch; and 3) n estimates of viscosity are obtained for each

[a] The total uncertainty in estimated values of modelled properties must include the contribution of model uncertainty, but this document focuses on the estimation, manipulation, and effects of composition uncertainty. For these purposes, Equation (7) is adequate for all attributes of material involved in the HLW vitrification process.

sample[a]. One method for checking this requirement is to test whether the true mean viscosity for the i-th batch, $\mu + \alpha_i$, is between 2 and 10[b]. To do this, $\mu + \alpha_i$ can be estimated with the batch mean[c]:

$$\overline{Y}_{i\cdot\cdot} \equiv \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} = \mu \div \alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{ij} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{ijk} \; .$$

Since the target of inference for a single-batch requirement is $\mu + \alpha_i$, the inference should be conditional on (i.e., taking as fixed) the true value of α_i. Thus, to obtain an estimate of the uncertainty in the batch mean, uncertainty due to β_{ij} and ε_{ijk} must be taken into account, but uncertainty in α_i is irrelevant. In this case, the uncertainty in the batch mean is

$$Var(\overline{Y}_{i\cdot\cdot} \mid \alpha_i) = \frac{\sigma_\beta^2}{b} + \frac{\sigma^2}{bn} \; .$$

[a] For example, the j-th sample might be split into n portions. Oxide composition would then be measured for each portion, and the CVS viscosity model would be used to estimate viscosity at 1150°C for each measured composition.

[b] Arguments can be made for other methods of testing viscosity and other attributes. For example, a method that focuses on some percentile of the distribution of the observed values for the attribute in the current batch (e.g., via a tolerance interval), rather than on the mean attribute value in the batch (e.g., via a confidence interval for the mean), might be used. The relative magnitudes of causes of within-batch uncertainty are important in choosing between these methods. The mean-based approach is more appropriate if sampling error is the major constituent of within-batch variability; the percentile-based approach is more appropriate if the true inhomogeneity in the batch is the major constituent of within-batch uncertainty. The preliminary Feed Test Algorithm (Bryan and Piepel, 1994) uses a mean-based approach, on the assumption of perfect mixing of the process batch. If data from actual processing indicates that perfect mixing is not an valid assumption, this issue should be re-visited.

[c] Since the batch effect is assumed to be random rather than fixed, an argument can be made that the best estimator for the quantity $\mu + \alpha_i$ is *not* the batch mean value, but a "shrunken" version of this value. Searle, et al. (1992, Chapter 7, esp. pp. 258-260) discuss this problem. The preliminary Feed Test Algorithm (Bryan and Piepel, 1994) ignores this complication. If testing of the PCC algorithms with the Plant Simulation Code indicates problems, this issue should be re-examined.

An estimate of this uncertainty can be constructed by substituting estimates of the individual variance components into this expression. Therefore, proper estimation of uncertainty in the batch mean requires estimates of the variance components.

Inference for multiple-batch requirements is described by Bryan and Piepel (1994) and by Bryan, Piepel, and Simpson (1994). These tests require an estimate of σ_α^2, the batch-to-batch variability. Thus, for both single-batch and multiple-batch requirements, proper estimation of the uncertainties required for inference requires estimation of individual variance components.

3.2 ESTIMATING VARIANCE COMPONENTS

Methods for estimating variance components are discussed in great detail by Searle et al. (1992). The discussion below focuses on general principles, applicability to HLW vitrification process/product control, and special features of the HLW vitrification process.

The most straightforward method of estimating variance components involves a designed experiment. Assume that such an experiment is performed, that a represents the number of batches examined, that b represents the number of samples taken from each batch, and that n estimates of the attribute value are obtained from each sample. The total number of observations is then abn. These data may be analyzed with the analysis of variance (ANOVA)[a], as in Table 1. The estimates of the individual variance components are

$$\hat{\sigma}_\alpha^2 = \frac{MSA-MSB}{bn}\ , \quad \hat{\sigma}_\beta^2 = \frac{MSB-MSE}{n}\ , \quad \hat{\sigma}^2 = MSE\ , \qquad (8)$$

[a] The analysis of variance, or ANOVA, is a well-known and widely-used statistical procedure. ANOVA is discussed in most books on basic applied statistics (e.g., Snedecor and Cochran, 1980); Graybill (1976) and Searle (1971) present extensive theoretical treatments of ANOVA.

TABLE 1. Analysis of Variance Table for the Two-Way Nested Random Model, $Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Batch-to-Batch	a-1	$SSA \equiv bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2$	MSA = SSA / (a-1)	$bn\sigma_\alpha^2 + n\sigma_\beta^2 + \sigma^2$
Within-Batch	a(b-1)	$SSB \equiv n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2$	MSB = SSB / a(b-1)	$n\sigma_\beta^2 + \sigma^2$
Analytical	ab(n-1)	$SSE \equiv \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	MSE = SSE / ab(n-1)	σ^2
Total	abn-1	$SST \equiv \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{\cdot\cdot\cdot})^2$		

where the symbols used above are defined in Table 1. These ANOVA-based estimators can be derived by setting the sample-based quantities MSA, MSB, and MSE equal to their expectations (the "Expected Mean Squares" of Table 1) and solving for σ_α^2, σ_β^2, and σ^2.[a]

The estimators of variance components given in Equation (8) belong to the class of *ANOVA estimators.*[b] Searle et al. (1992) discuss other methods for estimating variance components, including other ANOVA estimators, maximum likelihood estimation, restricted maximum likelihood estimation, minimum norm quadratic unbiased estimation (e.g., MINQUE and MIVQUE), and Bayes procedures. One application of the last approach is given in Section 9, but complete enumeration and elucidation of the wide variety of techniques for variance component estimation is beyond the scope of this document. The choice of estimation technique must depend on the structure of the available data and the assumptions about the process that are considered realistic at the time. Therefore, this choice must be made when data become available, as part of the data analysis process. Some considerations in the choice of specific technique are discussed below.

It is common to assume that random effects in a linear model follow a normal distribution. In fact, some distributional assumption is required for several of the variance component estimation techniques mentioned above (e.g., maximum likelihood and Bayes procedures). The ANOVA estimators are exceptions -- they are derived simply by equating mean squares to their expected values and therefore depend only on the moments of the

[a] The *method of moments* is a technique for deriving statistical estimators in which sample-based quantities are set equal to their expectations and the resulting equations are solved for the parameters in terms of the sample-based quantities (Lindgren, 1976). Thus, the ANOVA-based estimators discussed here are examples of *method of moments estimators*.

[b] This document follows the convention of Searle et al. (1992) in applying the term "ANOVA estimators" to any estimators derived by applying the method of moments to quantities involved in an ANOVA. In some cases (e.g., when the data are not balanced, as discussed at the end of this section), different types of ANOVA may be legitimately applied to the same data. Different method of moments may result from these different ANOVAs. Therefore, ANOVA estimators are not necessarily unique.

underlying distribution(s). The role of the normality assumption is discussed further in Section 10.1 and in Bryan and Piepel (1994).

In the preceding discussion, it was assumed that the variance components were to be estimated from *balanced data*; i.e., that the same number of samples were taken from each batch, and that the same number of analyses were performed on each sample. This assumption greatly simplifies the form of the ANOVA estimators. In fact, with balanced data, several of the other variance component estimation methods yield estimates identical or closely related to the ANOVA estimates. Unfortunately, this is not the case if the data are unbalanced; in fact, for unbalanced data, several reasonable types of ANOVA estimators exist.

ANOVA estimates of variance components can be negative. This is troubling, since the true values of variance components, by their nature as variances, must be nonnegative. Searle et al. (1992, pp. 129-131) discuss various options for dealing with this problem. One is simply to set a negative estimate to zero (thereby concluding that the related random effect contributes no uncertainty to the observed value); a second is to use one of the methods that guarantee nonnegative estimates (e.g., maximum likelihood and Bayes procedures).

4.0 ESTIMATING MULTIVARIATE UNCERTAINTY

For purposes of predicting melt/glass properties, the composition of an HLW vitrification process batch is currently planned to be quantified by the mass fractions of nine individual oxides and a catchall tenth category, "Others." As noted in Section 2.3, this is an example of multivariate data -- each observation of batch composition comprises measurements of 10 individual quantities. For multivariate data, interest lies in the simultaneous behavior of the individual components of each datum, and uncertainty is usually represented by a covariance matrix for the random vector (the set of individual random variables) of interest. Thus, composition uncertainty can be represented by the covariance matrix associated with the vector of oxide mass fractions. Methods for estimating the covariance matrix from a set of n IID observations are given in Equations (2) and (3) of Section 2.4.

Anderson and Piepel (1993) studied the effect of composition uncertainty on uncertainty in the resulting estimated property values by propagating three different composition covariance matrices through two property models. This study demonstrated the necessity of accounting for covariance in estimating uncertainty for property values. Ignoring covariance structure (i.e., accounting only for variances of the individual oxide mass fractions) led to underestimation of uncertainty in the estimated property value in nine of twelve cases considered. There was an average of 27% underestimation in those nine cases, and an average of 26% overestimation in the remaining three cases. Thus, ignoring covariance structure can seriously affect estimation of composition uncertainty (propagated into property units).

