Statistical regression modelling of glass properties – a tutorial

Alexander Fluegel

Pacific Northwest National Laboratory, Richland, WA 99352, USA

Manuscript received 5 August 2005
Revised version received 14 July 2006
Accepted 28 July 2006

Known statistical analysis methods are described in detail with the aim of developing a new and more accurate modelling approach for glass properties. It is shown that the combined analysis of historic and current data from, for example, the SciGlass and Interglad databases, often provides the basis for making property predictions that are more reliable than the raw data itself or a few test measurements. Targeted glass research and process modelling can be facilitated by the approach outlined.

Introduction

In glass research and technology, it is often necessary to reduce the costs of raw materials, to improve specific properties, or to design a glass composition for new applications. To meet these needs, the use of personal experience and published scientific literature is advantageous. In recent years, the creation of large glass property databases has facilitated systematic glass property modelling and property measurement evaluation. It is no longer required to search worldwide for numerous individual publications regarding specific properties. The commercially available systems SciGlass(1) and Interglad(2) combine hundreds of thousands of experimental findings from the majority of glass research papers from over a century, including the associated references. In addition, SciGlass gives details about measurement methods. Furthermore, predictions are possible in SciGlass through models published previously, and Interglad includes a predictive linear regression feature.

Despite the recent progress, there are several shortcomings in these commercial tools:

(1) Numerous experimental data from various investigators differ significantly, even within simple glass systems and for well investigated properties (see Figure 8 in the discussion of future work).

(2) Many predictive glass property models exist in the literature, and it is sometimes difficult to decide which model is the most appropriate.

(3) Industrial glasses have complex compositions, and it is not always possible to predict properties through commonly used linear property–composition relations.

(4) Some of the published models are based on scientific principles or derived assumptions about the details of chemical bonding within a glass. Therefore, experiments need to be interpreted (that lead to those scientific principles or to those derived assumptions about the details of chemical bonding within a glass) *before* some models can be established. This can be a source of error based on the accuracy of this interpretation.

(5) Some models consider the experimental findings of a few or only one single investigator, i.e. systematic errors of a few investigators easily can lead to incorrect conclusions. Systematic errors of whole data series are known in glass science as described below in Table 2, and referred to at the end of Step 7 of the statistical analysis (outlier analysis, data leverage).

(6) The prediction errors for models based on one investigator reflect the measurement precision in her/his laboratory, but it is not possible to conclude how such prediction errors will relate to those based on data from other laboratories.

In this work, an attempt is made to overcome these problems above for practical application in glass technology. The goal is a method that features quantitative prediction accuracy and simplicity. Therefore, it is not a requirement to have an expert knowledge of all details of the nature of glass, the published modelling techniques, or advanced mathematical procedures. It is also not necessary to acquire expensive software (besides common spreadsheet software like Excel) or a high performance computer. However, it is strongly advised to rely on experience about the subject matter, and, if possible, on dedicated software that can perform some of the procedures described in this tutorial automatically.

Now at xxxxxxxx. Email XXXXXXXXXX@xxxx

High accuracy can be obtained by using (1) a large number of experimental findings from the commercial SciGlass and Interglad databases and (2) well established statistical analysis techniques. The statistical approach outlined here is not original in general, as numerous publications can be consulted about the topic,(3–9) written over two centuries. In the current work, the statistical method largely follows general techniques that are applied successfully, for example by Harold S Haller & Company(10) in business consulting, by Bechtel Hanford Inc. for nuclear waste vitrification,(11) and for modelling of solar control glasses.(12) Statistical analysis is also firmly established in numerous areas in quality control, economy, biology, sociology, politics, etc., to mention just a few applications. The range of applications underlines the basic character of statistical analysis that has been ignored for too long in glass science.

Statistical analysis in glass science and technology

Ernst Abbe, Otto Schott, and A. Winkelmann initiated methodical studies of glass properties in the 19th century, creating the basis of modern glass science in Jena, Germany. They considerably improved the use of glasses for special applications, e.g. for optics, thermometers, and as a thermal shock resistant material. Winkelmann & Schott published the first model that allowed the prediction of glass properties based on the chemical composition, using the additivity principle,(13,14) i.e. multiple regression using linear functions. This principle is based on the assumption that the relation between the glass composition and a specific property is linearly related for all component concentrations. In this case all of the component influences can be summed as follows

$$\text{Property} = \beta_0 + \sum_{i=1}^{n} \beta_\mathrm{i} C_\mathrm{i} \qquad (1)$$

where β_0 in Equation (1) is the model intercept, n is the total number of significant glass components excluding the main component (usually silica), the i values are the individual numbers of the significant glass components, the β_i values are the component-specific coefficients, and the C_i values are the concentrations of the glass components (also called model factors or independent variables).

The additivity principle allows for very precise and accurate predictions within limited concentration ranges,(12,13,15–24) that cannot be reached by structural,(25–29) thermodynamic,(30–53) or molecular dynamic(54–56) modelling approaches. Because of the simplicity of the technique, the ease of interpretation, and good prediction results within specified limits, Equation (1) is most widely used for glass property modelling. The additivity principle cannot however be applied for modelling glass properties over wide concentration ranges because of component interactions.

The glass models summarised above are not different in principle. Procedures expressed through equations based on specific variables are always established for property–composition relationships. The variables and equations vary, however, and sometimes the variables are derived beforehand from basic principles (“*ab initio*”) which are in turn based on other (basic) observations such as the atomic mass, bonding distance, bond strength, electron affinity, stochiometry, or elemental charge. Therefore, all models are empirical by nature, while in *ab initio* models the empirical character is hidden within fundamental laws about property relations. Often, property modelling derived from other observed properties is less reliable (but scientifically more interesting) than direct modelling of observed properties. Because the nature of glass is not sufficiently well understood, few models exist where different kinds of properties can be mutually derived from each other, e.g. thermodynamic and rheological properties, even though rheological properties can be determined from thermodynamic properties. The final goal is to *understand* the meaning of all the variables and relationships in glass models in detail. In the current work, it is suggested that the problem should be approached stepwise, i.e. an empirical start is made using simple linear and polynomial relations between the glass composition and observed properties. Once all relevant phenomena have been recognised and organised by means of basic statistical analysis, the step from observation to interpretation can be performed with much higher confidence than without statistical analysis.

In the following paragraphs, simple statistical analysis will be developed for glass property modelling as the most basic tool of data organisation and interpretation.

Single linear regression using linear functions

$$\text{Property}=\beta_0+\beta_1 C \qquad (2)$$

The terms β_0 and β_1 in Equation (2) are *regression coefficients* and C is the *regression factor* or *independent variable* (glass component concentration). Equation (2) can be used for glass property modelling in binary systems within narrow concentration limits.

Single linear regression using polynomial functions

$$\text{Property}=\beta_0+\beta_1 C+\beta_2 C^2+\beta_3 C^3+\ldots \qquad (3)$$

Equation (3) can be used for modelling in binary glass systems within wide concentration ranges as long as sufficient data exist and sharp property extrema(57,58) caused by significant crystallisation, phase separation, or other effects do not occur.

Even though *nonlinear* polynomial functions are used in Equations (3), (6), and (7), those equations are still *linear* in the coefficients and they are solved using *linear* regression. All coefficients of the higher order terms in Equations (3), (6), and (7) are determined in exactly the same way (through Equation (12), see below) as the first order additive terms in Equations (1), (2), (4), and (5). An example for non linear regression is described below.

Multiple linear regression using linear functions

$$\text{Property}=\beta_0+\beta_1C_1+\beta_2C_2+\beta_3C_3+\ldots \qquad (4)$$

Equation (4) is identical to Equation (1). It may be applied for glass property modelling in multicomponent systems within narrow concentration ranges under exclusion of sharp property extrema such as caused by crystallisation, phase separation, and other effects. Component interactions are not considered in this model form.

It needs to be mentioned that the intercept in Equations (1) to (4) may be eliminated if the main glass component silica (SiO_2) is included in the equation. For example, in the binary system SiO_2–Na_2O Equation (2) may be modified as follows

$$\text{Property}=\beta_0+\beta_1C(\text{Na}_2\text{O})=\beta_2C(\text{SiO}_2)+\beta_3C(\text{Na}_2\text{O})+\ldots \qquad (5)$$

[Author: please check this equation, not sure about the second "="]

The approach using an intercept is called *"slack variable"* (SV) technique and the approach without results in a *"canonical"* or *"Scheffé"* type model.[59] Both approaches are very similar statistically; however, the canonical form of the model is expected to produce slightly more accurate predictions if the main component concentration varies significantly or even approaches zero. The slack variable technique will be the main focus of this study because it can be straightforwardly applied to common binary systems such as SiO_2–Na_2O using squared and cubic terms for Na_2O that are not allowed in the canonical technique, as discussed in the following section. Most commercial glasses contain silica as the main component (about 40 to 85 mol%) that can be excluded, and slack variable modelling results can be interpreted easily: a specific coefficient shows the property variation caused by the exchange of 1% of the considered component for silica. The multiple regression coefficients obtained with the slack variable (SV) technique are similar to the values commonly depicted in "spider graphs".[16,17] Except for the constant term, β_0, the SV coefficients must be interpreted as resulting from interactions with the excluded main component silica (Ref. 6, p15–18 and 333–343); while in a canonical model the coefficients show the extrapolated properties of the related pure components in the (theoretical) vitreous state. All equations described in this work can be applied for the SV and canonical modelling approaches as described below.

Multiple linear regression using polynomial functions

$$\text{Property}=\beta_0+\beta_1C_1+\beta_2C_2+\beta_3C_3+\ldots+\beta_4C_1{}^2+\beta_5C_2{}^2+\beta_6C_3{}^2+\ldots$$
$$+\beta_7C_1C_2+\beta_8C_2C_3+\ldots+\beta_9C_1{}^3+\ldots\beta_{10}C_1{}^2C_2+\beta_{11}C_1{}^2C_3+\ldots \qquad (6)$$

Equation (6) is analogous to Equations (2) to (5). The exact form of Equation (6) for the SV regression model using polynomial functions of the second order

$$\text{Property} = \beta_0 + \sum_{i=1}^{n}\left(\beta_\text{i}C_\text{i} + \beta_\text{i}C_\text{i}^2 + \sum_{k=i+1}^{n}\beta_\text{ik}C_\text{i}C_\text{k}\right) \qquad (7)$$

where β_0 is the intercept, β_i are the single component coefficients and the coefficients of squared influences, and β_ik are the coefficients of two component interactions. The variable *n* in Equation (7) is the total number of the significant glass components excluding silica; *j* and *k* are the indices of the significant glass components, and the *C*-values are the component concentrations (excluding silica) in mol or mass fraction or percent (preferably mol fraction or percent). Equation (7) or its canonical variation may be used for glass property modelling in multicomponent systems over wide concentration ranges.[11,60–66] Sudden property changes, such as changes caused by crystallisation and/or phase separation, are difficult to describe with multiple regression using polynomial functions;[58] therefore, advanced nonlinear functions should be applied in these cases. For example, only within very limited concentration ranges can glass liquidus temperatures be modelled with linear regression using polynomial functions,[22,23,67–69] as long as the primary crystal phase is constant or has little influence.

If *all* high order terms are examined in a second order canonical model, then only cross product terms ($C_\text{i}C_\text{k}$), and no squared terms ($C_\text{i}{}^2$), are allowed to avoid over-parameterisation caused by the mixture constraint $C_1+C_2+C_3+\ldots+C_\text{n}=100\%$. However, if only some of the second order terms are populated, then both cross product and squared terms can appear in a "reduced" model.[70]

Multiple nonlinear regression using advanced functions

It is possible to introduce advanced functions into Equation (7), or the structure of the equation may be changed completely. *Chemical equilibria* between species in glass may be considered.[36,43,44,71] The exact calculation of chemical equilibria based on equilibrium constants and total concentrations requires the solution of high order polynomials, e.g. a second order polynomial for a two component mixture including one interaction,[64] or a fifth order polynomial for a three component mixture including all three possible two component interactions. Multicomponent mixtures can be calculated numerically using a high performance workstation.

A widely applied nonlinear approach is *neural network regression*:[72] logistic activation functions such as $e^x/(1+e^x)$ or $\tanh(x)$ are used as "neurons" that make graduated yes/no "decisions" according to the input signal. Several neurons are connected following specified rules, partially comparable to neural networks in biology. The connections between the neurons are weighted, i.e. the signal is amplified or reduced. The "weights" are the fitting coefficients that are optimised using neural network regression. Dreyfus *et al*[58] demonstrate the application of neural network regression for glass property modelling. They show that neural network regression is advantageous compared to multiple regression using polynomial functions if the glass property does not change gradually with the chemical composition, but sharp property extrema do occur such as liquidus surfaces. Polynomial functions do not fit the sharp extrema well that may appear in glasses because of crystallisation or phase separation.

For building a neural network, component interactions have to be assumed and the total number of adjustable weights in neural networks regression needs to be always larger than two times the number of significant glass components. Therefore, neural network regression requires a very high number of data points within an experimental region where interactions can be investigated. Typically multiple training sets are used to "teach" the neural network prior to application to prediction. Each training set will require at least as many data points as adjustable weights used by the network to predict response.