Equation (7) of Section 3.1 presents a hierarchical model for uncertainty in univariate observations. A model of that form can be applied to each of the 10 components used to quantify the composition of an HLW vitrification process batch. In fact, the 10 individual models can be expressed concisely using matrix notation:

$$\underline{X}_{ijk} = \underline{\mu} + \underline{\alpha}_i + \underline{\beta}_{ij} + \underline{\varepsilon}_{ijk} \qquad (9)$$

where $\underline{X}$ is the vector of individual oxide mass fractions ($\underline{X}_i$, i = 1, ..., 10), and $\underline{\alpha}_i$, $\underline{\beta}_{ij}$, and $\underline{\varepsilon}_{ijk}$ are vectors of random effects. In this multivariate generalization of Equation (7), it is assumed that $\underline{\alpha}_i \sim (\underline{0},\Sigma_\alpha)$, $\underline{\beta}_{ij} \sim (\underline{0},\Sigma_\beta)$, $\underline{\varepsilon}_{ijk} \sim (\underline{0},\Sigma)$, and the random vectors $\underline{\alpha}_i$, $\underline{\beta}_{ij}$ and $\underline{\varepsilon}_{ijk}$ are uncorrelated. In analogy to the univariate case, the covariance matrices Σ_α, Σ_β, and Σ are known as *covariance components* (or *components of covariance*).

Just as for univariate variance components, the importance of multivariate covariance components lies in their crucial role in estimating uncertainty in values calculated from a set of observations, $\underline{X}_{ijk}$. Searle et al. (1992) discuss estimation of covariance components. The method to be used in HLW vitrification process/product control is based on the methods for univariate variance component estimation discussed in Section 3.2 and on the well-known formula for the variance of a sum of two random variables (see, for example, Lindgren, 1976, p. 137):

$$Var(X_i+X_j) = Var(X_i) + Var(X_j) + 2\ Cov(X_i,X_j)\ ,$$

from which is easily derived:

$$Cov(X_i,X_j) = \frac{1}{2}\{Var(X_i+X_j) - [Var(X_i)+Var(X_j)]\}\ .$$

To obtain estimates of the components of covariance between X_i and X_j, the three univariate variance components (σ_α^2, σ_β^2, σ^2; see Section 3.1) are estimated for each of X_i, X_j, and the sum, $X_i + X_j$, and the above formula is applied. Performing this estimation for each pair (i,j), j > i, "fills in" the upper half of each of the matrices of covariance components; the lower half of each matrix is derived from the symmetricity of covariance matrices.

5.0 ESTIMATING UNCERTAINTY FROM PROCESS KNOWLEDGE

The HLW process/product control strategy and algorithms require uncertainty estimates for each process batch, but the sampling effort for each process batch is unlikely to be extensive enough to support separate estimation of uncertainties for each batch. Therefore, reasonable estimates of uncertainties must be obtained from other sources during the early stages of plant operation.[a] Such estimates might be developed (using the methods for data-based estimation of uncertainty discussed in Sections 3 and 4) from data obtained from experiments, cold testing, and qualification runs performed during technology development and demonstration or during plant construction and testing, or from data on similar vitrification processes at other sites (e.g., Savannah River and West Valley). It is possible, however, that the data available before commencement of vitrification operations may be inadequate or unsuitable for proper estimation of uncertainty. In any case, such data are not currently available, and uncertainty estimates are required for development and testing of the PCC algorithms. One method of filling this void, estimating uncertainty from assumptions and/or knowledge of the vitrification and analysis process, is presented in this section.

To estimate composition uncertainty (or a component thereof) from process knowledge, the process must be modelled, and assumptions and knowledge about the process must be translated into numerical statements. The general method consists of: 1) formulating a mathematical model for the process by which data will be obtained (e.g., the sampling and analytical processes), 2) assigning a reasonable uncertainty to each parameter in this model, and 3) propagating the uncertainties in the parameters through the model to yield an estimate of composition uncertainty.

If the mathematical model is simple enough, composition uncertainty may be estimated analytically (i.e., using explicit mathematical error propagation via the general

[a] This problem will fade in importance as an operational database accumulates, at least within a single waste type. If the assumption of the stability of various uncertainties over the several waste types is untenable, the problem will reappear at the beginning of processing of each waste type.

method presented in Section 7). Unfortunately, due to the complicated nature of the measurement process and the multivariate nature of the data, it is likely that estimation of composition uncertainty from a model of the measurement process will have to be done by the Monte Carlo method. The Monte Carlo method involves: 1) constructing a large number of simulated compositions using the assumed mathematical model and parameter uncertainties, and 2) computing the resulting covariance matrix via Equation (3) of Section 2.4.[a] This method is illustrated in Section 5.1. A method for deriving a measure of strength of belief in a simulated uncertainty estimate is discussed in Section 5.2.

5.1 ESTIMATING UNCERTAINTY USING THE MONTE CARLO METHOD

The general method for estimating uncertainty from process knowledge requires formulating a detailed mathematical model of the measurement process. The exact measurement process to be used in HLW vitrification is not yet known, so a hypothetical process must be used. In fact, a model of the full measurement process (including both sampling and analysis) is not required for illustrating the general principles; a model of only the analytical process suffices to illustrate the principles involved. Therefore, the general method is here illustrated using a model of a hypothetical analytical process to yield a hypothetical estimate of analytical uncertainty. *What follows is only an illustration of the general method for estimating uncertainty from process knowledge -- there is no guarantee that the actual HLW analytical process will be as follows or that the process uncertainties assumed here are representative of the uncertainties in the actual HLW analytical process.*

The analytical process here assumed to be used for estimating composition from a single sample from a process batch is as follows:

1) Extract the i-th subsample and obtain its mass in grams, M_i.

[a] In fact, a slight modification of Equation (3) might be required in some cases. If the assumed measurement process is unbiased and the number of simulated compositions is large, this modification is trivial and is not considered further.

2) Prepare the subsample, e.g., by dissolution and dilution, to produce a solution of size C_i, where C_i is measured in liters (or ml).

3) Extract the j-th aliquot from this solution and quantify its size in liters, C_{ij}.

4) Measure the amounts of each of the p cations[a] of interest in the j-th aliquot of the i-th solution, yielding A_{ijk}, k = 1, ..., p. These amounts are assumed to be counts of atoms, expressed in moles.[b]

5) For each cation k, k = 1, ..., p, convert from moles per liter, A_{ijk}, to raw[c] (unnormalized) oxide mass fraction, χ_{ijk} [i.e., (mass of oxide) per (mass of subsample)], using a stoichiometric constant, λ_k:[d]

$$\chi_{ijk} = \frac{A_{ijk}\lambda_k C_i}{M_i C_{ij}} . \qquad (10)$$

6) Normalize[e] the oxide mass fractions so that they sum to one:

[a] The assumptions being made here include: 1) composition of the final glass is adequately represented solely in terms of oxides; 2) the analytical procedure produces measurements for all cations in the final glass; and 3) no significant losses of measured cations occur during processing. In fact, in the simulation, the component "Others" was treated as a single entity, with a "molecular weight" reflecting the proportions of minor species given by Hrma, Piepel, et al. (1992) for neutralized current acid waste.

[b] Measurements of cation amounts in units of mass (g or mg) are here assumed to be equivalent to moles, since atomic weights are known with much less error than that arising in almost all measurement processes. If the measurements are expressed in units of mass, the units associated with λ_k must be modified accordingly.

[c] These mass fractions are raw in the sense that, due to various errors in the measurement process, they are unlikely to sum to one.

[d] These constants relate oxide mass to cation quantity (moles or mass) [i.e., (mass of oxide) per (moles or mass of cation)] and are assumed to be known without error (or with very small error).

[e] Both logic and the statistical models used for estimating property values from composition data require that mass fractions sum to one. The normalization performed here is crude -- it does not take into account what is known about the variances and covariances of the quantities being normalized. An approach that takes account of variances is discussed by Deming (1943) and Mandel (1964). Unfortunately, this approach: 1) does not take account of covariances, and 2) is found, empirically, to produce larger variances of predicted property values. Proper techniques for normalizing compositional data require further study.

$$x_{ijk} = \frac{\chi_{ijk}}{\sum_{k=1}^{p} \chi_{ijk}} .$$

Two more sets of quantities must be specified in order to perform the simulation: 1) the "true" composition for which the hypothetical analytical process will be performed, i.e., the vector of assumed oxide mass fractions; and 2) the uncertainties introduced at each stage of the assumed analytical process.

The "true" composition used in this illustration is that of the CVS Internal Standard Glass (Hrma, Piepel et al., 1992). The oxide mass fractions for this composition appear in Table 2.