Advanced procedures for obtaining optimal regression fits

During regression analysis, often the "*ordinary least squares*" (OLS) technique is applied to derive coefficients for each factor, this technique minimises the sum of squares of the differences between the observed and calculated glass properties (*unexplained errors, residuals*). It is assumed that all unexplained errors are normally distributed with a mean of zero and with no mutual correlation of errors. Significant outliers that influence the result should not exist. If those conditions do not apply, then advanced procedures for obtaining optimal regression fits can be used, e.g. *robust regression*.[73,74]

For glass property modelling, it is commonly not required to evaluate further fitting methods besides OLS. In addition, the leverage analysis described below in the section "Step 7: Outlier analysis, data leverage" allows the detection and handling of outliers that influence the result.

Limits of regression analysis for glass property prediction

In principle, regression analysis can be applied to glass property data so long as a systematic relation exists between the experimental conditions such as concentrations and the resulting properties. However, even though regression analysis can be used in almost all cases, it may be used incorrectly. The most important issue is the possibility of sharp extrema in glass properties. Besides crystallisation and phase separation effects, sharp extrema may occur in glasses with a network former content higher than 85 to 90 mol%. For example, the Littleton softening point of 100% pure silica glass may be estimated as 1666±50°C from 54 datapoints in SciGlass.[1] If as little as 0·06 mol% sodium oxide is introduced, then the Littleton softening point decreases dramatically to 1280°C according to Leko.[57] In addition to high silica glasses, sharp property extrema may also be expected for glasses with high concentrations of B_2O_3 (based on modelling studies of the author; see also publications by Appen and Gan Fuxi), P_2O_5, and GeO_2 or if extreme compositions have to be quenched fast to prevent crystallisation (e.g. 50 mol% Na_2O plus 50 mol% SiO_2). Sharp property extrema also appear to exist in alkali aluminosilicate glasses with high Al_2O_3 concentrations,[75,76] especially at low temperatures if the molar ratio Al/Na is approximately 1 to 1·2.

If sharp property extrema occur that cannot be described through the inverse of concentrations, then advanced regression techniques must be applied. Equation (7) must be modified substantially, or the model application range must be narrowed.[22,23,67–69]

Many property data in glass related scientific publications were not obtained using optimal experimental designs, and most papers were not published in mutual cooperation. This fact leads to sometimes strong correlations among the model variables (see below, section "Step 4: Correlation analysis"), i.e. variables partly depend on each other. Strong correlations can be excluded through the procedure described below, but even weaker correlations always need to be considered during interpretation of the model results (see below).

For models based on simple assumptions, advanced interpretations are hardly possible. For example, it is completely out of the question to derive atomic radii, chemical equilibrium constants, or properties of pure silica from polynomial models based on a few soda–lime–silica glasses. Model application limits must be followed as described below in the section "Step 9: Model application". Sometimes, interesting interpretations can be derived from multiple regression models, for example about the mixed alkali effect and the influence of batch materials on the glass viscosity.[77] Advanced regression models may allow further interpretation.

Selection of the appropriate regression technique

In general, regression analysis should be used according to the available data. Within narrow concentration limits, the linear additivity approach is most appropriate whereas wider concentration limits require polynomial functions. If property extrema appear (e.g. through crystallisation, phase separation, or in glasses with a network former content higher than 90 mol%) that cannot be described via the inverse of concentrations, then advanced regression techniques should be applied. The transition between the regression approaches is gradual. Most important for deciding which technique should be used is the study of the linear correlation matrix (see below, Step 4: Correlation analysis) of the available concentration data as well as the analysis of binary, ternary, and other related systems for evaluating the crystallisation and phase separation tendency, the property extrema, and general property trends.[78] In addition, all the modelling techniques not based on regression analysis that are summarised in the introduction should be evaluated according to scientific experience.

Regarding commercial silicate and borosilicate glasses, it is recommended in this paper, for practical reasons, to initially apply the additive multiple regression technique according to Equation (1) or to introduce selected higher order terms according to Equation (7). It will be explained below, beginning with the section "Step 3: Selection of the model type; establishment of all possible variables", how the higher order terms are selected.

Analysis procedure using common spreadsheet software

Step 1: Selection of the source data

The glass property of interest must be defined. According to general previous experience a glass composition region should be selected. The compositional region does not need to be very specific at first. It is helpful to extract all available values from the databases SciGlass[1] and Interglad,[2] and other data sources. If the desired property was never or only seldom investigated in the compositional area of interest, then new experiments should be designed and performed.

The uniformity of the selected source data (the frequency of outstanding or special glass compositions within the source data) can be evaluated with the leverage analysis described below (h-value). In many cases, outstanding glass compositions can be recognised empirically by sight without further calculations. If possible, outstanding glass compositions should be excluded from the model, or it should be verified that the reported properties are correct through repeated experiments.

All concentration values must be converted to either mol%, wt%, or concentration ratios, where mol% may be preferred, especially over relatively wide concentration ranges. Models based on mol% can be interpreted more easily, for example regarding the mixed alkali effect.[77]

Step 2: Calculation of property fixpoints and error normalisation

For obtaining optimal results, it is beneficial to stay as close to the original data as possible. For example, it is better to fit original viscosity data than the derived constants of the Vogel–Fulcher–Tammann equation, as demonstrated by Fluegel and co-workers.[20,77] Furthermore, properties should be preferred that require as limited a number of measurements as possible to reduce error propagation. For instance, glass corrosion rates may be expressed in mm/day, which requires measurement of only the corrosion layer thickness and time. Glass corrosion rates may also be expressed in g/m^{2}day, where in addition to the corrosion layer thickness and time, the glass density also has to be determined. Three measurements are likely to contribute more error to the desired property than two measurements.

A property fixpoint must be selected for all glasses that need to be analysed. If possible, not only the directly given data may be considered, but also safe interpolations should be taken into account; e.g. if a property is known at 800°C, 900°C, and 1100°C, it is in many cases possible to estimate for 1000°C.

For selecting an appropriate scale for the property fixpoint, it is required initially to estimate the overall error distribution. For example, it can be assumed that all density measurements expressed in g/cm^3 have a similar error, independent from the absolute value; i.e. a glass density of 2·5 g/cm^3 can be measured with approximately the same precision as a glass density of 5 g/cm^3. This is not the case for all properties. For example, a glass viscosity measurement at 100 Poise is related to a much lower absolute error than a viscosity measurement at 10^{13} Poise. However, the relative errors of all viscosity measurements are closer to a constant, i.e. a constant percentage of the absolute value. Therefore, viscosities are commonly expressed on a logarithmic scale, also called *data transformation*, which normalises relative errors to absolute errors. In Figure 1 the influence of logarithmic transformation on the observed property is demonstrated.

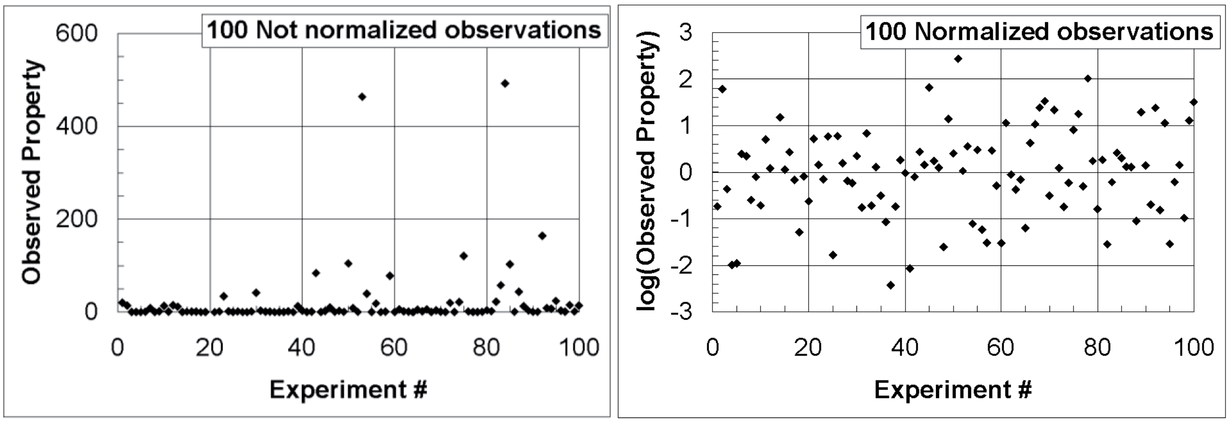

Figure 1. Non transformed and transformed data

In glass technology, it is recommended either to model the desired glass properties directly without data transformation or to determine the logarithm (natural or decadic) of the property.

Figure 2 illustrates the systematic approach concerning instances when logarithmic transformation for error normalisation should be considered; after modelling a plot of the observed property values versus the differences of observed and calculated property values (residuals) shows increasing residual variance with increasing property value. In other words if large observed property values appear apparently as outliers as seen in Figure 2 logarithmic transformation for error normalisation is advised.

Besides the logarithmic approach, further error normalisation techniques are not common for glass property modelling. In a few cases reciprocal normalisation is possible.

Step 3: Selection of the model type; establishment of all possible variables

It is proposed to select the model type from Equation (1) for narrow concentration ranges if only few experimental data are available, or Equation (7) otherwise. In glass technology, it is seldom required to apply more advanced functions.

For facilitating practical application, a statistical analysis example will be demonstrated below. The example is entirely artificial. Ten experiments, produced in two laboratories, will be analysed with Equation (1) within a five component glass system. Table 1 gives the concentrations of the five components A, B, C, D, and E, the related property, and the laboratory designation. It may be known to the statistical analyst that in Laboratory 1 a new measurement method was attempted that is assumed to be as sensitive to glass composition changes as the well established method used in Laboratory 2, but also that is assumed to not give the correct absolute property value.

If component interactions need be analysed, several columns need to be added to Table 1 that contain the concentration products of interest, such as B.C, B.D, and B.E as well as squared terms, e.g. B^2. In this example, nonlinear terms will not be considered.

If the glass compositions to be analysed contain elements in various oxidation states, every oxidation state should be introduced as a separate variable. Often

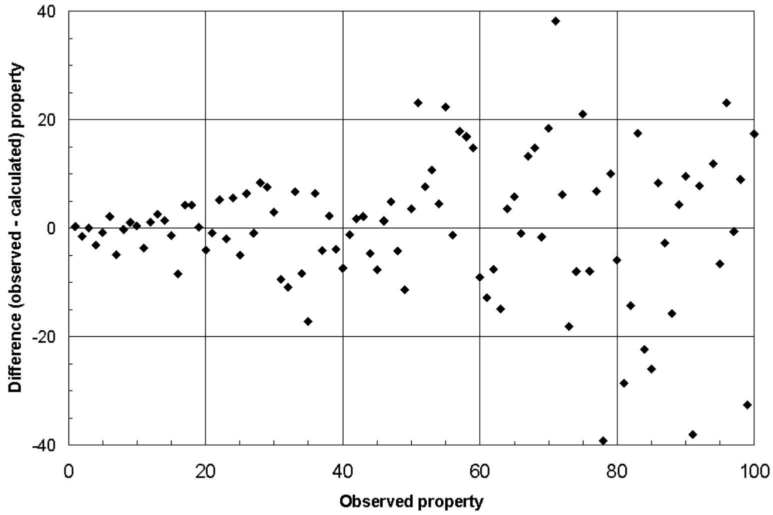

Figure 2. Logarithmic error trend

Table 1. Example glass composition–property data for statistical analysis

	Compositions in %					*Property*
	A	*B*	*C*	*D*	*E*	
Laboratory 1	88	1	0	1	10	69·4
	78	3	3	8	8	15·8
	73	5	7	9	6	18·6
	76	7	8	5	4	30·8
	88	9	0	3	0	25·7
Laboratory 2	75	2	10	4	9	71·1
	77	4	7	5	7	62·5
	84	6	1	4	5	53·3
	83	8	2	5	2	64·1
	74	10	9	7	0	93·4

however, the exact ratios of the oxidation states of an element in glass are unknown. Then, it is only possible to consider the sum of all oxidation states as one variable. For example, in most cases, the exact ratio of Fe(III)/Fe(II) in glass is not measured, and Fe(III) plus Fe (II) must be taken as a single variable (considering two Fe ions in Fe_2O_3 and one Fe ion in FeO).

If glass component interactions are well known from the literature, e.g. mixed alkali effects or boron oxide anomalies, it should be confirmed that corresponding interaction variables are established. Examples of this are the concentration products $Na_2O.K_2O$ and $Na_2O.B_2O_3$. All possible interaction variables also may be added, in case some of them turn out to have a significant influence. (Variable deselection will be described later based on correlation analysis and factor significance.)

If reasonable, further variables could be created, such as the inverse of concentrations to account for extreme end term properties, concentration ratios, etc, Ref. 6, p 286. It is not beneficial in many cases for common glasses with $C(SiO_2)$=40 to 85 mol% to consider the mentioned unusual variables because most of them can be reduced to linear influences of the glass components or component products within the concentration ranges studied. Unusual or complicated variables should be used only if they reduce the number of variables significantly, improve the model fit, or if there are other reasons to evaluate their impacts.