TABLE 2. Oxide Mass Fractions and Hypothetical Uncertainty Estimates for the CVS Internal Standard Glass

Oxide	Mass Fraction	Standard Deviation	Relative Standard Deviation (%)
SiO_2	0.5328	0.005383	0.99
B_2O_3	0.1048	0.002209	2.10
Na_2O	0.1129	0.002366	2.04
Li_2O	0.0373	0.000834	2.14
CaO	0.0082	0.000187	1.22
MgO	0.0084	0.000192	1.19
Fe_2O_3	0.0733	0.001585	2.05
Al_2O_3	0.0235	0.000529	2.13
ZrO_2	0.0392	0.000874	2.04
Others	0.0596	0.001306	2.18
Total	1.0000		

The calculation of unnormalized mass fraction for cation k, χ_{ijk}, involves four uncertain quantities: 1) mass of the i-th subsample, M_i, 2) mass of the i-th solution, C_i, 3) mass of the j-th aliquot, drawn from the i-th solution, C_{ij}, and 4) moles of the k-th cation, A_{ijk}. Uncertainty in the measured composition is introduced at each of these stages, and, to the extent that characteristics are measured on the same aliquot (or solution, or subsample), correlation is likely to be introduced. For this illustration, a 2% relative standard deviation (RSD) in the measurement of each uncertain quantity was assumed, and these quantities were assumed to follow normal distributions and to be independent (since the measurement processes are quite distinct). The independence of underlying uncertainties *does not imply that the resulting measurements of oxide mass fractions are independent* -- if the same subsample, solution, or aliquot is used to produce data for more than one oxide, some correlation is introduced. In addition, correlation is introduced in the normalization process. For this illustration, it was assumed that measurements of cation quantities were taken from the same aliquot (which implies the same solution and subsample as well). Since measurements taken from the same aliquot (or solution, or subsample) are expected to be more strongly related than those taken from separate aliquots (or solutions, or subsamples), this assumption should lead to the "strongest" correlation patterns.

The simulation was carried out by generating 100,000 "observations" from the analytical process described above and then computing the empirical (Monte Carlo) covariance and correlation matrices. Standard deviations and RSDs appear in Table 2; the correlation matrix appears in Table 3[(a)]. These results are used as "true" values for illustrations and examples in the sections that follow.

The uncertainty estimates presented in Tables 2 and 3 model only analytical uncertainty. Similar techniques could be used to produce estimates of batch-to-batch variability and within-batch uncertainty.

[(a)] Due to greater ease of interpretation, standard deviations and correlation matrices, rather than covariance matrices, are presented here and below. It is possible to recover a covariance matrix from the associated correlation matrix and standard deviations; and *vice versa.*

TABLE 3. Correlation Matrix from Hypothetical Analytical Process

	SiO_2	B_2O_3	Na_2O	Li_2O	CaO	MgO	Fe_2O_3	Al_2O_3	ZrO_2	Others
SiO_2	1.000	-0.598	-0.618	-0.447	-0.382	-0.384	-0.524	-0.415	-0.447	-0.492
B_2O_3	-0.598	1.000	0.092	0.156	0.172	0.170	0.125	0.162	0.148	0.140
Na_2O	-0.618	0.092	1.000	0.149	0.167	0.170	0.120	0.161	0.145	0.127
Li_2O	-0.447	0.156	0.149	1.000	0.216	0.217	0.176	0.206	0.195	0.180
CaO	-0,382	0.172	0.167	0.216	1.000	0.230	0.194	0.218	0.214	0.199
MgO	-0.384	0.170	0.170	0.217	0.230	1.000	0.194	0.223	0.212	0.204
Fe_2O_3	-0.524	0.125	0.120	0.176	0.194	0.194	1.000	0.178	0.174	0.158
Al_2O_3	-0.415	0.162	0.161	0.206	0.218	0.223	0.178	1.000	0.206	0.189
ZrO_2	-0.447	0.148	0.145	0.195	0.214	0.212	0.174	0.206	1.000	0.180
Others	-0.492	0.140	0.127	0.180	0.199	0.204	0.158	0.189	0.180	1.000

5.2 QUANTIFYING STRENGTH OF BELIEF IN SIMULATED UNCERTAINTY ESTIMATES

Methods for combining and updating uncertainty estimates (discussed in Sections 6 and 9, respectively) require some measure of the strength of belief[a] in the uncertainty estimates in order to assign relative weights. Therefore, if Monte Carlo estimates of uncertainty are to be used in HLW process/product control, measures of strength of belief in these estimates must be developed. One approach to assigning strength of belief to univariate uncertainty estimates is discussed below. If simulated estimates of multivariate uncertainty are to be used in PCC testing or HLW operations, this or some other approach should be extended to the multivariate situation.

Consider the problem of estimating variance from a random sample of n observations, X_i, i = 1, ..., n. Under the assumption of normality [i.e., X_i, i = 1, ..., n, ~ IID $N(\mu,\sigma^2)$], the standard estimator of variance, S^2, follows (a multiple of) a chi-square distribution with $f = n\text{-}1$ degrees of freedom (Lindgren, 1976, p. 334, Theorem 4). Since the mean and variance of a chi-square distribution are f and 2f, respectively, the RSD of S^2 is

[a] The discussion here is phrased in terms of "strength of belief," because discussion of "uncertainty in estimates of uncertainty" is more likely to lead to confusion. The word "confidence" is also avoided in this context, since "confidence" has a specific technical meaning in statistical applications.

$$rsd(s^2) \equiv \frac{\sqrt{Var(s^2)}}{E(s^2)} = \sqrt{\frac{2}{n-1}} = \sqrt{\frac{2}{f}} .$$

For example, the RSD of a variance estimate based on 50 degrees of freedom is 20%. Thus, the relative precision (as measured by RSD) in an estimated variance is a function of the associated degrees of freedom. This line of thought can be reversed to assign a strength of belief, measured by the number of degrees of freedom, to a simulated uncertainty estimate: if it is believed that the simulation is yielding a variance within 100p% of the true value, $f = 2/p^2$ would be taken as the associated number of degrees of freedom. This approach could be refined by taking into account the strength of belief in the estimate of relative precision (see the discussion in Section 8.0).

6.0 COMBINING SOURCES OF UNCERTAINTY

When observed data are subject to more than one source of uncertainty, proper estimation of the uncertainty in a function (e.g., the mean) of these observations requires combining variance components (for example, see Section 3.1). In addition, estimating variance components often requires combining mean squares (see Section 3.2), and the PCC algorithms must combine model uncertainty with composition uncertainty. In many cases (including all so far identified for HLW vitrification process/product control), the required combination of mean squares or variance components is a weighted sum, where the weights are related to the distribution of sampling effort (e.g., the number of samples per batch and the number of analyses per sample) or strength of belief in the individual variance components. In general, such weighted sums take the form:

$$s_c^2 = \sum_{j=1}^{p} c_j s_j^2 \ ,$$

where s_c^2 is the required combination of the individual variance components, s_j^2, with weights c_j.

Some measure of the quality of s_c^2 must be available in order to use this estimate to draw inferences. As discussed in Section 5.2, the quality of a variance estimate is often quantified by the associated degrees of freedom. The weighted sum above incorporates several variance estimates, each with an associate number of degrees of freedom, f_j. What number of degrees of freedom should be associated with the combined variance estimate, s_c^2? The answer to be used by the PCC algorithms is that given by Satterthwaite (1946); the degrees of freedom to be associated with s_c^2 is

$$f_c = \frac{\left(\sum_{j=1}^{p} c_j s_j^2\right)^2}{\sum_{j=1}^{p} \frac{\left(c_j s_j^2\right)^2}{f_j}} = \frac{\left(s_c^2\right)^2}{\sum_{j=1}^{p} \frac{\left(c_j s_j^2\right)^2}{f_j}} \ .$$

Satterthwaite's approximation, as the above formula is known, was derived under the assumption of normality. Caution should be exercised in applying this formula when some of the c_j are negative (which is often the case when estimating variance components). Methods and additional requirements in this case are discussed by Gaylor and Hopper (1969), who show that the approximation is adequate when the component (or the sum of the several components) being subtracted is relatively small.

7.0 PROPAGATING MULTIVARIATE UNCERTAINTY

Many of the batch and glass attributes that must be estimated and checked as part of HLW vitrification process/product control will be calculated as functions of more than one uncertain quantity (e.g., oxide mass fractions, other process measurements, empirical model coefficients). In order to check compliance of these attributes with process and product specifications, an estimate of the total (univariate) uncertainty associated with each such attribute value must be obtained. Therefore, a procedure for combining multivariate uncertainties to yield univariate uncertainties is required. The role of the resulting univariate uncertainty estimates in constructing tests for the acceptability of a feed batch is discussed by Bryan and Piepel (1994).

The procedure described in Section 7.1 is one form of *error propagation* (or *propagation of error*). The general procedure can be used to estimate uncertainty for a wide variety of functions of uncertain quantities. For HLW process/product control, the uncertain quantities fall into two categories: 1) composition and other process measurements, and 2) empirical model coefficients. This document focuses on the former category, but the method for incorporating uncertainty due to the latter category is briefly discussed. Section 7.2 examines the contribution of composition uncertainty to overall (univariate) uncertainty for several modelled properties. Section 7.3 discusses an alternative approach to estimating univariate uncertainty for attributes that are calculated as functions of more than one oxide mass fraction or other process measurement.