In cases where various data series are combined, as seen in the example given in Table 1, (Laboratories A and B), it is possible that so-called "block effects" occur. For example, it might happen that experimental results produced in one laboratory (or study) are systematically different from those produced in another laboratory (or study). This could be caused by a different calibration procedure or by different expertise. It also might be possible that a newly introduced measurement technique results in systematically different findings than previously established techniques. In the beginning of the statistical analysis procedure, it is best to assume that block effects might be present in *all* cases; even so, it could turn out later that most of them do not exist in fact. This means the creation of "categorical" or "dummy" variables[3,4,7] in Equations (1) or (7). In glass property modelling, block effects should only be introduced into the calculation with the *forward selection* approach (see below in section "Step 6: Calculation of the model standard error, the coefficient errors and significance"). All block effects must be excluded from the model initially and only considered if significant and reasonable.

Table 2. Analysis of block effects

Series	*Chemical Composition*			*Offset Variable*	*Property*
Laboratory A	…	…	…	1	P1
Laboratory A	…	…	…	1	P2
Laboratory A	…	…	…	1	P3
Laboratory A	…	…	…	1	P4
Laboratory A	…	…	…	1	P5
Laboratory B	…	…	…	0	P6
Laboratory B	…	…	…	0	P7
Laboratory B	…	…	…	0	P8
Laboratory B	…	…	…	0	P9
…	…	…	…	0	…

$$\text{Property observed}=\text{Equation (7)}+\beta(\text{offset}) \qquad (8)$$

$$\text{Property observed}=\text{Equation (7)} +\text{Property observed}\times\beta(\text{trend}) \qquad (9)$$

As a result, it would be required to add further columns to the regression dataset in Table 1, according to the number of data series examined. Following Equation (8), these columns would contain the value of "1" for each experiment within the series, and the value of "0" for each experiment of other series (see Table 2). Following Equation (9), the "1" would be replaced by the property value itself.

If only two data series are examined and block effects occur, it is not possible to decide which one of the data series should be given the block variable (see example in Ref. 20). Further knowledge may help one to decide.

Equation (9) should be applied with caution, and *only* if measurement trends are reasonable based on knowledge of the subject matter and if several other data series without a measurement trend including a high number of experiments exist. If Equation (9) is used for most experiments, the result would be a perfect fit with a trend coefficient of 1, even if all the experimental findings were seriously incorrect. A trend according to Equation (9) can be converted to an offset according to Equation (8) through logarithmic normalisation described in the previous section "Step 2: Calculation of property fixpoints and error normalisation".

Step 4: Correlation analysis

When using multiple regression, it is always important to evaluate possible factor correlations at the beginning. The *linear correlation matrix* is made up of the simple, or two way, correlation coefficients. They are denoted by the letter "r" and have a range of $-1<r<+1$ (Pearson's "r"). The correlation coefficient for two factors (variables) is a measurement of the linear relationship between the two factors. If r is close to 1 then a plot of the two factors against one another would look like a straight line with positive slope. If r is close to −1 then the plot of the two factors against one another would look like a straight line with a negative slope. If r is close to zero then a plot of the two factors would have no discernible linear trend.

For selecting the appropriate model variables, correlations between changes in the components' concentrations and/or their interactions (concentration cross products) have to be considered if the data were not collected using a statistical design that is "orthogonal" (not correlated) for all of the factors of interest. If the absolute value of r is larger than approximately 0·5 to 0·6, then the influences of the two considered factors are "partially correlated" (i.e. linked but not completely aliased) and may be difficult to separate. If the absolute value of r is larger than approximately 0·8 to 0·9, then the influences are correlated so strongly that they are likely not to be separated in most cases, and one of three actions should be taken: (1) the factors should be combined, (2) one factor should be excluded, (3) additional experimental data should be used to lower the correlation. The value of r can be calculated using

$$r=\frac{\sum xy-\sum x\sum y/n}{\left[\left(\sum x^2-\left(\sum x\right)^2\right)/n\right]\left[\left(\sum y^2-\left(\sum y\right)^2\right)/n\right]^{1/2}} \qquad (10)$$

where n is the number of experimental datapoints and x and y the variables that need to be tested for correlation.

A correlation is considered statistically significant if

$$t_{\alpha.\mathrm{DF}}<\left[|r|(n-2)^{1/2}\right]/\left(1-r^2\right)^{1/2}$$

where $t_{\alpha,\mathrm{DF}}$ is the t-distribution value depending on the confidence level α and the degrees of freedom DF with $DF=n-2$. Squared or cubic component influences are often strongly correlated to the corresponding linear influences. To follow the system of factor hierarchy (see below in this section), squared and cubic terms should only be considered after analysing simpler terms (linear effects and 2-factor interactions), and if the component concentrations vary widely.

If several data series with various glass compositions need to be analysed, systematic differences between the chemical compositions of the series can be detected by analysing the correlation coefficients between block effect variables (Table 2) and glass components.

Table 3. Linear correlation matrix; "Offset L1" stands for the offset or dummy variable for data from Laboratory 1

	B	*C*	*D*	*E*	*Offset L1*
B	1				
C	0·014	1			
D	0·208	0·472	1		
E	**−0·991**	0·050	−0·145	1	
Offset L1	−0·174	−0·298	0·044	0·148	1

The correlation matrix of the compositions given in Table 1 is displayed in Table 3, where the main component A (usually silica) is excluded according to the slack variable approach explained above.

It can be concluded from Table 3 that the components B and E are very strongly correlated in such a way that E usually decreases as B increases. Therefore, it is difficult to decide if property changes may be caused by component B or E, and consequently either B or E should be excluded from the calculation, or both may be added to form a new variable. If B and E have a similar chemical nature, like Na_2O and K_2O or MgO and CaO, the terms should be added.

In most cases, it is sufficient to delete the variable from the model of the lower hierarchy (see below) or the variable that is less represented, i.e. the one that occurs less often at lower concentrations and with little concentration variation.* Therefore, in the example based on Table 1, the component E will be excluded from further calculations.

Even after excluding the glass component E, some weaker correlations remain in Table 3, e.g. $r(C-D)$=0·472. This and similar effects always need to be considered when model results are interpreted, such as when the components C and D are not completely statistically independent. Therefore, it is preferable to interpret model predictions rather than individual coefficients.

For models including high order terms, the *factor hierarchy* should be considered. This principle implies that higher order terms such as B.C.D or B^2 may only be introduced if the significance (see below, section "Step 6: Calculation of the model standard error, the coefficient errors and significance") of all the related lower order terms was evaluated first (e.g. B, C, D, B.C, B.D, C.D). Low order terms are always preferred over corresponding high order terms in cases where they are partially correlated and significant (see below). Statistical modelling techniques follow the *principle of simplicity* (also called *principle of parsimony* or *Occam's razor*), a scientific approach that accepts as correct the simplest of several possible interpretations of a phenomenon.

The system of factor hierarchy is not strictly mandatory, however, if it does not make sense according

* A glass component with low concentration variation is one that almost always is present at a constant concentration level (its concentration hardly changes regardless of all other glass components).

Table 4. Example glass composition–property data for statistical analysis following the slack variable technique after correlation analysis

Compositions in %			*Offset*	*Property*
B	*C*	*D*	*L1*	
1	0	1	1	69·4
3	3	8	1	15·8
5	7	9	1	18·6
7	8	5	1	30·8
9	0	3	1	25·7
2	10	4	0	71·1
4	7	5	0	62·5
6	1	4	0	53·3
8	2	5	0	64·1
10	9	7	0	93·4

to previous experience in the field it is possible in the glasses studied that a component shows a strong influence only in combination with another component, without influencing the property by itself significantly. For example, B_2O_3 does not have a strong effect on the viscosity in borosilicate glasses in the transition range, however, the $B_2O_3.Na_2O$ interaction greatly increases the viscosity.[77]

Step 5: Determination of the model coefficients

After excluding the main component A (following the slack variable technique) and the component E (because of correlation) from Table 1, it is necessary to analyse the remaining data in Table 4.

The factor matrix **X** and the property matrix **Y** may be defined as seen in Table 5. If no intercept would be included, the first column would contain the main component A instead.

Given a linear regression model, the **co**efficient matrix **CO** can be estimated using the ordinary least square (OLS) method according to Equation (12).[79] The symbol "^{T}" stands for the matrix transpose operation, "$^{-1}$" indicates matrix inversion, and the sign "." means the scalar product. The term $(\mathbf{X}^T.\mathbf{X})^{-1}$ is the *inverse information matrix*. Multiplied by S^2 (see Equation (13) below) it is called a *variance–covariance matrix* because it contains variable variances (square of standard deviations) of all the model variables as its diagonal elements (printed bold in Table 6), and covariances in all other matrix positions.

Table 5. Example factor and property matrices

Factor matrix **X**					*Property matrix* **Y**
Intercept	*B*	*C*	*D*	*Offset L1*	
1	1	0	1	1	69·4
1	3	3	8	1	15·8
1	5	7	9	1	18·6
1	7	8	5	1	30·8
1	9	0	3	1	25·7
1	2	10	4	0	71·1
1	4	7	5	0	62·5
1	6	1	4	0	53·3
1	8	2	5	0	64·1
1	10	9	7	0	93·4

Table 6. Example inverse information matrix $(\mathbf{X}^T.\mathbf{X})^{-1}$

	Intercept	*B*	*C*	*D*	*Offset L1*
Intercept	**1·05733**	−0·06684	−0·03086	−0·05546	−0·32364
B	−0·06684	**0·01358**	0·00224	−0·00552	0·01960
C	−0·03086	0·00224	**0·01125**	−0·00957	0·02891
D	−0·05546	−0·00552	−0·00957	**0·02881**	−0·03233
Offset L1	−0·32364	0·01960	0·02891	−0·03233	**0·48966**

$$\mathbf{CO}=(\mathbf{X}^T.\mathbf{X})^{-1}.\mathbf{X}^T.\mathbf{Y} \tag{12}$$

From the factor matrix **X** in Table 6, the inverse information matrix $(\mathbf{X}^T.\mathbf{X})^{-1}$ should be determined first. Table 6 gives the inverse information matrix that is obtained in this case.

The coefficient matrix **CO**, determined from **X** and **Y** can be seen in Table 6. Hence the preliminary model would be

Property=81·5743+0·3096B+1·7997C−4·9981D
−31·5514 Offset for data from Laboratory 1

Step 6: Calculation of the model standard error, the coefficient errors and significance

The *model standard error S* can be derived from the *model residuals* Δ and the *degrees of freedom* DF. The model residuals are the differences between the experimentally observed and the calculated property values. Given the example in Table 1 and the model in Table 7, Table 8 displays the residuals. (This and the following tables contain more significant figures than necessary to supply an example that can be worked through and checked to see if exactly the same answers are obtained.)

The degrees of freedom DF is the difference between the number of independent experimental datapoints and the number of variables including the intercept. For the example in Table 1 the number of experimental datapoints is 10, and the number of variables including the intercept is 5, i.e. DF=5. The model standard error S (also called "*root mean square error*") can now be calculated using Equation (13)

$$S=[\Sigma(\Delta^2)/DF]^{1/2} \tag{13}$$

The model standard error S should be larger than the standard deviation of repeated measurements; otherwise, the model is "over fitted" (an unrealistically good fit is obtained). In addition, S should not be significantly larger (about 1·7 times) than the standard deviation of repeated experiments from several investigators ("under fitting"). An F-test can

Table 7. Example model coefficients

Intercept	81·5743
β_B	0·3096
β_C	1·7997
β_D	−4·9981
Offset L1	−31·5514

Table 8. Example model residuals

Property		
observed	*calculated*	*residual*
69·4	45·3345	24·0655
15·8	16·3663	−0·5663
18·6	19·1862	−0·5862
30·8	41·5975	−10·7975
25·7	37·8155	−12·1155
71·1	80·1982	−9·0982
62·5	70·4203	−7·9203
53·3	65·2396	−11·9396
64·1	62·6605	1·4395
93·4	65·8814	27·5186

be performed for quantitative evaluation of over/under fitting.

The model standard error S derived from the residuals in Table 8 and with five degrees of freedom is 19·4453. Approximately 68% of all residuals fall within the limits of ±S. The *standardised residual* is the quotient of the residual and the model standard error (the first standardised residual in Table 8 is 1·2376; the second is –0·0291, etc...).

The standard errors of the coefficients S_β are determined using

$$S_\beta=SC_{jj}^{1/2}=\beta/t_\beta \qquad (14)$$

where C_{jj} are the diagonal elements (marked in bold) of the inverse information matrix $(\mathbf{X}^T.\mathbf{X})^{-1}$ in Table 6. The elements C_{jj} are given in Table 9. The t-value is the ratio of the coefficient and its standard error, as seen in Equation (14). Table 10 summarises all coefficients, including their errors and t-values.

The absolute value of the t-value (also called the *t-statistic* or *t-ratio*) is a measure of a coefficient being equal to zero. It is an indicator of the significance of a model factor (variable, component concentration or concentration product (interaction factor)) in slack variable models. In other words, it is a measure of how much information a factor adds to the model. In general, a t-value with an absolute value greater than or equal to two is considered to be significant, with a statistical confidence level of approximately 95%. (This confidence level at t=2 will change slightly with the degree of freedom, but it is at least 90% for DF>4.) Most minor components are insignificant, i.e. their influence is less than the standard error ("noise"). If the t-value of a glass component indicates that its coefficient is insignificant, it should be concluded that its influence on the property being modelled is insignificant within the studied composition range.