7.1 A METHOD OF ERROR PROPAGATION

The basis for the error propagation method to be used by the PCC algorithms is as follows. Let y represent the characteristic of interest, and assume that y = f($\underline{z}$), where $\underline{z}$ is a random vector with mean $\underline{\mu}_z$ and covariance matrix Σ_z. Then, using a Taylor series expansion about $\underline{\mu}_z$ to approximate f($\underline{z}$), an approximation to the variance of y, σ_y^2, can be derived:

$$\sigma_y^2 \approx \underline{d}_z^T \Sigma_z \underline{d}_z \ ,$$

where $\underline{d}_z$ is the gradient of f (i.e., the vector of partial derivatives with respect to $\underline{z}$), evaluated at the observed value of $\underline{z}$.

As discussed above, two distinct sources of uncertainty enter into the uncertainty associated with a modelled batch or glass property (y): one associated with the estimated coefficients ($\underline{b}$) of the empirical model, the other associated with the estimated composition ($\underline{x}$).[a] Model uncertainty will be represented by the covariance matrix, Σ_b, for the vector of estimated model coefficients (which will be obtained from CVS; e.g., Hrma, Piepel, et al., 1994). For simplicity of presentation, it is assumed here that a single covariance matrix representing composition uncertainty, Σ_x, is available. The case of several covariance components for feed composition is discussed at the end of this section.

The general method of error propagation discussed above can be applied to the case in which the random vector $\underline{z}$ consists of two distinct subvectors, e.g., the case in which y = f($\underline{x}$,$\underline{b}$). Denote the gradients of f($\underline{x}$,$\underline{b}$) with respect to $\underline{x}$ and $\underline{b}$ by $\underline{d}_x$ and $\underline{d}_b$, respectively. If $\underline{x}$ and $\underline{b}$ are *uncorrelated* random vectors (a reasonable assumption unless $\underline{x}$ is part of the data used to estimate $\underline{b}$), the approximate variance of y divides neatly into two parts, one attributable to composition uncertainty, the other attributable to model uncertainty:

$$\sigma_y^2 \approx \underline{d}_x^T \Sigma_x \underline{d}_x + \underline{d}_\beta^T \Sigma_\beta \underline{d}_\beta \ .$$

For the special case where the function f($\underline{x}$,$\underline{b}$) is linear in both the data, $\underline{x}$, and the estimated coefficients, $\underline{b}$, this formula takes on an even simpler form. For this case, y = $\underline{x}^T\underline{b}$, $\underline{d}_x = \underline{b}$, $\underline{d}_b = \underline{x}$, and

$$\sigma_y^2 \approx \underline{\beta}^T \Sigma_x \underline{\beta} + \underline{x}^T \Sigma_\beta \underline{x} \ .$$

Since the (approximate) uncertainty in y can be separated into two parts (one due to composition uncertainty; the other due to model uncertainty), the PCC algorithms will calculate the two contributions to uncertainty in y separately, to produce two univariate

[a] If the property model is second-order, the vector $\underline{x}$ contains not only the individual mass fractions, but also some cross-products.

estimates of components of uncertainty in the estimated property value. These two univariate uncertainty estimates will then be combined and an associated number of degrees of freedom assigned, as described in Section 6. This approach has the advantage of easy generalization to the case of several covariance components relevant to composition uncertainty: the composition covariance components will be propagated separately and the resulting univariate variance components will be combined to form a univariate estimate of overall composition uncertainty (in property units). Again, the method given in Section 6 will be used to combine the univariate components of composition uncertainty and to assign an associated number of degrees of freedom.

7.2 CONTRIBUTION OF COMPOSITION UNCERTAINTY TO OVERALL UNCERTAINTY

The method of error propagation described above was used to investigate the contribution of composition uncertainty to uncertainty in estimated property values for each of five properties for which CVS is developing models: viscosity at 1150°C, electrical conductivity at 1150°C, and normalized release of boron, lithium, and sodium from the Product Consistency Test (PCT). Model uncertainties were not accounted for in these calculations. The models used in the error propagation were the first-order CVS models given by Hrma, Piepel, et al. (1994). These models actually predict the *natural logarithm* (ln) of each property. Since the standard deviation of ln(Y) can be shown to be approximately equal to the RSD of Y, error propagation using models that predict ln(Y) yields estimates of RSDs on the original property scales.

Two composition covariance matrices were used in this investigation (and in investigations discussed in later sections). The first was the hypothetical analytical covariance matrix described in Section 5.1. The second was that derived from the Corning RR6 data, as described by Anderson and Piepel (1993). The mean oxide mass fractions and standard deviations from the Corning RR6 data set appear in Table 4; the associated correlation matrix appears in Table 5. Like the hypothetical covariance matrix, the Corning RR6 covariance matrix is an estimate of analytical uncertainty, but the Corning laboratory seems to be very

TABLE 4. Oxide Mass Fractions and Uncertainty Estimates for the Corning RR6 Data

Oxide	Mass Fraction	Standard Deviation	Relative Standard Deviation (%)
SiO_2	0.4787	0.000878	0.17
B_2O_3	0.0866	0.000416	0.46
Na_2O	0.1148	0.000354	0.26
Li_2O	0.0321	0.000162	0.31
CaO	0.0143	0.000081	0.00
MgO	0.0085	0.000047	0.00
Fe_2O_3	0.1402	0.000648	0.43
Al_2O_3	0.0472	0.000170	0.21
ZrO_2	0.0013	0.000047	0.00
Others	0.0763	0.000212	0.26
Total	1.0000		

TABLE 5. Correlation Matrix from Corning RR6 Data

	SiO_2	B_2O_3	Na_2O	Li_2O	CaO	MgO	Fe_2O_3	Al_2O_3	ZrO_2	Others
SiO_2	1.000	-0.243	-0.421	-0.337	-0.510	-0.652	-0.632	0.567	-0.367	-0.804
B_2O_3	-0.243	1.000	0.240	0.154	-0.307	-0.141	-0.430	-0.144	0.004	0.102
Na_2O	-0.421	0.240	1.000	0.306	-0.104	0.036	-0.127	-0.310	-0.104	0.059
Li_2O	-0.337	0.154	0.306	1.000	-0.201	-0.054	-0.079	-0.042	0.060	0.173
CaO	-0.510	-0.307	-0.104	-0.201	1.000	0.629	0.733	-0.344	0.290	0.492
MgO	-0.652	-0.141	0.036	-0.054	0.629	1.000	0.697	-0.512	0.414	0.682
Fe_2O_3	-0.632	-0.430	-0.127	-0.079	0.733	0.697	1.000	-0.525	0.279	0.602
Al_2O_3	0.567	-0.144	-0.310	-0.042	-0.344	-0.512	-0.525	1.000	-0.026	-0.459
ZrO_2	-0.367	0.004	-0.104	0.060	0.290	0.414	0.279	-0.026	1.000	0.384
Others	-0.804	0.102	0.059	0.173	0.492	0.682	0.602	-0.459	0.384	1.000

precise (note the very low RSDs in Table 4). Uncertainty estimates derived from the Corning RR6 data are probably too optimistic (i.e., too low) for HLW vitrification operations and should be viewed as lower bounds.

Table 6 presents the contributions of composition uncertainty to uncertainty in modelled property values for each of the two composition covariance matrices. The RSDs yielded by the hypothetical covariance matrix range from 3.5% to 8.6%, while those yielded by the Corning RR6 covariance matrix range from 0.6% to 1.5%, reflecting the high precision attributed earlier to the Corning laboratory.

TABLE 6. RSDs for Modelled Properties Derived from Hypothetical and Corning RR6 Covariance Matrices

Property	**Relative Standard Deviation Corresponding to**	
	Hypothetical Covariance Matrix	**Corning RR6 Covariance Matrix**
Viscosity at 1150°C	8.61%	1.53%
Electrical Conductivity at 1150°C	3.51%	0.61%
PCT B	7.29%	1.51%
PCT Li	5.74%.	1.25%
PCT Na	7.20%	1.47%

7.3 AN ALTERNATIVE TO ERROR PROPAGATION FOR COMPOSITION UNCERTAINTY

The major purpose of error propagation in HLW vitrification process/product control is collapsing various multivariate uncertainties (composition and model uncertainties) to a univariate uncertainty for each attribute that is a function of more than one uncertain quantity (e.g., oxide mass fractions, other process measurements, and empirical model coefficients). For estimating the contribution of composition uncertainty to the uncertainty of such an attribute, a possible alternative is the following: 1) apply the function (e.g., the CVS property model) to each measured composition separately, yielding several estimates of the attribute value; and 2) use the methods of univariate uncertainty estimation (discussed in Section 3) to estimate the uncertainty directly from the estimated values, or use the method of updating a univariate uncertainty estimate (discussed in Section 9) to combine the current information with information from previous batches.

If this alternative method is employed, error propagation will still have to be performed for modelled properties, in order to estimate the contribution of model uncertainty to the overall uncertainty in each property. In addition, the requirement to perform variance component estimation for each univariate uncertainty may impose a significant computational burden. A method for updating the composition covariance matrix (such as that provided by the Kalman filter and/or the existing Measurement Error Model), combined with the general error propagation method of Section 7.1 (for both composition and model uncertainty) and the method of assessing strength of belief in the propagated uncertainty (as detailed in Section 6) is probably preferable.