It must be noted that t-values may also be calculated and used for models that do not include an intercept following the canonical approach by Scheffé as seen in Equation (5). The t-values are a measure of a coefficient to equal zero in models with and without intercept. However, it is demonstrated in a paper by Piepel[80] that it is not a meaningful hypothesis to assume that a coefficient equals zero in models without intercept. An alternative component slope mixture model is presented in Ref. 80. For practical application it is advised in this work to always use t-values for evaluating the probability for a coefficient to equal zero for *all* model forms, with appropriate interpretation of the meaning of the t-values in models *without* intercept.

Table 9. Example C_{jj} values from Table 6

Intercept	1·05733
β_B	0·01358
β_C	0·01125
β_D	0·02881
Offset L1	0·48966

Table 10. Example model coefficients, coefficient standard errors, and t-values

Variable	*Coefficient*	S_β	t_β
Intercept	81·5743	19·9949	4·0798
β_B	0·3096	2·2659	0·1367
β_C	1·7997	2·0628	0·8725
β_D	−4·9981	3·3008	−1·5142
Offset L1	−31·5514	13·6070	−2·3188

Another way of looking at a coefficient significance is possible through consideration of p-values. A p-value reflects the probability of a coefficient being equal to zero, derived from the t-value and the degree of freedom, and is generally based on normal error distribution. The p-value should be lower than 0·05 for a 95% confidence level.

A t-value reflects the significance of a coefficient within the model, but a special application might require a more strict limitation of the composition area than the model is valid for. The narrower the concentration range of a given component, the more and more a coefficient becomes practically insignificant.

It is obvious in Table 10 that the errors of the coefficients β_B, β_C, β_D are larger than their absolute values or of comparable size, and that the absolute values of their t-values are smaller than two. It is not possible with reasonable certainty to conclude whether the glass components B, C, and D have a significant influence within the examined composition range. Consequently, the glass components B, C, and D may be excluded from further calculations (Table 11). This exclusion must be performed stepwise because based on correlations it frequently happens that after exclu-

Table 11. Example factor matrix, after exclusion of the components B, C, and D

Intercept	*Offset L1*
1	1
1	1
1	1
1	1
1	1
1	0
1	0
1	0
1	0
1	0

Table 12. Example inverse information matrix ($(\mathbf{X}^T\cdot\mathbf{X})^{-1}$, after exclusion of the component B, C, and D

	Intercept	*Offset L1*
Intercept	**0·2**	–0·2
Offset L1	–0·2	**0·4**

sion of one insignificant variable another previously insignificant one becomes significant. In the example described here this is not the case.

The exclusion of components B, C, and D in the given example is called *stepwise backward elimination*, as initially all possible variables were included in the model, and the insignificant ones were excluded stepwise. The opposite approach would be *stepwise forward selection*, i.e. the most significant variable is included first in the model, followed by the next significant one, until no further significant variables can be found. During stepwise backward elimination and forward selection the factor hierarchy described above in the section "Step 4: Correlation analysis" should be taken into account.

For analysing large glass databases it is recommended to proceed stepwise as follows: (1) selection of all significant single glass components through backward elimination; (2) selection of the most significant and scientific reasonable two-component interaction factors (boron anomaly, mixed alkali effect) through forward selection; (3) analysis of systematic offsets in whole data series from selected laboratories and if necessary exclusion of those series that are incomparable with the majority of all other series;[81,82] (4) stepwise deletion of outliers and simultaneous selection/elimination of variables according to their significance and scientific reason until no further outlier can be found.

Models that include insignificant variables are termed *over fitted*, and often are unrealistically well fitted.

Next, it is necessary to repeat the procedure, beginning with Step 5 (determination of the model coefficients). The size of the inverse information matrix decreases to 2×2 (Table 12).

Table 13 shows the new coefficients, coefficient standard errors, and *t*-values. The new model standard error *S* is 18·6893, which is insignificantly lower than the previous value of 19·4453. Table 14 gives the new model residuals.

Step 7: Outlier analysis, data leverage

From the residuals in Table 14, it can be concluded that some of them seem to stand out from others. It is possible that experimental or data entry problems caused a datapoint to show an anomalous response or that the composition is outside the range where the linear approximation of component effects is valid. The latter can be checked easily by comparing the compositions of the outliers to other glasses in two dimensions (see Figure 3), or through leverage analysis, described below in this section.

Table 13. Example model coefficients, coefficient standard errors, and t-values, after exclusion of the component B, C, and D

Variable	*Coefficient*	S_β	t_β
Intercept	68·8800	8·35811	8·2411
Offset L1	−36·8200	11·8202	−3·115

Regression analysis assumes that the residuals are normally distributed. Thus, a datapoint may be regarded as an *outlier*:[10]

(1) if the absolute of the residual is larger than about three times the model standard error (=absolute of standardised residual larger than about three),
(2) if the largest residual is higher than about 1·5 times the next largest residual, or
(3) if the *externally studentised* residual* is higher than about three.

It is always possible to refer to the *p*-statistic (see above in the previous section, Step 6) to determine exact critical outlier limits using the desired confidence level such as 99·7%. However, in most cases, the approximate limits stated here are sufficient.

Understanding of the outlier conditions (1) and (2) can be derived from the explanations above. To clarify the meaning of the outlier condition (3), more details must be given about data leverage in statistical analysis.

The *h-value* measures how much "leverage" a particular row of data could have on the resulting overall correlations. It is also known as the Hat diagonal value, or diagonal value h_i of the *Hat-matrix* **H**. It is a measure of how good the experimental design is for all the variables in the model. In Figure 3 is a two-dimensional example of a point with a high *h*-value. One glass composition with high *h*-value stands out from all other compositions. The Hat-matrix **H** and

Table 14. Example model residuals, after exclusion of the component B, C, and D

Property observed	*calculated*	*residual*
69·4	32·0600	37·3400
15·8	32·0600	−16·2600
18·6	32·0600	−13·4600
30·8	32·0600	−1·2600
25·7	32·0600	−6·3600
71·1	68·8800	2·2200
62·5	68·8800	−6·3800
53·3	68·8800	−15·5800
64·1	68·8800	−4·7800
93·4	68·8800	24·5200

* The expression "studentised" is not a typographic error; "studentised" is different from "standardised." The term is in honor of the English statistician William Sealey Gosset (1876–1937) who published under the pseudonym "Student."

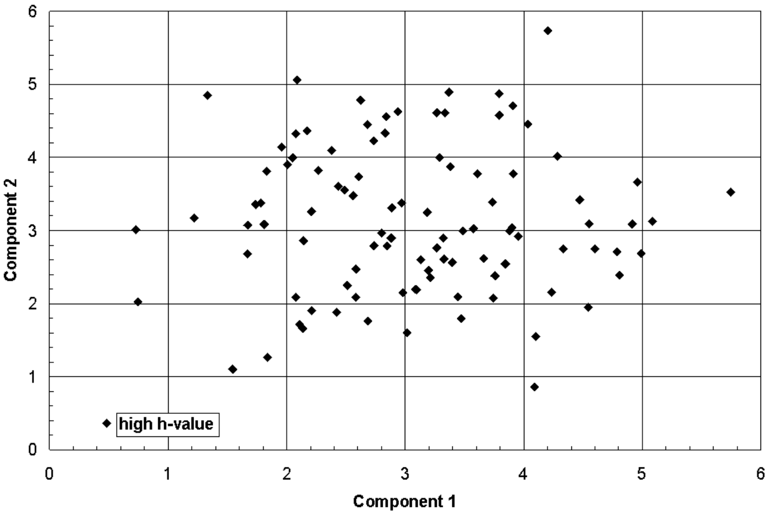

Figure 3. h-value demonstration

Table 15. Example Hat-matrix **H**, *derived from Table 11*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	**0·2**	0·2	0·2	0·2	0·2	0	0	0	0	0
2	0·2	**0·2**	0·2	0·2	0·2	0	0	0	0	0
3	0·2	0·2	**0·2**	0·2	0·2	0	0	0	0	0
4	0·2	0·2	0·2	**0·2**	0·2	0	0	0	0	0
5	0·2	0·2	0·2	0·2	**0·2**	0	0	0	0	0
6	0	0	0	0	0	**0·2**	0·2	0·2	0·2	0·2
7	0	0	0	0	0	0·2	**0·2**	0·2	0·2	0·2
8	0	0	0	0	0	0·2	0·2	**0·2**	0·2	0·2
9	0	0	0	0	0	0·2	0·2	0·2	**0·2**	0·2
10	0	0	0	0	0	0·2	0·2	0·2	0·2	**0·2**

the h-value h_i are defined by

$$\mathbf{H}=\mathbf{X}.(\mathbf{X}^{T}.\mathbf{X})^{-1}\mathbf{X}^{T}$$

$$h_i=\mathbf{x}_i^{T}.(\mathbf{X}^{T}.\mathbf{X})^{-1}.\mathbf{x}_i \qquad (15)$$

Table 15 shows the example Hat-matrix determined from the factor matrix in Table 11. The h-values h_i derived from the factor matrix in Table 11, are marked in bold and underlined. In the example described here, all experiments have the same leverage by chance. Please note that the sum of all h-values is exactly two, the number of variables in the model, including the intercept. The rule-of-thumb is that a glass composition has a high amount of leverage when the h-value is higher than two times the quotient of the number of variables including the intercept, over the number of experimental datapoints.[83] In the example above, it would mean that experiments have a large leverage if h is higher than 2×2/10, i.e. 0·4.

It is not recommended that a glass composition be deleted solely on the basis of its h-value.

While the *leverage of a glass composition on the overall correlations* is quantified by its h-value, the *leverage of one experiment on the model result* (i.e. the coefficients) is measured with the *Cook value*, defined in Equation (16).[84,85] The Cook value compares the model result (coefficients) with and without each experiment. Generally, an experiment with a Cook value larger than one has a high leverage. One should make sure that an experiment with a high Cook value is accurate, e.g. a NIST (National Institute for Standards and Technology) or DGG (German Society of Glass Technology) glass property standard; otherwise, the whole model might suffer from one single high leverage datapoint

$$\mathrm{Cook_i} = \frac{\Delta_i^2 h_i}{pS^2\,(1-h_i)^2} \qquad (16)$$

where p is the number of model variables including the intercept (=2 in Table 13). Table 16 lists the Cook values of all experiments from the model in Tables 11 to 14. The first experiment has a much higher Cook value (Cook=0·6237) than all others, but which is

Table 16. Example model statistics, model from Table 13

Property observed	*calculated*	*Residual Δ*	*h value*	*Cook value*	*Press residual*	S_i	*ES residual*
69·4	32·0600	37·3400	0·2	0·6237	46·6750	13·2324	3·1549
15·8	32·0600	−16·2600	0·2	0·1183	−20·3250	18·7521	−0·9695
18·6	32·0600	−13·4600	0·2	0·0810	−16·8250	19·0956	−0·7881
30·8	32·0600	−1·2600	0·2	0·0007	p1·5750	19·8167	−0·0711
25·7	32·0600	−6·3600	0·2	0·0181	−7·9500	19·6629	−0·3616
71·1	68·8800	2·2200	0·2	0·0022	2·7750	19·8036	0·1253
62·5	68·8800	−6·3800	0·2	0·0182	−7·9750	19·6619	−0·3628
53·3	68·8800	−15·5800	0·2	0·1086	−19·4750	18·8421	−0·9245
64·1	68·8800	−4·7800	0·2	0·0102	p5·9750	19·7327	−0·2708
93·4	68·8800	24·5200	0·2	0·2690	30·6500	17·2919	1·5854

smaller than one, i.e. it is still not significant. If the first experiment were removed from the discussed model example, the coefficients would change the most.

The *Press residual* is the quotient of the residual over (1−*h*), as given in Table 16. Press is the residual that would occur for a given experiment if that experiment were not included in the regression analysis calculation, but rather if it were predicted by a model including all other experiments. Hence, an experiment with high leverage on the model (Cook value) would have a high Press residual as well.

The *h*-value also makes it possible to determine the model standard error if one experiment were excluded (S_i) using

$$S_i=\{[(DF+1)S^2-\Delta_i^2/(1-h_i)]/DF\}^{1/2} \qquad (17)$$

In Table 16, it becomes clear that if the first experiment were excluded from the model, then the model standard error would decrease much more than for all other experiments. In the final step of the outlier analysis, the externally studentised residual (*ES residual*) may be derived from S_i using

$$\text{ES residual}=\Delta_i/[S_i(1-h_i)^{1/2}] \qquad (18)$$

The ES residual is a more sensitive outlier indicator than the residual Δ alone, because the data leverage is considered. The analysis of the ES residuals makes consideration of the generally employed *studentised residuals*, which do not take into account S_i from Equation (17) and which do not have a true *t*-distribution, obsolete. Table 16 displays all ES residuals from the model example above.

From the statistical indicators in Table 16, it can be concluded that the first experiment is probably an outlier. It should be evaluated in detail by an expert in the subject and excluded from the source database if deemed reasonable; the modelling procedure should then be started all over again, beginning with Step 4 (correlation analysis). Table 17 shows the new correlation matrix, Table 18 shows the final model, and Table 19 shows further statistical indicators. No further outlier can be detected. In the final model the last experiment has a relatively high leverage (Cook=0·9230), which is still lower than one, i.e. it is still considered to have an insignificant leverage on the model. However, during model validation it should be evaluated if the experimental result is reliable.

In the given example it turns out that there exists a systematic difference between the data from Laboratory 1 and Laboratory 2. The calibration procedure for the new measurement method attempted in Laboratory 1, stated above in connection with Table 1, must be re-evaluated.

Systematic errors in whole measurement series that lead to an offset are known in the glass science literature.[65,86]

Step 8: Goodness-of-fit evaluation, validation, model improvement

The goodness of a model fit (the measure of how well the glass properties can be described with the given glass compositions) is generally expressed in R^2 values, defined by

$$R^2=1-\{\Sigma(\Delta^2)/\Sigma[(\text{Observation}-\text{Average observation})^2]\} \qquad (19)$$

$$R^2\text{, adjusted}=R_A{}^2=1-[(1-R^2)(n-1)]/DF \qquad (20)$$

$$R^2\text{, predicted}= R_P{}^2=1-\{(\text{Press}^2)/\Sigma[(\text{Observation}-\text{Average observation})^2]\} \qquad (21)$$

R^2, also known as *coefficient of determination*, is a general goodness-of-fit indicator for models including an intercept (slack variable) that shows the fraction of the total variance in the dependent variable (glass property) that is explained by the model. If R^2 is close to one, it either means that the model is good or that it could be the result of over fitting with insignificant and correlated variables.

R^2 can be calculated and interpreted for the models described in this work with and without an intercept (Equation (5)).[87] However, the reader should be aware of the fact that in some model forms with a forced intercept R^2 is not defined,[88,89] and some software packages calculate R^2 incorrectly or do not permit its calculation at all. For glass property modelling described in this study R^2 should only be determined using Equation (19).

$R_A{}^2$, originally established for models without an intercept (Ref. 6, p 532), is adjusted for the degrees of freedom in multiple regression. If $R_A{}^2$ is significantly lower than R^2, it can be concluded that the model is over fitted. Therefore, $R_A{}^2$ is often a better goodness-of-fit indicator than R^2 for multiple regression.

$R_P{}^2$, also called $R^2{}_{\text{PRESS}}$, is calculated in the same way as R^2, but using the Press residuals instead of the residuals. $R_P{}^2$ shows the expected fraction of the variance in the dependent variable (glass property) that can be explained by the model for predicting new experiments. If $R_P{}^2$ is significantly lower than R^2 and $R_A{}^2$, it means that some experiments have a high leverage.

The model discussed above has relatively high R^2 values, with R^2=0·9678, $R_A{}^2$=0·9485, and $R_P{}^2$=0·8743; i.e. the model fit appears to be good. $R_P{}^2$ is somewhat lower than R^2 and $R_A{}^2$ because of the last high leverage datapoint. The last datapoint may be re-evaluated.

In glass property modelling using statistical analysis the values of the R^2 indicators are often higher than 0·9,[16–21,60,61,64,65] especially for properties that are relatively easy to measure such as the viscosity. For properties that are difficult to obtain, e.g. gas solubilities in glass melts, R^2 tends to be lower.[20] In

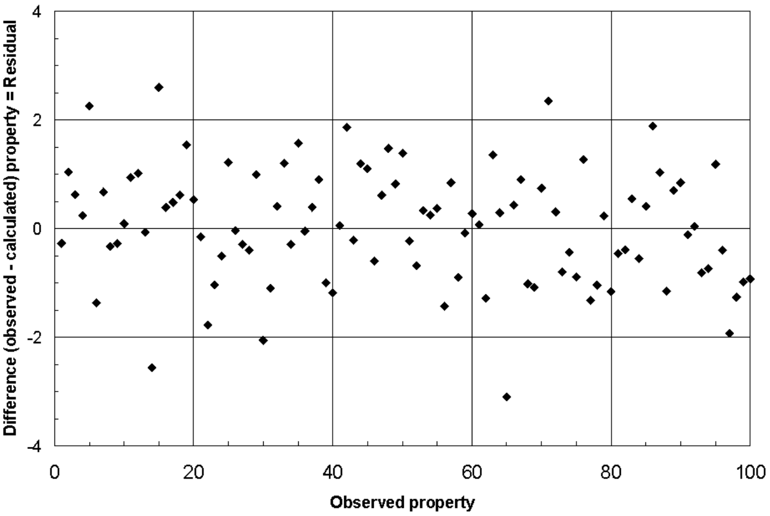

Figure 4. Normal residual distribution

general, R^2 values higher than 0·8 may be considered as good.

A good model fit could be a sign of reproducible experiments; further hidden variables are unlikely to be discovered (correlations not considered). It is not clear yet, however, which one of the strongly correlated glass components B and E is causing the observed property changes. Add on experiments should be performed for decorrelating the glass components B and E. Table 17 also shows that some weaker correlations still remain, which always needs to be considered during interpretation of the model results. Some variables are not completely statistically independent. Therefore, it should be preferred to interpret model predictions rather than individual coefficients.

For model evaluation, the residuals should be plotted against the experimentally observed property values and against the calculated property values. Ideally, the result would be as shown in Figure 4, i.e. the residuals are normally distributed. However, if a residual variance trend occurs as displayed in Figure 2, logarithmic data transformation for error normalisation may be evaluated. If a plot is obtained as in Figure 5, the model may still not include an important variable.

The residuals also should be plotted against each variable. A pattern as seen in Figure 6 indicates that a squared term for the considered variable (glass component 1) should be introduced.

If stepwise residual trends occur, "block effects" may be present. Some data series may be different than others (different property values), or the errors within various data series may be different.

Table 17. New linear correlation matrix

	B	*C*	*D*	*E*	*Offset L1*
B	1				
C	–0·269	1			
D	–0·159	0·298	1		
E	**–0·990**	0·321	0·210	1	
Offset L1	0·000	–0·183	0·328	–0·016	1

For statistical *model validation*, the differences between *precision* (*repeatability*), *reproducibility*, and *accuracy* must be taken into account. The precision reflects the consistency and repeatability within a data series of one experienced investigator, generally using one measurement technique. The reproducibility is a measure of how well other experienced investigators in other laboratories can reproduce the experiment. The accuracy shows the similarity to the "true" or "mean" value in case the absolute truth is known. It is often assumed that experiments reproduced by several experienced and independent investigators are very close to being accurate, e.g. NIST or DGG glass property standards.

Consequently, for models based on one single investigator, a reproducibility and accuracy can *not* be established; only the precision may be evaluated. However, in high quality publications that contain experimental data the author is always using external values for calibration and/or comparison. Therefore, even for some models based on one single study repeatability and accuracy can be established. For models based on several investigators, the reproducibility may be determined, which can be assumed to come close to accuracy if many investigators agree. *Statistical model validation* can be obtained by:

(1) Splitting of the source data into one set for modelling and a second set for comparing of predicted and experimental data;
(2) Comparing the model predictions to experimental data from another investigator;
(3) Comparative modelling of two data series from different investigators where coefficients and residual trends are compared with and without the second series;
(4) Comparative modelling of several data series from various investigators including careful

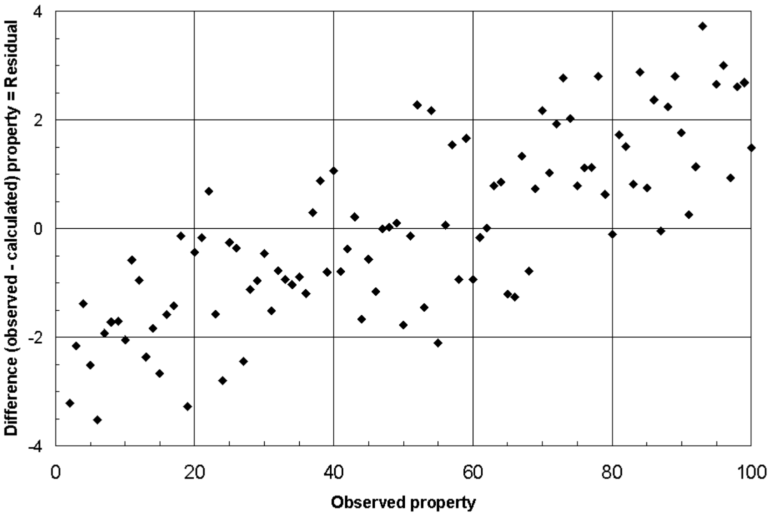

Figure 5. Residual trend

analysis of correlations, over/underfitting, systematic trends, and data leverage;

(5) Developing two independent models based on data series from different investigators in similar composition regions and comparing the model coefficients considering correlations, and

(6) Developing two independent models, including all possible component interactions based on data series from different investigators in different composition regions (compositions in mol%) and comparison of the model coefficients considering correlations.

If the standard errors are comparable to the errors found during model evaluation and correlations/trends are considered, it can be assumed that the model is accurate, i.e. it is "validated." Method (1) can be used for an internal validation of the model precision, and methods (2) to (6) allow conclusions to be drawn concerning total accuracy by comparing the results with other investigators.

In addition to standard error comparison, it is also possible to determine R^2, validation= R_V^2. R_V^2 is calculated in the same way as R^2 in Equation (19), but instead of the model residuals, the residuals from the validation test are used. For example, a model may be established based on 80% of all data, and then predictions are made for the remaining 20%, and only the residuals of the 20% of the data are considered for the R_V^2 calculation.

An *experimental model validation* is superior to the statistical validation explained above. In addition, it is often possible with experimental model validation to improve the model considerably, resulting in a prediction error reduction up to 20 to 50% according to the experience of the author. Experimental validation and model improvement should be approached as follows:

(1) First, the reasonability of dataset-specific categorical offset variables must be examined. It has to be investigated experimentally if and how findings from different laboratories are comparable. The most contradictory findings from different

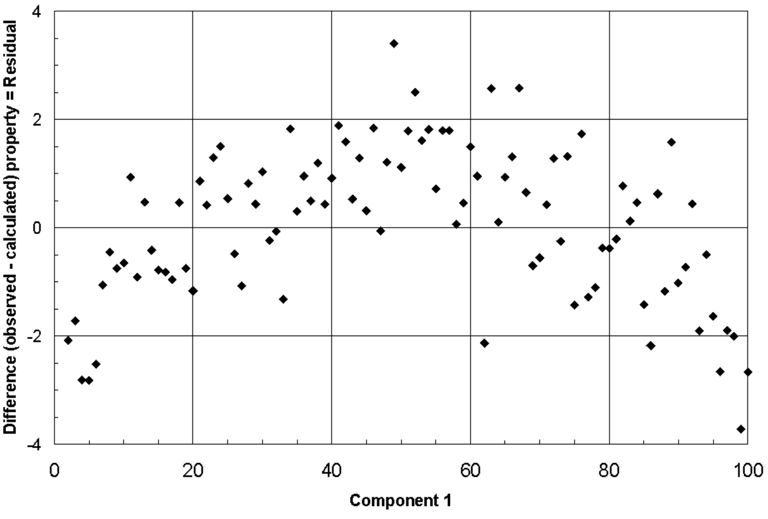

Figure 6. Residual trend versus a variable (glass component 1)

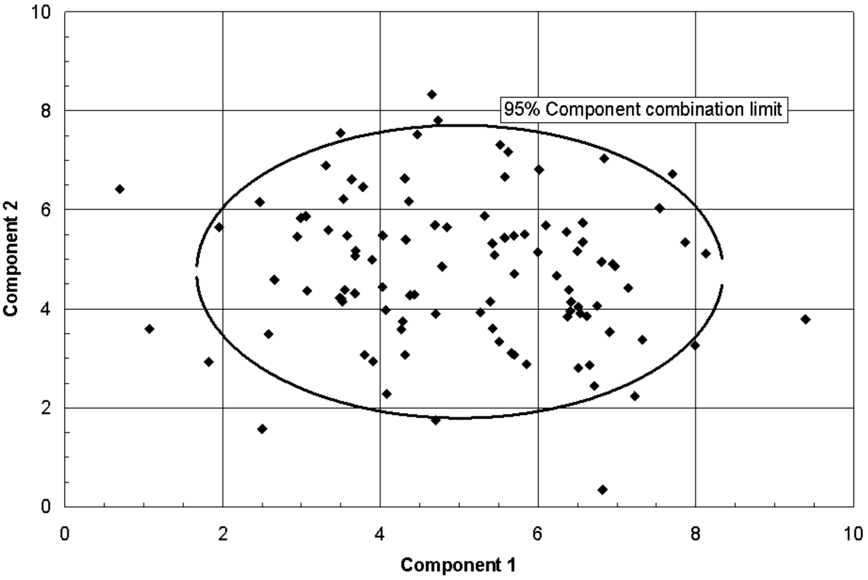

Figure 7. Binary 95% component combination limits; 95% of all data are within the limits of the ellipse. Please note that the limits are not characterised by a rectangle instead of the ellipse shown. The corners of a rectangle would be outside the source data, which is not considered in almost all glass property models

laboratories should be reproduced.

(2) The glasses with the greatest influence on the overall understanding (high Cook values) need to be re-analysed and/or remelted and excluded from the model if crystallisation or phase separation occurs, if undissolved batch material remains in the glass, if the bubble content is very high, if extremely strong striations are visible, or if other similar irregularities can be observed.

(3) The model result (glass component influences on the property) should evaluated from a glass science standpoint. It must be estimated whether the model is reasonable and makes any scientific sense. Surprising or unusual glass component influences need to be traced back to the individual glasses that are the most significant causes of these influences, and those glasses have to be remelted and measured. For example, if all glass components influence the property for up the 10 property units per mol%, and one minor component appears to influence the property for 5000 property units per mol%, then it is possible that this phenomenon is rather difficult to understand scientifically. The mentioned minor component with the exceptional influence should be analysed experimentally.