8.0 SAMPLE SIZES REQUIRED FOR ESTIMATING UNCERTAINTY

Understanding the methods used to calculate the number of samples (observations, measurements, etc.) needed to estimate a parameter requires understanding the ways in which a statistical estimation procedure can fail. Many estimation procedures produce an interval of values within which the parameter is claimed to lie. These procedures can fail in two ways: 1) the interval may not contain true value of the parameter, or 2) the interval may be too wide to be useful. To properly estimate sample size, the rates of both types of failure should be controlled. Therefore, at least two quantities must be specified: 1) the desired precision in the result, i.e., the maximum acceptable width of the interval, and 2) the desired level of statistical *confidence* in the result.

Statistical confidence is a measure of the success rate of the statistical procedure, i.e., the probability that the procedure produces a correct answer. For example, a properly-designed procedure to produce a 95% confidence interval for some parameter must have at least a 95% chance (probability ≥ 0.95) of producing an interval that actually traps the unknown true value of the parameter.[a]

The actual width of a statistical interval is a function of the data upon which the interval is based, and therefore this width is a random variable. The best procedures for estimating sample size also require specification of a probability with which the desired precision is to be achieved. This probability might be called the *stability* of the interval.

[a] Statistical confidence is a measure of the long-run performance of the statistical procedure, not of the probability that any particular application of the procedure results in success. In other words, confidence rests in the procedure, not in any particular result of the procedure. To illustrate this distinction, assume that data has been collected and used to construct a 95% confidence interval for the mean value of the characteristic of interest. Once this procedure is carried out, the calculated interval either does or does not contain the true mean, *and it is not known which is the case.* On the other hand, if the experiment is repeated many times, resulting in a large number of 95% confidence intervals, then approximately 95% of these intervals will have trapped the true mean.

8.1 SAMPLE SIZES FOR ESTIMATING A VARIANCE

Consider estimating the variance (or standard deviation) of a $N(\mu,\sigma^2)$ population, given a random sample of size n. The standard estimator of variance, S^2, follows a multiple of a chi-square distribution with n-1 degrees of freedom (Lindgren, 1976, p. 334, Theorem 4); specifically,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) .$$

This fact can be used to compute sample sizes required for estimating variance or standard deviation with specified precision. The method detailed below, which focuses on estimating standard deviation, is an adaptation of methods described by Hahn and Meeker (1991, pp. 141-144).

It is desired to find the smallest sample size n such that, with probability at least equal to 1-α, the sample standard deviation, S, is within 100p% of σ, i.e.,

$$Pr\{(1-p)\sigma \leq S \leq (1+p)\sigma\} \geq 1-\alpha .$$

In this case, 100(1-α)% is the statistical confidence and 100p% is the precision[a] associated with the estimation procedure.

By taking advantage of the known distribution of S^2, the problem can be reduced to a search for the smallest n such that

$$1-\alpha \leq Pr\{(n-1)(1+p)^2 \leq \chi^2(n-1)\} - Pr\{(n-1)(1-p)^2 \leq \chi^2(n-1)\} .$$

Sample size results for various combinations of 100(1-α)% (confidence) and p (precision) appear in Table 7. For example, to estimate σ^2 with 20% precision and 90% confidence, a sample of size 20 is required.

[a] Note that, in this context, higher precision corresponds to lower percentages. For example, a precision of 1% is higher than a precision of 20%.

TABLE 7. Sample Sizes Required for Estimating Standard Deviation with Specified Levels of Precision and Confidence

Precision	Sample Sizes Required for Confidence =			
	80%	90%	95%	99%
1%	3488	8174	13507	27084
5%	132	322	538	1088
10%	31	79	134	274
20%	7	20	34	70
30%	2	9	16	32
40%	2	5	9	19
50%	2	4	6	12

8.2 SAMPLE SIZES FOR ESTIMATING A COVARIANCE MATRIX

In a multivariate situation, interest lies in an entire covariance matrix, not just in a single variance. When batch composition is estimated in terms of nine oxide mass fractions and "Others," 10 variances and 45 covariances[(a)] must be estimated. . Although the sample size calculations above hold for each variance (and, as indicated by simulation results, each covariance) individually, the problem of *simultaneous inference* arises in estimating the entire covariance matrix. When desired levels of precision and confidence must be obtained for several quantities simultaneously, the sample size required is larger than that required for a

[(a)] Although a 10 × 10 matrix has 100 elements, the symmetric nature of covariance and correlation matrices reduces the number of distinct elements that must be estimated to 45. In fact, as long as the measured compositions sum to one, only 35 covariances must be estimated -- the remaining 10 covariances can be estimated from the fact that any row or column of the covariance matrix must sum to zero. This is related to the inherent singularity of the covariance matrix for compositional data, which is discussed in Section 10.1. However, this technical point is of little importance here.

single quantity, and this required sample size increases rapidly as the number of estimated quantities increases.

To investigate the extent of the simultaneous inference problem for 10 × 10 matrices, simulation studies were performed to quantify the performance of estimation for two "true" correlation matrices: the hypothetical correlation matrix of Section 5.1, and the Corning RR6 correlation matrix discussed in Section 7.2. Sample sizes of 10[a], 100, and 1000 were used in the simulation. For each combination of "true" correlation matrix and sample size, 1000 data sets were generated. The empirical covariance and correlation matrices were computed for each data set, and the maximum absolute difference between elements of the empirical correlation matrix and corresponding elements of the "true" correlation matrix was recorded.[b] In addition, each empirical covariance matrix was propagated through the first-order CVS property models for (the natural logarithms of) five properties, as discussed in Section 7.2.

Table 8 reports the mean (over the 1000 generated data sets) maximum absolute difference[c] between elements of the estimated and "true" correlation matrices. These results must be interpreted in terms of the absolute magnitudes of the "true" correlations (which appear in Tables 3 and 5). The "true" correlations range between 0.004 and 0.804 in absolute value. With a sample size of only 10, the maximum absolute difference, 0.72 for both "true" correlation matrices, is large enough to imply that some empirical correlations are

[a] In general, a minimum sample size of 10 is required for proper estimation of a 10 × 10 covariance matrix. Due to the special nature of compositional data (the inherent singularity of the covariance matrix), the minimum sample size in this case is 9.

[b] Correlation matrices, rather than covariance matrices, were chosen for these comparisons because the elements of a correlation matrix are constrained to lie between -1 and 1, so that the scale is fixed and absolute differences are easily interpreted. The scale of elements in a covariance matrix depends on the underlying variances, which hinders comparison of differences. A disadvantage of using correlation matrices is that variances are not included in the comparisons.

[c] For investigating the quality of *simultaneous* estimation, attention should be paid to the larger deviations, rather than to the mean deviation.

TABLE 8. Maximum Absolute Deviations in Estimation of Correlation Matrices

Sample Size	Hypothetical Correlation Matrix	Corning RR6 Correlation Matrix
10	0.72	0.72
100	0.23	0.21
1000	0.07	0.07

routinely quite different from the corresponding "true" values. Even with a sample size of 100, estimation is relatively poor: the maximum absolute difference is approximately 0.20 for both "true" correlation matrices.

These results imply that reasonably precise simultaneous estimation of a single 10 × 10 correlation (or covariance) matrix requires large sample sizes. Since required sample sizes increase rapidly as the number of estimated quantities increases, simultaneous estimation of covariance components (i.e., simultaneous estimation of *several* 10 × 10) is expected to require *very* large sample sizes. This is supported by the results of another simulation study (not reported here).

These results for estimation of covariance matrices and covariance components might seem discouraging, but it must be remembered that interest lies not in the multivariate uncertainties (covariances and components of covariance) *per se*, but only in their effects on the uncertainties in modelled properties. As part of the simulation study discussed above, each empirical covariance matrix was propagated through five first-order CVS property models. As discussed in Section 7.2, the CVS models actually predict the natural logarithm of each property. Since the standard deviation of ln(Y) can be shown to be approximately equal to the RSD of Y, error propagation using models that predict ln(Y) yields estimates of RSDs on the original property scales. The relationship of sample size to precision in estimation of RSD for each property can be examined by calculating the standard deviation (over the 1000 replications) of each estimated property RSD for each sample size. This was

TABLE 9. Relative Precision of RSDs[(a)] of Modelled Melt/Glass Property Values for Three Sample Sizes and Two "True" Covariance Matrices

Covariance Matrix	Property	Relative Precision of Melt/Glass Property RSD		
		n = 10	n = 100	n = 1000
Hypothetical	Viscosity at 1150°C	22.3%	7.1%	2.2%
Hypothetical	Electrical Conductivity at 1150°C	22.3%	7.0%	2.2%
Hypothetical	PCT B	22.5%	7.2%	2.2%
Hypothetical	PCT Li	22.5%	7.2%	2.2%
Hypothetical	PCT Na	22.5%	7.2%	2.2%
Corning RR6	Viscosity at 1150°C	23.1%	7.2%	2.2%
Corning RR6	Electrical Conductivity at 1150°C	23.1%	7.2%	2.2%
Corning RR6	PCT B	23.6%	7.0%	2.3%
Corning RR6	PCT Li	23.4%	6.9%	2.3%
Corning RR6	PCT Na	23.6%	6.9%	2.2%

(a) RSDs were obtained by propagating composition covariance matrices through CVS first-order ln(property) models. Standard deviations of ln(property) values are approximately RSDs of property values on the original scale (i.e., without logarithmic transformation).

done for each of the two "true" covariance matrices. The resulting relative precisions (the empirical standard deviation of the estimated RSDs, divided by the known "true" values, taken from Table 6) appear in Table 9.