(4) Finally, add-on experiments can be performed to reduce mutual correlations of glass components, and to expand the model into new composition areas.

In general, a good multiple regression model has the following properties:[10]

- All factors in the model are significant (absolute value of *t*-values>2), and all excluded factors are insignificant (absolute value of *t*-values<2), i.e. there are no over/under fitting occurs.
- Accurate predictions can be made using the model. The standard error of the model *S* is not significantly (no more than about 1·7 times) larger than the standard deviation of repeated experiments from several investigators.
- The standard error of the model *S* is higher than the standard deviation of repeated experiments from several investigators, i.e. the model is not over-fitted.*
- The coefficients are reasonable, according to the judgment of experts familiar with the modelled property.
- Follow-up experiments within the model application limits agree with the model predictions.

Step 9: Model application

For all glass property models published in the literature, the compositional area is defined for which the model is valid. Those composition limits are stated as minima and maxima of all glass components considered in the specific model. However, in many cases, it may happen when two or more concentrations are set to extrema that those extreme *combinations* were not investigated in fact, even though the extreme single component influences were. Consequently, models may result in inaccurate predictions for

* It is often impossible to estimate the standard deviation of repeated experiments from several investigators because of the lack of data for multicomponent glasses. Therefore, it is advised to collect all published values available from binary glass systems.[78] A polynomial fit of the second or third degree often describes those data well. The standard error obtained from binary systems may be assumed to be similar to the error concerning multicomponent glasses.

certain component combinations if the combination (i.e. interaction) limits are not well defined.

The *component combination limits* for glass property models state the maxima and minima of all component combinations that are covered by the model. The component combination limits need to be specified in addition to the concentration limits of all glass components.

Given the uncorrelated components 1 and 2 with normally distributed concentration values, the binary component combination limits may be expressed as seen in Figure 7 and defined using

$$U_2=s_2/s_1[(t_{\alpha.\mathrm{DF}}s_1)^2-(c_1-a_1)^2]^{1/2}+a_2 \qquad (22)$$

$$L_2=-s_2/s_1[t_{\alpha.\mathrm{DF}}s_1)^2-(c_1-a_1)^2]^{1/2}+a_2 \qquad (23)$$

where U_2, L_2 are the upper and lower concentration limits of component 2, s_1, s_2 are the standard deviations of all concentration values of components 1 and 2, a_1, a_2 are the average concentrations of the components 1 and 2 and c_1 is the concentration of component 1. Ternary component combination limits can be similarly derived from a spherical function like $f(x,y)=(\text{radius}^2-x^2-y^2)^{1/2}$.

If the components 1 and 2 in Equations (22) and (23) are linearly correlated the constant term a_2 in the Equations (22) and (23) would need to be replaced by a linear function $c_2=\text{slope}c_1+\text{intercept}$.

Equations (22) and (23) can be applied in many cases for glass property modelling if the component 1 is silica (SiO_2), i.e. most binary component combination limits [(SiO_2)–(other component)] can be quantified through Equations (22) and (23). However, it is not possible to use Equations (22) and (23) for component combination limits of most of the remaining glass components besides SiO_2 because of a non-normal distribution of the data. For practical application, it is beneficial to define the component combination limits as follows:

(1) Maxima and minima of component concentration products (minima mostly zero)=interaction variables in Equation (7),

(2) Maxima and minima of component concentration sums (e.g. for excluding binary glasses like SiO_2–CaO or ternary glasses like SiO_2–B_2O_3–Al_2O_3 from model predictions that are not part of the source data),

(3) Maxima of the terms (SiO_2+Na_2O).SiO_2, (SiO_2+K_2O).SiO_2, (SiO_2+PbO).SiO_2, and similar terms for excluding predictions in the binary systems SiO_2–Na_2O, SiO_2–K_2O, SiO_2–PbO, etc. at high concentrations of SiO_2.

In addition, model predictions should be considered unreliable if the 95% prediction confidence interval SCI given below in Equation (26) is higher than three times the model standard error S.

Model predictions can be made using Equations (1) or (7).

Table 18. Final example model coefficients, coefficient standard errors, t-values, and model standard error

Variable	*Coefficient*	S_β	t_β
Intercept	36·4174	7·3308	4·9678
β_B	3·1107	0·8195	3·7960
β_C	2·3791	0·6113	3·8915
Offset L1	−43·0622	4·1720	−10·3218
β_D	0	insignificant	
S=6·1053			

The *standard prediction error of the mean* for a glass composition of interest (*PE*, **p**rediction **e**rror) can be determined using

$$PE=S[\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0]^{1/2} \qquad (24)$$

with $\mathbf{x}_0$ being the factor 1-column matrix derived from the glass composition of interest, and $\mathbf{x}_0^{\mathrm{T}}$ being its 1-row transpose. *PE* is generally lower than *S*. For example, if the *PE* is determined for a glass from Laboratory 2 containing 3% B and 5% C, the matrix $\mathbf{x}_0^{\mathrm{T}}$ would be [1|3|5|0], and the prediction is 57·6448±2·3579 (*PE*=2·3579). This means that there is a confidence of about 68% ($t_{\alpha,\mathrm{DF}}$~1) that the average property of multiple glass samples with the desired composition would be as predicted with an error of ±2·3579. *PE* is also a measure for the *model sensibility limit*, i.e. how large model prediction differences must be to represent a real difference, and below which any small differences may be considered as zero. With a model prediction difference of PE and $t_{\alpha,\mathrm{DF}}$>3, there is a confidence of about 62% that the different predictions are equal; with a difference of about 0·12×PE this confidence increases to 95%, and with a difference of about 3·9×PE the confidence decreases to 5%. It should be the goal to keep this confidence low. That means for slightly different model predictions it is desirable to make sure that the predictions are *not* equal.

The standard *confidence interval* of the mean model prediction is obtained by multiplying the standard prediction error *PE* by the *t* distribution value $t_{\alpha,\mathrm{DF}}$. For a 95% confidence and *DF*>15, $t_{\alpha,\mathrm{DF}}$ can be approximated as 2. *DF*=5 for the model described above.

Naturally, the standard error for predicting a single future experiment (*PEF*, **p**rediction **e**rror for a single **f**uture experiment) is higher than the standard error for the predicting the mean response (*PE*). The *PEF* may be estimated using

$$PEF=(S^2+PE^2)^{1/2} \qquad (25)$$

Comparing *PE* and *PEF* demonstrates the fact that single experiments from one laboratory are less valuable than the modelled mean of a number of reproduced experiments for evaluating property values.

The standard prediction confidence interval of the mean for *multiple* glass compositions or the **s**imulta-

Table 19. Final example model statistics, model from Table 18

Property		*Residual*	*h*	*Cook*	*Press*	S_i	*ES*
observed	*calculated*	*Δ*	*value*	*value*	*residual*		*residual*
15·8	9·8244	5·9756	0·4678	0·3955	11·2277	5·5957	1·4638
18·6	25·5620	−6·9620	0·3123	0·2147	−10·1236	5·5348	−1·5168
30·8	34·1624	−3·3624	0·4166	0·0928	−5·7633	6·3917	−0·6887
25·7	21·3511	4·3489	0·5159	0·2792	8·9836	6·0759	1·0287
71·1	66·4295	4·6705	0·5416	0·3770	10·1882	5·9341	1·1625
62·5	65·5136	−3·0136	0·2689	0·0306	−4·1217	6·4997	−0·5422
53·3	57·4604	−4·1604	0·4310	0·1546	−7·3120	6·2166	−0·8872
64·1	66·0608	−1·9608	0·3610	0·0228	−3·0684	6·5975	−0·3718
93·4	88·9357	4·4643	0·6850	**0·9230**	14·1736	5·6635	1·4045

neous **c**onfidence **i**nterval of the mean

$$SCI=PE(pF_{\alpha.\mathrm{p.DF}})^{1/2} \tag{26}$$

reflects the certainty that all of several predicted values are within the specified range with the desired confidence (*S*-method[90]). SCI should be preferred over PE in glass technology because it shows the confidence related to mass production. The 95% SCI for a glass containing 3% B and 5% C following the model in Tables 18 and 19 is 57·6448±10·7454 (SCI=10·7454 with $F_{95\%,4,5}$=5·1922). If the glass containing 3% B and 5% C were mass produced, then 95 out of 100 glasses are predicted to have the desired property with an error of ±10·7454 because of model uncertainty.

The prediction errors *PE*, *PEF*, and *SCI* described above are valid only if the chemical glass composition is known exactly. However, the chemical composition within a glass melting tank may vary, and the accuracy of a chemical analysis of a sample from the glass tank may vary as well. For a good estimation of the predicted property error, the chemical composition variation within a glass melting tank must be quantified through chemical analysis. If the glass composition in the tank is systematically different in specific locations, those different compositions should be used for property prediction within those locations. In case the glass composition varies randomly, several measurements should be taken, and the standard deviation of each glass component, which is the result of random variations of the chemical composition in the tank and the analytical measurement error, should be calculated. The glass composition uncertainty can now be converted to a property **p**rediction **c**onfidence **i**nterval due to **c**hemical uncertainty (*PCIC*) using

$$PCIC=t_{\alpha.\mathrm{DF}}(\mathbf{CO}^{\mathrm{T}}.\mathbf{SC}.\mathbf{CO})^{1/2} \tag{27}$$

with **CO** being the 1-column **co**efficient matrix from Equation (12). The degrees of freedom *DF* in Equation (27) may be assumed to equal the number of samples used to determine the **SC** (see below) minus one.

The chemical composition variance matrix **SC** (S for σ^2, C for **c**hemical) in Equation (27) is defined in Table 20. It consists of a diagonal matrix that contains the square of the standard deviation σ^2 (=*variance*) for each glass component included in the model as diagonal elements, and zero in all other matrix positions. In case the model included squared or interaction variables according to Equation (7), they need to be considered in the chemical composition variance matrix as well. If the main glass component silica (SiO_2) were excluded following the slack-variable technique in Equations (1) and (7) the matrix **CO** should be recalculated according to the canonical model form in Equation (5) following equations listed by Piepel (Ref. 91, p 190) and Cornell (Ref. 6, p 15–18 and 333–343).

The total prediction confidence interval CI_{total} caused by model uncertainty and the uncertainty of the chemical glass composition to be predicted is the sum of *SCI* and *PCIC*

$$\mathrm{CI}_{\mathrm{total}}=\mathrm{SCI}+\mathrm{PCIC} \tag{28}$$

The technique described in this work can be successfully applied for glass property modelling as demonstrated in previous papers.[20,64,77]

Analysis summary

The advantages of the presented statistical analysis method may be summarised as follows:

Superior accuracy: The large number of source data points allows a higher accuracy than the error arising from a few experiments.

Time and financial savings: Calculations can be completed within minutes. In contrast, one single experimental investigation may take several hours or days, including large personnel and equipment expenses, and the measurement ac-

*Table 20. Chemical composition variance matrix (*SC*)*

	Component 1	*Component 2*	*Component 3*	*Component 4*	*Component 5*
Component 1	σ_1^2	0	0	0	0
Component 2	0	σ_2^2	0	0	0
Component 3	0	0	σ_3^2	0	0
Component 4	0	0	0	σ_4^2	0
Component 5	0	0	0	0	σ_5^2

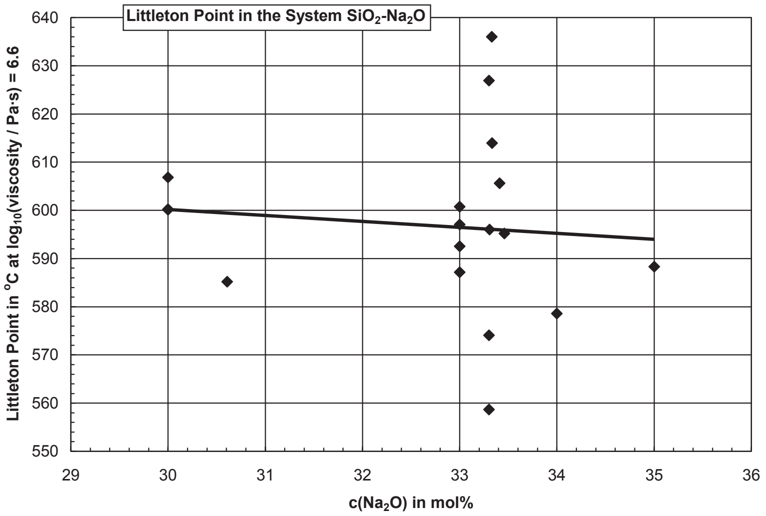

Figure 8. Littleton softening point in the binary system SiO_2– Na_2O for Na_2O=30–35 mol%

curacy still needs to be confirmed statistically afterwards by comparing results with those of other investigators.

Compatibility: Multiple regression makes data comparable over wide composition ranges, different measurement techniques, and from various investigators. Previous glass property models can be integrated (e.g. Lakatos *et al*; Hrma *et al*).[20] Without multiple regression, only data in binary glass systems are directly comparable.

Broad application range: The accuracy of new measurements mostly cannot be derived directly from similar results in the scientific literature or from NIST or DGG glass property standards. The multiple regression modelling approach may include several glass types with different chemical compositions; it makes them comparable. Consequently, the model allows predictions in composition areas that are not covered by common industrial models that are valid for only one specific glass type. The results of experiments inside and outside conventional ranges can be predicted and compared economically.