These results are more encouraging. The relative precision in each estimated RSD (i.e., the univariate measures of uncertainty) improves much more quickly with increasing sample size than does that associated with multivariate uncertainty estimation.[a] Even with a sample size of only 10, the individual univariate uncertainty estimates have relative precisions of about 23%.

An apparent contradiction is lurking in these results. As discussed in Section 4, Anderson and Piepel (1993) demonstrated that ignoring covariances can lead to serious underestimation of uncertainty in modelled properties. However, the results above indicate that propagation of even quite poor estimates of multivariate uncertainty can lead to acceptable precision of uncertainty estimation for univariate properties. The resolution of this dilemma seems to lie in the tendency of poor estimates of individual elements of a covariance matrix to offset one another when propagated. Further pursuit of this theoretical point is beyond the scope of this document.

For some insurance against the simultaneous inference problem, it is suggested that uncertainties be estimated from sample sizes greater than 20 (which, according to Table 7, corresponds to a 90% confidence and 20% precision).

[a] Basic statistical theory for simple estimation problems suggests that the standard deviation of an estimator should decrease (i.e., precision should increase) at a rate proportional to the square root of sample size. In other words, the standard deviation of an estimator based on n_1 samples should be approximately $\sqrt{(n_2/n_1)}$ times the standard deviation of an estimator based on n_2 samples. It is interesting to note that this pattern appears in Table 9.

9.0 UPDATING ESTIMATES OF COMPOSITION UNCERTAINTY

Estimates of composition uncertainty (variances, covariances, and matrices thereof) should be updated to reflect the information that becomes available with each process batch. One obvious way to achieve this updating is to maintain a database of composition and other measurements, sample sizes, tank level measurements, and other results for each process batch, and to re-estimate all required quantities at each step, using the methods described in Sections 3 and 4. Implementing this approach may be quite cumbersome (due to the computational burden of re-computing components of variance and covariance).

One alternative is use of a method for updating composition uncertainties (covariance matrices and components of covariance) for each batch. These updated composition uncertainties would then be propagated through property models to yield uncertainty estimates for modelled properties (as in Section 7.1; see also the discussion in Section 7.3). The existing Measurement Error Model (MEM) updates covariance matrices (in addition to its main function of data reconciliation). It may be possible to improve the MEM by combining it with a Kalman filter or some other Bayesian approach (Bayesian approaches are discussed below). Adams (1994) describes and compares the basic Kalman filter and the MEM. This approach would require more investigation and development.

A computationally simpler univariate alternative is described in this section. The illustration concerns estimating a univariate variance. The relationship between variances and covariances discussed in Section 4 should allow application of this technique to covariance estimation. In addition, the technique may be used for updating variance and covariance components.

The relative performance of these alternative methods could be investigated (e.g., by simulation), but this would require extensive additional effort. The PCC algorithms will implement the simple technique described below, and testing (using the Plant Simulation Code, as described by Bryan and Piepel, 1993) will indicate the potential rewards of developing and implementing one of the alternative methods.

In updating estimates, prior information must be combined with information contained in a current data set. Combining prior and current information is one application of the branch of statistics known as *Bayesian statistics.* In the Bayesian approach, as in other branches of statistics, data are modelled as realizations of a random variable, with an associated statistical distribution, known as the *likelihood.* The parameters of the likelihood (or functions thereof) are usually the target of inference. In Bayesian statistics, these parameters are also modelled as random variables. Hence, a statistical distribution, known as the *prior distribution,* is associated with each parameter. This prior distribution is chosen to reflect information and beliefs about the parameter of the likelihood. One method is to choose a general distributional form, a mean, and a standard deviation for the parameter. In this case, the mean value represents the best guess of the true value of the parameter, and the standard deviation reflects the uncertainty about this guess.

The Bayesian approach combines prior information (in the form of the prior distribution) with the current data (in the form of the likelihood) to produce a *posterior distribution,* an updated statistical distribution for the parameter. Both the new estimate of the parameter and the new estimate of the uncertainty about the true value are drawn from this posterior distribution.

To illustrate the principles of Bayesian statistics, consider the problem of estimating a univariate variance. Assume that n current observations, X_i, i = 1, ..., n, are available, where X_i, i = 1, ..., n, ~ IID N(0, σ^2). (In this case, the normal distribution serves as the likelihood.) In addition, assume that both a prior estimate of σ^2, denoted s^2, and a prior estimate of the standard deviation of σ^2, denoted e, are available. Define

$$\gamma \equiv \frac{\left(s^2\right)^2}{e^2} + 2 = \frac{s^4}{e^2} + 2$$

$$\delta \equiv s^2 \left(\frac{s^4}{e^2} + 1\right).$$

Based on these definitions, an updated estimate of σ^2 (one incorporating both the data and the prior information) can be constructed from one possible Bayes estimator:

$$s_u^2 = \frac{\delta_u}{\gamma_u - 1} , \tag{11}$$

where

$$\delta_u = \delta + \frac{1}{2}\sum_{i=1}^{n} x_i^2 ,$$

and

$$\gamma_u = \gamma + \frac{n}{2}$$

(in each case, the subscript "u" is used to denote an updated estimate). In addition, an updated estimate of the standard deviation of σ^2 can be constructed:

$$e_u = \frac{\delta_u}{(\gamma_u - 1)\sqrt{\gamma_u - 2}} = \frac{s_u^2}{\sqrt{\gamma_u - 2}} . \tag{12}$$

The derivation of the updated estimators in Equations (11) and (12) are given in the Appendix. That derivation generally follows Lehmann (1983, pp. 246-247). Slightly different approaches to this problem are given by Berger (1985, p. 287, Problem 8), DeGroot (1970), and Searle et al. (1992, pp. 94-96).

One cost of using a Bayesian approach is the requirement to specify a prior distribution. The updating method given above is based on the use of a *conjugate prior*, which is a prior distribution that, when combined with the likelihood, yields a posterior distribution of the same family as the prior. This is quite a handy feature when updating must be done for each of several steps in a process (e.g., for each batch of material to be vitrified), since the posterior distribution (and associated parameter estimates) of the preceding step serves as the prior distribution for the next step, and the estimation procedure is essentially unchanged. Automating such a procedure is quite simple.

Initial estimates of γ and δ must be furnished for the first batch in the process (or waste type). The strength of knowledge about the process will be taken into account in choosing the initial γ and δ to be used by the PCC algorithms.

A multivariate version of the updating scheme discussed in this section appears in Anderson (1984, p. 272). As might be expected, the multivariate approach requires much more prior information. If PCC testing indicates problems, and if adequate prior information can be obtained, the univariate approach (applied to each measured or estimated quantity individually) may be replaced with a multivariate method (applied to all measured and estimated quantities simultaneously). One example of such a multivariate method is the hybrid MEM/Kalman filter mentioned above.

10.0 MISCELLANEOUS TOPICS

This section deals briefly with additional issues related to the nature of the data in the HLW vitrification process (Section 10.1) and to the nature and use of variance and covariance components (Section 10.2).

10.1 COMPOSITIONAL DATA

Batch compositions will be the major type of multivariate data used in HLW process/product control. These compositions are currently planned to be expressed as vectors of 10 mass fractions, corresponding to the nine major oxides and the category "Others." These mass fractions are proportions (or percentages) and therefore must lie between 0 and 1 (or 0% and 100%). In addition, the 10 mass fractions in a single composition should sum to one (or 100%). These characteristics define *compositional data.* Aitchison (1986) discusses the nature of compositional data and statistical techniques for such data. Two consequences of these characteristics of compositional data are of interest here: 1) individual mass fractions cannot be normally distributed, and therefore the joint distribution of the 10 mass fractions cannot be multivariate normal, and 2) covariance matrices for compositions are singular.[a] These facts complicate statistical modelling and manipulation of compositions.

Since compositions cannot be normally distributed, the applicability of statistical techniques that assume normality must be questioned. This issue is discussed by Bryan and Piepel (1994). Briefly, the strict nonnormality of compositional data does not imply that statistical techniques based on the assumption of normality must necessarily perform poorly for compositional data. In fact, preliminary investigations indicate that such techniques

[a] The definition of matrix singularity is somewhat involved; for a full discussion of singularity, see Searle (1982). Briefly, a matrix is singular if there exists some exact linear relationship among the rows or columns of the matrix. The singularity of covariance matrices for compositions follows directly from the unit-sum restriction. The specific manifestation is that each row (and column) of a covariance matrix for compositional data sums to zero. The relevance of the singularity of a covariance matrix for compositional data should become clear in the discussion below.

perform adequately for individual components and quite well for certain functions of compositions. Specifically, property values estimated from CVS models seem to follow distributions that are almost indistinguishable from normal distributions. More insight into performance of techniques based on normality will be gleaned during testing of the PCC algorithms.