Future work

It is recommended that the demonstrated method be applied to all experimental values in the databases SciGlass[1] and Interglad[2] for "cleaning up" and organising the large volume of collected data, as in the work started by the author.[64,77] For example, up to now, such a well investigated glass property as the Littleton softening point in a simple binary system like SiO_2–Na_2O is not known exactly, as demonstrated through Table 21 and Figure 8. For other less investigated properties and multicomponent glasses, the uncertainties are more significant. It is necessary to quantify property values and the corresponding errors in detail.

The quality of the experimental data published by various authors should be evaluated based on statistical procedures described in this work and knowledge of the subject matter, including systematic offsets of whole series and the residual scattering compared with other investigators. A "quality rating" for publications, authors, or institutions would help to significantly improve the accuracy of property predictions.

In the future, systematic experiments should be performed to fill the numerous remaining "blank spots" of glass properties. For instance, little or nothing is known about the influences of binary glass component interactions on properties besides very limited information in confined systems concerning mixed alkali effects, boron anomalies, or alumina interactions.

Some steps of the statistical analysis procedure can be automated; hence, it is possible to introduce them into the databases SciGlass[1] and Interglad[2] directly,

Table 21. Littleton softening point in the binary system SiO_2–Na_2O for Na_2O=30–35 mol%

Author	*Year*	*c(Na_2O) in mol%*	*L.P. in °C*
K. S. Evstropiev	1940	34·00	578·6
G. S. Meiling	1967	33·46	595·2
K. S. Evstropiev	1968	33·30	558·7
Shvaiko-Shvaiko	1968	30·00	606·8
Shvaiko-Shvaiko	1968	35·00	588·3
O. G. Ivanov	1969	30·00	600·2
O. V. Mazurin	1970	30·00	600·1
K. Matusita	1973	33·30	574·1
U. E. Schnaus	1976	33·00	600·8
W. H. Dumbaugh	1978	33·31	596·0
N. A. Ghoneim	1984	33·33	636·0
Y. Shiraishi	1987	33·00	597·0
R. Ota	1991	33·30	626·9
R. Ota	1995	33·33	613·9
M. Liska	1996	33·41	605·6
D. Ehrt	1997	33·00	592·5
D. B. Dingwell	1998	30·61	585·2
D. Ehrt	2001	33·00	587·1

Reference: SciGlass Database and Information System 4·0 especially for well investigated properties and within often examined glass composition ranges. It is not recommended to automate *all* steps described in this work; glass science and statistical expertise must be considered.

The analysis procedure presented in this study is a powerful and economic tool for data organisation and modelling, based on empirical and impartial phenomenology. However, the technique does *not* allow a detailed scientific understanding of all states and processes within glass, e.g. on the atomic scale. To obtain a better insight into the nature of glass, the simple linear and polynomial equations applied in this work may be replaced by equations with physical meaning such as those described in the nonlinear regression section above.

Conclusions

A statistical analysis method has been presented that enables the modelling of glass properties with high accuracy. It is possible for any user to perform all steps of the analysis in commonly available spreadsheet software.

Statistical analysis allows measurement accuracy to be established through the combined modelling of a high number of original data from different sources. The modelled accuracy is often superior to a few test measurements within one laboratory because of the elimination of the influence of systematically different experimental conditions and/or systematic errors.

The detailed knowledge of glass property–composition relations makes targeted investigations possible with minimal time and financial investment.

The presented statistical analysis technique does not allow a detailed physical understanding of all states and processes within glass, e.g. on the atomic scale. Further work is necessary in this area.

For obtaining highly accurate glass property predictions it is recommended to systematically analyse all available glass databases, and to introduce a "quality rating" for publications.

Acknowledgments

The author would like to thank David Earl (Alfred University) for teaching the basics of statistics and providing very helpful guidance and advice. In addition, special thanks are directed to Alexander Priven (Russia) for reviewing this paper.

References

1. SciGlass 6.5 Database and Information System, 2005. http://www.sciglass.info/
2. International Glass Database System INTERGLAD Ver.6; New Glass Forum, Tokyo, Japan; http://www.ngf.or.jp
3. Montgomery, D. G. *Design and Analysis of Experiments*, John Wiley & Sons, 2001.
4. Dowdy, S. & Wearden, S. *Statistics for Research*, John Wiley & Sons, 1983.
5. Myers, R. H. & Montgomery, D. C. *Response surface methodology*, 2nd edition, John Wiley & Sons, 2002.
6. Cornell, J. A. *Experiments with Mixtures - Designs, Models and the Analysis of Mixture Data*, Wiley Series in Probability and Statistics, Wiley-Interscience, third edition, January 2002.
7. Draper, N. R. & Smith H. *Applied regression analysis*, John Wiley & Sons, 1998.
8. Dempster, A. P. *Elements of Continuous Multivariate Analysis*, Addison-Wesley, New York, New York, 1969.
9. Neter, J., Kutner, M. H., Wasserman, W. & Nachtsheim, Ch. J. *Applied Linear Regression Models*, third edition, McGraw-Hill/Irwin, 1996.
10. Harold S Haller & Company; 5 Ashley Court; Cleveland, Ohio 44116; USA. http://www.haroldhaller.com/; Bowles, R. L. & Haller: H. S. *Advanced Process Improvement Methods*; Harold S Haller & Company, Cleveland, Ohio, 1993. User guide of the MCA software, Haller Information Technology System.
11. Piepel, G. F. & Redgate, T. Mixture Experiment Techniques for Reducing the Number of Components Applied for Modeling Waste Glass Sodium Release, *J. Am. Ceram. Soc.*, 1997, **80** (12), 3038–3044. Piepel, G. F., Szychowski, J. M. & Loeppky, J. L. Augmenting Scheffé Linear Mixture Models with Squared and/or Crossproduct Terms, *J. Qual. Technol.*, July 2002, **34** (3), 297–314.
12. Blume, R. & Drummond III, C. Modeling and optimization of solar-control glasses, *J. Am. Ceram. Soc.*, May 2002, **85** (5), 1070–1076.
13. Winkelmann, A. Über die specifischen Wärmen verschieden zusammengesetzter Gläser (Heat capacity of various glasses), *Annal. Phys. Chemie*, 1893, **49**, 401–420; Winkelmann, A. & Schott, O. Über die Elastizität und über die Druckfestigkeit verschiedener neuer Gläser in ihrer Abhängigkeit von der chemischen Zusammensetzung (Dependence of the elasticity and strength of various new glasses from the chemical composition), *ibid*, 1894, **51**, 697–730; Winkelmann, A. & Schott, O. Über thermische Widerstandscoefficienten verschiedener Gläser in ihrer Abhängigkeit von der chemischen Zusammensetzung (Dependence of the thermal resistance of various glasses from the chemical composition), *ibid*, 1894, **51**, 730–746; Winkelmann, A. Über die Elasticitätscoefficienten verschieden zusammengesetzter Gläser in ihrer Abhängigkeit von der Temperatur (Temperature dependence of the elasticity coefficient of various glasses), *ibid*, 1897, **61**, 105–141 (in German).
14. Volf, M. B. *Mathematical Approach to Glass*, Glass Science and Technology, vol. 9, Elsevier, 1988.
15. Gehlhoff, G. & Thomas, M. *Z. techn. Phys.*, 1925, **6**, 544; *Z. techn. Phys.*, 1926, **7**, 105 and p 260; *Lehrbuch der technischen Physik (Applied Physics)*, J. A. Barth-Verlag, Leipzig, 1924, p 376 (in German).
16. Lakatos, T., Johansson, L.-G. & Simmingsköld, B. Viscosity temperature relations in the glass system SiO_2–Al_2O_3–Na_2O–K_2O–CaO–MgO in the composition range of technical glasses, *Glass Technol.*, June 1972, **13** (3), 88–95.
17. Öksoy, D., Pye, D. L. & Boulos, E. N. Statistical analysis of viscosity-composition data in glassmaking, *Glastech. Ber. Glass Sci. Technol.*, 1994, **67** (7), 189–195.
18. Hrma, P. R., Piepel G. F. *et al* Property/Composition Relationships for Hanford High-Level Waste Glasses Melting at 1150ºC; PNL Report 10359 to the US Department of Energy, vol. 1 and 2, Contract DE-AC06-76RLO 1830, December 1994. http://www.osti.gov/dublincore/gpo/servlets/purl/10121755-P8oQTl/webviewable/; http://www.osti.gov/dublincore/gpo/servlets/purl/10121752-cDjMo0/webviewable/; Vienna, J. D., Hrma, P. R. *et al* Effect of Composition and Temperature on the Properties of High Level Waste (HLW) Glass Melting above 1200ºC (Draft); PNNL Report 10987 to the US Department of Energy, Contract DE-AC06-76RLO 1830, February 1996. http://www.osti.gov/dublincore/gpo/servlets/purl/212394-mv0A6T/webviewable/; Hrma, P. & Robertus R. J. Waste glass design based on property composition functions, *Ceram. Eng. Sci. Proc.*, 1993, **14** (11/12), 187–203. Hrma, P., Piepel, G. F., Redgate, P. E., Smith, D. E., Schweiger, M. J., Vienna, J. D. & Kim, D. S. Prediction of processing properties for nuclear waste glasses; *Ceram. Trans.*, 1995, **61**, 505–513.
19. Fluegel, A., Varshneya, A. K., Seward, T. P. & Earl, D. A. Viscosity of commercial glasses in the softening range in Proc. Seventh Int. Conf. Advances in Fusion and Processing of Glass III; *Ceram. Trans.*, Volume 141; Eds. J. R. Varner, T. P. Seward & H. Schaeffer, Rochester, New York, USA, 27-31 July 2003, p 379–386; The American Ceramic

Society, Westerville, Ohio, 2004.