The singularity of the covariance matrix for compositional data and the nonnormality of such data complicate the generation of random compositions (which will be required for testing PCC algorithms). The technique used to generate random compositions for the Monte Carlo studies reported in this document is based on the technique for generating observations from a multivariate normal distribution presented by Kennedy and Gentle (1980, pp. 228-231). This technique uses the Cholesky decomposition of the covariance matrix of the target multivariate normal distribution to transform a vector of IID normal random variables. The singularity of the covariance matrix for compositional data necessitates a minor modification of this technique -- one row and column of the covariance matrix (usually the row and column corresponding to the "Others" component) is dropped, in order to eliminate the singularity. The Cholesky decomposition of this reduced covariance matrix is used to produce nine components of the composition, and the tenth component is calculated by subtracting the sum of the nine components from one.

10.2 VARIANCE AND COVARIANCE COMPONENTS

The remaining issues to be addressed in this section relate to the nature and use of variance and covariance components in HLW vitrification process/product control:

- Once estimates of variance components have been obtained, these components can be used to optimize allocation of sampling effort. Essentially, optimal allocation entails concentrating sampling effort at those stages in the hierarchy of uncertainty (batch-to-batch variability, within-batch uncertainty, analytical uncertainty) that contribute most to the overall uncertainty. Cochran (1977) discusses optimal allocation of sampling effort.

- The methods presented in this document assume a certain stability in the hierarchy of uncertainty. For example, within-batch and analytical uncertainties are assumed to be constant or changing only slowly. This assumption seems reasonable and greatly reduces what would otherwise be an inachievable level of sampling for each batch and analysis for each sample. Still, this assumption should be investigated during HLW vitrification operations. Such investigation would be a natural part of a process monitoring scheme.

- The linear models used for attributes of feed, melt, and glass [Equations (7) and (9)] could be modified to include other variance (covariance) components, e.g., separation of within-batch heterogeneity from sampling error, or separation of variability induced during sample preparation from analytical error (Bryan and Piepel, 1994).

11.0 APPLICATIONS AND FURTHER WORK

The topics covered in this document affect several stages of HLW vitrification process/product control. The specific applications of this material will depend on the course of development of the process/product control system and the vitrification plant itself. Some examples of applications are presented in Section 11.1. A number of possible future investigations are briefly discussed in Section 11.2.

11.1 POSSIBLE APPLICATIONS

Uncertainty estimates must be available when processing begins. If no data are available from which to estimate various uncertainties (or if the available data are inadequate), the methods of Section 5 should be used to construct uncertainty estimates. If data are available (e.g., from Savannah River or West Valley operations, or from development and testing of the Hanford HLW vitrification plant itself), the methods of Section 3 and 4 should be used to estimate uncertainties.

During processing of each batch, composition data will be obtained and used to judge batch quality. As part of this, uncertainty estimates will be calculated, updated, combined, and propagated; these steps draw on the material in Sections 3 and 4, 9, 6, and 7, respectively.

Testing of the PCC algorithms (e.g., with the Plant Simulation Code) will probably require the ability to produce random vectors ("observations") that follow known and reasonable covariance patterns. A method for generating multivariate observations that follow a given covariance structure is presented in Section 10.1. Two possible analytical covariance matrices are discussed in this document: the hypothetical covariance matrix developed in Section 5.1, and the Corning RR6 covariance matrix (discussed in Section 7.2). In addition, the methods of Section 5 can be adapted to produce a range of covariance matrices with which to exercise the PCC algorithms.

The results presented in Section 8 suggest that reliable uncertainty estimates should be based on 20 or more observations. If a designed experiment, such as the one described in Section 3.1, is to be conducted to estimate variance (or covariance) components, a minimum of 20 batches should be examined, with two samples drawn from each batch, and two analyses run for each sample. Such an experiment would entail 20 x 2 x 2 = 80 observations and would provide 19, 20, and 40 degrees of freedom for estimating batch-to-batch variability, within-batch uncertainty, and analytical uncertainty, respectively. Unfortunately, since uncertainty estimates are required for each of the tanks and steps in the HLW vitrification process, several to many such experiments might be required. However, the replicated analytical effort might be unnecessary for later experiments, providing analytical uncertainty can be assumed to be unaffected by the source of the analyzed material.

11.2 POSSIBLE FUTURE INVESTIGATIONS

A number of possible future investigations were mentioned in preceding sections. For ease of reference and comparison, these are recapitulated below, along with the section in which the topic arose:

- Identify and collect information from Savannah River's DWPF and/or the West Valley Demonstration Project that is suitable for estimating composition uncertainty, and perform this estimation (Section 1).

- Evaluate the assumption of perfect mixing, alternatives, and implications for choosing between mean-based statistical procedures and percentile-based statistical procedures (Section 3.1).

- Investigate necessity for and applicability of "shrunken" estimators of attribute values (Section 3.1).

- Develop and compare techniques for normalizing measured compositions, including effects on uncertainty (Section 5.1).

- Improve simulation of analytical uncertainty (Section 5.1).

- Develop simulated within-batch and batch-to-batch covariance matrices (Section 5.1).

- Examine sensitivity of simulated covariance matrices to "truè" composition, precision of various steps in the measurement process, and the measurement of several components from the same or different aliquots, solutions, or subsamples (Section 5.1). Note that such sensitivity analyses may be valuable in optimizing allocation of sampling effort.

- Extend the method for constructing measures of strength of belief for simulation-based estimates of univariate uncertainty to the multivariate situation (Section 5.2).

- Develop methods for constructing measures of strength of belief in simulation-based uncertainty estimates that incorporate a confidence coefficient (Section 5.2).

- Develop the multivariate Bayes and hybrid Kalman filter/Measurement Error Model approaches to updating uncertainty estimates; compare each to the existing MEM and the univariate Bayesian updating method (Section 9).

- Investigate natural and reasonable statistical models for compositional data, including the Dirichlet and logistic normal classes of distributions discussed by Aitchison (1986) (Section 10.1).

- Develop optimal allocation of sampling effort in HLW vitrification operations (Section 10.2).

- Examine adequacy of the suggested experimental design and sample sizes for estimating uncertainty, under several sets of assumptions about the true uncertainties (Section 11.1).

The necessity and benefits of each activity should be judged relative to operating experience and data (such as that from Savannah River and West Valley) and testing of the existing PCC algorithms (e.g., with the Plant Simulation Code).

12.0 REFERENCES

Adams, T.L. 1994. Application of the HWVP Measurement Error Model and Feed Test Algorithms to Pilot Scale Feed Testing, PHTD-C93-05.01B, Rev. 0, Pacific Northwest Laboratory, Richland, Washington.

Aitchison, J. 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, New York.

Anderson, D.N., and G.F. Piepel. 1993. Preliminary Investigation of Glass Composition Covariance Matrices and Glass Property Models, PHTD-C92-05.01D, Rev. 0, Pacific Northwest Laboratory, Richland, Washington.

Berger, J.O. 1985. Statistical Decision Theory and Bayesian Analysis, second edition. Springer-Verlag, New York.

Bryan, M.F. and G.F. Piepel. 1993. Strategy for Product Composition Control in the Hanford Waste Vitrification Plant, PHTD-C93-05.01F, Rev. 0, Pacific Northwest Laboratory, Richland, Washington.

Bryan, M.F. and G.F. Piepel. 1994. Preliminary Feed Test Algorithm for the Hanford Waste Vitrification Plant Product Composition Control System, PHTD-C93-05.01A, Rev. 0, Pacific Northwest Laboratory, Richland, Washington.

Bryan, M.F., G.F. Piepel, and D.B. Simpson. 1994. Demonstrating Compliance with WAPS 1.3 in the Hanford Waste Vitrification Plant Process, PHTD-C93-05.01K, Pacific Northwest Laboratory, Richland, Washington. To be issued.

Cochran, W.G. 1977. Sampling Techniques, third edition. John Wiley and Sons, New York.

DeGroot, M.H. 1970. Optimal Statistical Decisions. McGraw-Hill Book Company, New York.

Deming, W.E. 1943. Statistical Adjustment of Data. John Wiley and Sons, Inc., New York.

Gaylor, D.W., and F.N. Hopper. 1969. "Estimating the Degrees of Freedom for Linear Combinations of Mean Squares by Satterthwaite's Formula," Technometrics, 11:691-706.

Graybill, F.A. 1976. Theory and Application of the Linear Model. Duxbury Press, North Scituate, Rhode Island.

Hahn, G.J., and W.Q. Meeker. 1991. Statistical Intervals: A Guide for Practitioners. John Wiley and Sons, Inc., New York.

Hrma, P.R., G.F. Piepel, M.J. Schweiger, D.E. Smith, P.E. Redgate, J.W. Johnston, and D.J. Bates. 1992. Property/Composition Relationships for Hanford Waste Vitrification Plant Glasses--Preliminary Results Through CVS-II Phase 2, PHTD-92-03.01/K897, Pacific Northwest Laboratory, Richland, Washington.

Hrma,.P.R., G.F. Piepel, M.J. Schweiger, D.E. Smith, D.-S. Kim, P.E. Redgate, J.D. Vienna, C.A. LoPresti, D.B. Simpson, and D.K. Peeler. 1994. Property/Composition Relationships for Hanford Waste Vitrification Plant Glasses--Results Through CVS-II Phase 4, PVTD-93.01C, Pacific Northwest Laboratory, Richland, Washington, draft, May 1994.

Kennedy, W.J., Jr., and J.E. Gentle. 1980. Statistical Computing. Marcel Dekker, Inc., New York.

Lehmann, E.L. 1983. Theory of Point Estimation. John Wiley and Sons, New York.

Lindgren, B.W. 1976. Statistical Theory, third edition. MacMillan Publishing Co., Inc., New York.

Mandel, J. 1964. The Statistical Analysis of Experimental Data. John Wiley and Sons, Inc., New York.

Postles, R.L. and K.G. Brown. 1991. Savannah River Site Defense Waste Processing Facility Product Composition Control System Statistical Process Control Algorithms, SCS-PMC-91097, Westinghouse Savannah River Company, Aiken, SC.

Randklev, E.R. 1993. Hanford Waste Vitrification Plant Project Waste Form Qualification Program Plan, June 1993, WHC-EP-0522, Westinghouse Hanford Company, Richland, Washington.

Satterthwaite, F.E. 1946. "An Approximate Distribution of Estimates of Variance Components," Biometrics Bulletin, 2:110-114.

Searle, S.R. 1971. Linear Models. John Wiley and Sons, New York.

Searle, S.R. 1982. Matrix Algebra Useful for Statistics. John Wiley and Sons, New York.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance Components. John Wiley and Sons, Inc., New York.

Snedecor, G.W., and W.G. Cochran. 1980. Statistical Methods, seventh edition. Iowa State University Press, Ames, Iowa.

U.S. Department of Energy. 1993. Waste Acceptance Product Specifications for Vitrified High-Level Waste Forms, EM-WAPS, Rev. 0, February 1993, Office of Environmental Restoration and Waste Management, U.S. Department of Energy, Germantown, Maryland.

WSRC. 1993. Defense Waste Processing Facility Waste Form Qualification Report, Volume 5: Technical Bases for the DWPF Glass Product Control Program, WSRC-IM-91-116-5, Westinghouse Savannah River Company, Aiken, SC.

APPENDIX

DERIVATION OF A BAYESIAN ESTIMATOR OF VARIANCE

APPENDIX

DERIVATION OF A BAYESIAN ESTIMATOR OF VARIANCE

The problem considered here is that of estimating (or updating an existing estimate of) a univariate variance, σ^2, using Bayesian statistics. The form of the Bayes estimator is affected by

- the prior distribution, a statistical distribution that embodies information on the parameter of interest, in this case, the variance;
- the likelihood, the statistical distribution from which the data are assumed to be drawn; and
- the loss function, which characterizes the penalty associated with incorrect estimates of the parameter.

The squared error loss function is used in this derivation; for squared error loss, the Bayes estimator is simply the mean of the posterior distribution. Further discussion of loss functions is beyond the scope of this document; see Berger (1985), DeGroot (1970), Lehmann (1983), or Lindgren (1976) for more information.

A normal likelihood is used in this development. Specifically, it is assumed that n observations, X_i, i = 1, ..., n, are available, where X_i, i = 1, ..., n, ~ IID $N(0, \sigma^2)$.

In addition, it is assumed.that both a prior estimate of σ^2, denoted s^2, and a prior estimate of the standard deviation of σ^2, denoted e, are available. These estimates are used to formulate a prior distribution below.

Define:

$$\gamma \equiv \frac{(s^2)^2}{e^2} + 2 = \frac{s^4}{e^2} + 2$$

$$\delta \equiv s^2 \left(\frac{s^4}{e^2} + 1 \right).$$

Based on these definitions, an updated estimate of σ^2 (one incorporating both the data and the prior information) can be constructed from one possible Bayes estimator:[a]

$$s_u^2 = \frac{\delta_u}{\gamma_u - 1}, \tag{A.1}$$

where

$$\delta_u = \delta + \frac{1}{2}\sum_{i=1}^{n} x_i^2 ,$$

and

$$\gamma_u = \gamma + \frac{n}{2}$$

(in each case, the subscript "u" is used to denote an updated estimate). In addition, an updated estimate of the standard deviation of σ^2 can be constructed:

$$e_u = \frac{\delta_u}{(\gamma_u - 1)\sqrt{\gamma_u - 2}} = \frac{s_u^2}{\sqrt{\gamma_u - 2}} . \tag{A.2}$$

The development of a prior distribution from s^2 and e and the use of this prior distribution in deriving the estimators in Equations (A.1) and (A.2) are now considered. Under normality, one (slightly nonstandard) representation of the joint density of the X's is

$$f(\underline{x} \mid \tau) \equiv f(x_1, \ldots, x_n \mid \tau) = C_1 \tau^r e^{-\tau \sum_{i=1}^{n} x_i^2} = C_1 \tau^r e^{-\tau y} ,$$

[a] Since the prior distribution has not yet been fully specified, many Bayes estimators are possible.

where

$$y \equiv \sum_{i=1}^{n} x_i^2 \,, \quad \tau \equiv \frac{1}{2\sigma^2} \,, \quad r \equiv \frac{n}{2} \,,$$

C_1 is a normalizing constant (to ensure that the density function integrates to one; this constant is of no interest in this discussion), and $\underline{x}$ denotes the vector containing x_i, i = 1, ..., n. In this development, a prior distribution is placed on τ,[(a)] rather than on σ^2. Specifically, τ is assumed to follow the gamma density $\Gamma(g,1/\alpha)$:

$$\pi(\tau) = C_2 \tau^{g-1} e^{-\alpha\tau} \,, \tag{A.3}$$

where C_2 is another normalizing constant (again, of no interest in this discussion).

Reasonable values for the parameters of this gamma density (g and α) must now be derived. To do so, the following results for the $\Gamma(g,1/\alpha)$ distribution are used:

$$E(\tau) = \frac{g}{\alpha}, \qquad E(\tau^2) = \frac{g(g+1)}{\alpha^2} \,,$$

$$E\left(\frac{1}{\tau}\right) = E\left(2\sigma^2\right) = \frac{\alpha}{g-1}, \qquad E\left(\frac{1}{\tau^2}\right) = E\left(4\sigma^4\right) = \frac{\alpha^2}{(g-1)(g-2)} \,.$$

From these results, the following can be derived:

$$E(\sigma^2) = \frac{\alpha}{2(g-1)} \,,$$

$$E(\sigma^4) = \frac{\alpha^2}{4(g-1)(g-2)} \,,$$

$$V(\sigma^2) \equiv E\left(\sigma^2 - E(\sigma^2)\right)^2 = E(\sigma^4) - \left[E(\sigma^2)\right]^2$$

[(a)] The parameter τ, known as the *precision*, is sometimes used in place of σ^2, the variance, as a parameter of the normal distribution.

$$= \frac{\alpha^2}{4(g-1)(g-2)} - \left[\frac{\alpha}{2(g-1)}\right]^2 = \frac{\alpha^2}{4(g-1)^2(g-2)} \; .$$

The prior estimates are now equated to the corresponding moments of the proposed gamma prior [i.e., $s^2 = E(\sigma^2)$; $e^2 = V(\sigma^2)$], the resulting equations are solved for g and α, and the relationships to γ and δ (defined above) are noted, as follows:

$$g = \frac{(s^2)^2}{e^2} + 2 = \frac{s^4}{e^2} + 2 = \gamma \; ,$$

$$\alpha = 2s^2(g-1) = 2s^2\left(\frac{s^4}{e^2}+1\right) = 2\delta \; .$$

The elements necessary for a Bayesian approach to updating the estimate of σ^2 are now available. The updating proceeds by forming the posterior distribution of τ as the (properly normalized) product of the prior distribution and the likelihood:

$$\pi(\tau \mid \underline{x}) = C_3 f(\underline{x} \mid \tau)\pi(\tau) = C_4 \tau^{r+g-1} e^{-\tau(\alpha+y)} = C_{\tau}^{r+\gamma-1} e^{-\tau(2\delta+y)} \; ,$$

where C_3 and C_4 are normalizing constants (again, of no interest in this discussion). By comparison with Equation (A.3) above, the posterior distribution of τ is $\Gamma[r+\gamma,\ 1/(2\delta+y)]$. Thus, the posterior distribution is of the same family as the prior distribution (the gamma family), and the update is reflected by the change in the parameters of the gamma distribution. The Bayes estimator of $\sigma^2 = 1/2\tau$ is the posterior mean of $1/2\tau$:

$$s^2 = E\left(\frac{1}{2\tau} \mid \underline{x}\right) = \frac{2\delta+y}{2(r+\gamma-1)} = \frac{\delta+\frac{1}{2}\sum_{i=1}^{n} x_i^2}{\gamma+\frac{n}{2}-1} = \frac{\delta_u}{\gamma_u-1} \; ,$$

as in Equation (A.1) above. Similarly, the posterior standard deviation of 1/2t is used as the estimator of the posterior standard deviation of σ^2:

$$e_u = \sqrt{\frac{(2\delta+y)^2}{4(r+\gamma-1)^2(r+\gamma-2)}} = \frac{s^2}{\sqrt{\gamma+\frac{n}{2}-2}} = \frac{\delta_u}{(\gamma_u-1)\sqrt{\gamma_u-2}} \, ,$$

as in Equation (A.2) above.