20. Fluegel, A., Earl, D. A., Varshneya, A. K. & Öksoy, D. Statistical analysis of viscosity, electrical resistivity, and further glass melt properties, Chapter 9 in: *High temperature glass melt property database for process modelling*, Eds. T. P. Seward III & T. Vascott, The American Ceramic Society, Westerville, Ohio, 2005. http://www.ceramics.org/glassmelt/ http://www.osti.gov/bridge/purl.cover.jsp?purl=/809193-hMVo0M/native/
21. Kucuk, A., Clare, A. G. & Jones, L. An estimation of the surface tension of silicate glass melts at 1400°C using statistical analysis, *Glass Technol.*, Oct 1999, **40** (5), 149–153.
22. Crum, J. V., Schweiger, M. J., Hrma, P. & Vienna, J. D. Liquidus temperature model for Hanford high-level waste glasses with high concentration of zirconia, *Proc. 1996 MRS Fall Meeting*, 2–6 Dec. 1996, Boston, MA, Materials Research Society, Vol. 465, Scientific Basis for Nuclear Waste Management XX, 1997, p 79–85.
23. Hanni, J. B., Pressly, E., Crum, J. V., Minister, K. B. C., Tran, D., Hrma, P. & Vienna, J. D. Liquidus Temperature Measurements for Modeling Oxide Glass Systems Relevant to Nuclear Waste Vitrification, *submitted to Journal of Materials Research.*
24. Choudhary, M. K. & Potter, R. M. Heat Transfer in Glass-Forming Melts, Chapter 9 in: Properties of Glass-Forming Melts, Eds L. D. Pye, A. Montenaro & I. Joseph, CRC Press, Boca Raton, Florida, 2005.
25. Appen, A. A. *Khimiya Stekla (Glass Chemistry)*, Leningrad, 1959. (In Russian).
26. Gan, Fuxi New system of calculation of some physical properties of silicate glasses, *Sci. Sinica*, 1963, **12**, 1365–1391. (In Russian). Gan, Fuxi New system of calculation of properties of inorganic oxide glasses, *Sci. Sinica*, 1974, **17**, 533–551. (In Russian).
27. Demkina, L. I. *Physicochemical Principles of the Manufacture of Optical Glass*, Izd. Chimia, Leningrad, 1976. (In Russian).
28. Priven, A. I. General method for calculating the properties of oxide glasses and glass-forming melts from their composition and temperature, *Glass Technol.*, Dec 2004, **45** (6), 244–254. http://www.ingentaconnect.com/content/sgt/gt/2004/00000045/00000006/art00001; http://www.sciglass.info/Publications/Priven.pdf and: Priven, A. I. Doctoral Thesis, St. Petersburg, 2002. (In Russian).
29. Priven, A. I. & Mazurin O. V. Comparison of methods used for the calculation of density, refractive index and thermal expansion of oxide glasses, *Glass Technol.*, August 2003, **44** (4), 156–166. http://www.ingenta.com/isis/searching/Expand/ingenta?pub=infobike://sgt/gt/2003/00000044/00000004/art00003; http://www.sciglass.info/Publications/PrivenMazurin.pdf
30. Makishima, A. & Mackenzie, J. D. Direct calculation of Young's modulus of glass, *J. Non-Cryst. Solids*, 1973, **12**, 35–45; and Calculation of bulk modulus, shear modulus, and Poisson's ratio of glass, *ibid*, 1975, **17**, 147–157; Calculation of thermal expansion coefficient of glasses, *ibid*, 1976, **22**, 305–313.
31. Bottinga, Y., Weill, D. F. & Richet, P. Thermodynamic modeling of silicate melts in *Thermodynamics of minerals and melts*, Eds R. C. Newton, A. Navrotsky & B. J. Wood, Springer-Verlag, Heidelberg, 1981, p. 207–245.
32. Nemilov, S. V. *Thermodynamic and kinetic aspects of the vitreous state*, CRC Press, Boca Raton, 1995.
33. Pelton, A. D. & Blander, M. Thermodynamic analysis of ordered liquid solutions by a modified quasichemical approach - application to silicate slags, *Metallurg. Trans. B (Process Metallurgy)*, Dec 1986, **17B** (4), 805–815.
34. Pelton, A. D. & Wu, P. Thermodynamic modeling in glass-forming melts, *J. Non-Cryst. Solids*, 1999, **253** (1–3), 178–191. http://dx.doi.org/10.1016/S0022-3093(99)00352-X
35. Chartrand, P. & Pelton, A. D. Modeling the charge compensation effect in silica-rich Na_2O-K_2O-Al_2O_3-SiO_2 melts; *Calphad*, June 1999, **23** (2), 219–230. http://dx.doi.org/10.1016/S0364-5916(99)00026-7
36. Kress: V. C. On the mathematics of associated solutions, *Am. J. Sci.*, Oct 2003, **303**, 708–722. http://earth.geology.yale.edu/~ajs/2003/Oct/08.03.02Kress.pdf
37. Conradt, R. Thermodynamic Approach to viscosity in the glass transition, *Glastech. Ber. Glass Sci. Technol.*, 1994, **67** (11), 304–311.
38. Conradt, R. Modeling of the thermochemical properties of multicomponent oxide melts, *Z. Metall./Mater. Res. Adv. Tech.*, October 2001, **92** (10), 1158–1162.
39. Björkvall, J., Sichen, Du, Stolyarova, V. & Seetharaman S. A model description of the thermochemical properties of multicomponent slags and its application to slag viscosities, *Glass Phys. Chem.*, Mar.–Apr. 2001, **27** (2), 132–147. http://dx.doi.org/10.1023/A:1011332410674
40. Vedishcheva, N. M., Shakhmatkin, B. A. & Shultz, M. M. A simulation of the thermodynamic properties of oxide melts and glasses, *Ceram. Trans.*, Vol. 29, Eds A. K. Varshneya, D. F. Bickford & P. P. Bihuniak, The American Ceramic Society, 1993, p 283–288.
41. Vedishcheva, N. M. & Shakhmatkin, B. A. Thermodynamic studies of oxide glass-forming liquids by the electromotive force method, *J. Non-Cryst. Solids*, July 1994, **171** (1), 1–30.
42. Shakhmatkin, B. A., Vedishcheva, N. M., Shultz, M. M. & Wright, A. C. Thermodynamic properties of oxide glasses and glass-forming liquids and their chemical structure, J. *Non-Cryst. Solids*, 1994, **177** (1), 249–256.
43. Vedishcheva, N. M., Shakhmatkin, B. A., Shultz, M. M. & Wright, A. C. Thermodynamic modelling of glass properties: A practical proposition? *J. Non-Cryst. Solids*, March 1996, **196**, 239–243.
44. Shakhmatkin, B. A., Vedishcheva, N. M. & Wright, A. C. Can thermodynamics relate the properties of melts and glasses to their structure? *J. Non-Cryst. Solids*, November 2001, **293–295**, 220–226. http://dx.doi.org/10.1016/S0022-3093(01)00674-3
45. Shakhmatkin, B. A. Vedishcheva, N. M. & Wright, A. C. Thermodynamic modeling of the structure of glasses and melts: single-component, binary and ternary systems, *J. Non-Cryst. Solids*, November 2001, **293–295**, 312–317. http://dx.doi.org/10.1016/S0022-3093(01)00683-4
46. Schneider, J., Mastelaro, V. R., Zanotto, E. D., Shakhmatkin, B. A., Vedishcheva, N. M., Wright, A. C. & Panepucci, H. Qn distribution in stoichiometric silicate glasses: Thermodynamic calculations and 29Si high resolution NMR measurements, *J. Non-Cryst. Solids*, Sept 2003, **325** (1–3), 164–178. http://dx.doi.org/10.1016/S0022-3093(03)00332-6
47. Besmann, T. M., Spear, K. E. & Beahm, E. C. Thermochemical models for nuclear waste glass subsystems - MgO-CaO and MgO-Al_2O_3, *Materials Research Society Symp. Proc.*, 1999, **556**, 383–389.
48. Spear, K. E., Besmann, T. M. & Beahm, E. C. Thermochemical modeling of glass: Application to high-level nuclear waste glass, *MRS Bull.*, April 1999, **24** (4), 37–44.
49. Besmann, T. M. & Spear, K. E. Thermochemical modeling of oxide glasses, *J. Am. Ceram. Soc.*, Dec 2002, **85** (12), 2887–2894.
50. Besmann, T. M., Spear, K. E. & Vienna, J. D. Extension of the modified associate species thermochemical model for high-level nuclear waste: Inclusion of chromia, *Materials Research Society Symp. Proc.*, 2003, **757**, 195–200.
51. Allendorf, M. D. & Spear, K. E. Thermodynamic analysis of silica refractory corrosion in glass-melting furnaces, *J. Electrochem. Soc.*, Feb 2001, **148** (2), B59–B67. http://dx.doi.org/10.1149/1.1337603
52. Nilson, R. H., Griffiths, S. K., Yang, N., Walsh, P. M., Allendorf, M. D., Bugeat, B., Marin, O., Spear, K. E. & Pecoraro, G. Analytical models for high-temperature corrosion of silica refractories in glass-melting furnaces, *Glass Sci. Technol.*, May/June 2003, **76** (3), 136–151.
53. Strachan, D. M. & Croak, T. L. Compositional effects on the long-term dissolution of borosilicate glass, *J. Non-Cryst. Solids*, 2000, **272**, 22–33.
54. Tischendorf, B. C., Alam, T. M., Cygan R. T., *et al* The structure and properties of binary zinc phosphate glasses studied by molecular dynamics simulations, *J. Non-Cryst. Solids*, Feb 2003, **316** (2–3), 261–272. http://dx.doi.org/10.1016/S0022-3093(02)01795-7
55. Kresse, G. *Ab initio* molecular dynamics: recent progresses and limitations, *J. Non-Cryst. Solids*, Oct 2002, **312–314**, 52–59. http://dx.doi.org/10.1016/S0022-3093(02)01649-6
56. Cormack, A. N. & Du, Jincheng Molecular dynamics simulations of soda-lime-silicate glasses, *J. Non-Cryst. Solids*, Nov 2001, **293–295**, 283–289. http://dx.doi.org/10.1016/S0022-3093(01)00831-6
57. Leko, V. K., Gusakova, N. K., Meshcheryakova, E. V. & Prokhorova, T. I. The effect of impurity alkali oxides, hydroxyl groups, Al_2O_3, and Ga_2O_3 on the viscosity of vitreous silica, *Glass Phys. Chem.*, May–June 1977, **3** (3), 204–219.
58. Dreyfus, C. & Dreyfus, G. A machine learning approach to the estimation of the liquidus temperature of glass-forming oxide blends, *J. Non-Cryst. Solids*, 2003, **318**, 63–78. http://dx.doi.org/10.1016/S0022-3093(02)01859-8
59. Scheffé, H. Experiments with Mixtures, *J. R. Statistical Soc. B*, 1958, **20**, 344–360.
60. Lyon, K. C. Prediction of the Viscosities of Soda-Lime Silica Glasses, *J. Res. Nat. Bur. Standards A, Phys. Chem.*, July–Aug 1974, **78A** (4), 497–504.
61. Kucuk, A. Structure and the physicochemical properties of glasses and glass melts, Doctoral Thesis, Alfred University, New York, 1999.
62. Rao, Q., Piepel, G. F., Hrma, P. & Crum, J. V. Liquidus temperatures of HLW glasses with zirconium-containing primary crystalline phases, *J. Non-Cryst. Solids*, Oct 1997, **220** (1), 17–29. http://dx.doi.org/10.1016/

S0022-3093(97)00227-5

63. Backman, R., Karlsson, K. H., Cable, M. & Pennington, N. P. Model for liquidus temperature of multicomponent silicate glasses, *Phys. Chem. Glasses*, 1997, **38** (3), 103–109.
64. Fluegel, A., Earl, D. A., Varshneya, A. K. & Seward, T. P. Statistical analysis of glass melt properties for high accuracy predictions: Density and thermal expansion of silicate glass melts, Proceedings CD, DGG Meeting, Würzburg, Germany, 23–25 May 2005. http://glassproperties.com/density/
65. Fluegel, A., Varshneya, A. K., Earl, D. A., Seward, T. P. & Oksoy, D. Improved composition-property relations in silicate glasses, part I: viscosity, *Ceram. Trans.*, **170**, Melt Chemistry, Relaxation, and Solidification Kinetics of Glasses - Proceedings of the 106th Annual Meeting of the American Ceramic Society, 2005, p 129–143. http://glassproperties.com/viscosity/
66. Russell, J. K. & Giordano, D. A model for silicate melt viscosity in the system $CaMgSi_2O_6$-$CaAl_2Si_2O_8$-$NaAlSi_3O_8$, *Geochim. Cosmochim. Acta*, 2005, **69** (22), 5333–5349.
67. Karlsson, K. H., Backman, R., Cable, M., Peelen, J. & Hermans, J. Estimation of liquidus temperatures in silicate glasses, *Glass Sci. Technol. (Glastech. Ber.)*, July 2001, **74** (7), p 187–191.
68. Vienna, J. D., Hrma, P., Crum, J. V. & Mika, M. Liquidus temperature-composition model for multi-component glasses in the Fe, Cr, Ni, and Mn spinel primary phase field, *J. Non-Cryst. Solids*, November 2001, **292** (1–3), 1–24.
69. Hrma, P., Smith, D. E., Matyáš, J., Yeager, J. D., Jones, J. V. & Boulos, E. N. Effect of Float Glass Composition on Liquidus Temperature and Devitrification Behavior, *accepted for publication by Journal of Glass Science and Technology.*
70. Redgate, P. E., Piepel, G. F. & Hrma, P. R. Second-Order Model Selection in Mixture Experiments, p 104–110 in 1992 Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association, Alexandria, VA, 1992.
71. Hastie, J. W. & Bonnell, D. W. A predictive phase-equilibrium model for multicomponent oxide mixtures: Part II. Oxides of Na-K-Ca-Mg-Al-Si, *High Temp. Sci.*, June 1984, **19** (3), 275–306.
72. Bishop, C. *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
73. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression & Outlier Detection*, Wiley, 1987.
74. Huber, P. J. *Robust Statistics*, Wiley, 1981.
75. Hunold, K. & Brückner, R. Physikalische Eigneschaften und struktureller Feinbau von Natrium-Alumosilicatgläsern und -schmelzen (Physical properties and structural details of sodium aluminosilicate glasses and melts), *Glastechn. Ber.*, 1980, **53** (6), 149–161.
76. Leko, V. K. & Mazurin, O. V. Analysis of Regularities in Composition Dependence of the Viscosity for Glass-Forming Oxide Melts: II. Viscosity of Ternary Alkali Aluminosilicate Melts, *Glass Phys. Chem.*, 2003, **29** (1), 16–27. http://www.ingentaconnect.com/content/maik/gpac/2003/00000029/00000001/00463582
77. Fluegel, A. Accurate Glass Viscosity Calculation Based on a Global Statistical Modeling Approach, *Proc. 8th Int. Conf. Advances in the Fusion and Processing of Glass*, 12–14 June 2006, Dresden, Germany. http://glassproperties.com/viscosity/
78. Mazurin, O. V. Glass properties: compilation, evaluation, and prediction, *J. Non-Cryst. Solids*, 2005, **351**, 1103–1112.
79. The least-square method was established in 1805 by A. M. Legendre and in 1809 by C. F. Gauss. see e.g.: Placket, R. L. The discovery of the method of least-squares. *Biometrica*, 1972, **59**, 239–251.
80. Piepel, G. F. A Note Comparing Component-Slope, Scheffé, and Cox Parameterizations of the Linear Mixture Experiment Model, *J. Appl. Statistics*, May 2006, **33** (4), 397–403.
81. Mazurin, O. V. & Gankin, Yu. About testing the reliability of glass property data in binary systems, *J. Non-Cryst. Solids*, 2004, **342**, 166–169. http://www.sciglass.info/Publications/MazurinGankin1.pdf
82. Mazurin, O. V. History, prospects, and problems of measurement and calculation of glass properties, *Third Balkan Conference on Glass Science and Technology*, Varna, Bulgaria, 2005. http://glassproperties.com/prospect/ http://www.sciglass.info/Publications/Mazurin1.pdf
83. Belsley, A. D., Kuh, E. & Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York, 1980, p 17.
84. Cook, R. D. Detection of Influential Observation in Linear Regression, *Technometrics*, Feb 1977, **19** (1), 15–18.
85. Cook, R. D. Influential Observations in Linear Regression, *J. Am. Statistical Assoc.*, Mar 1979, **74** (365), 169–174.
86. Fluegel, A. From Data Collection to Data Organization – the First Step for the Scientific Interpretation of Glass Melts, *Proc. 8th ESG Conf.*, Sunderland, UK, 10–14 Sept 2006.
87. Marquardt, D. & W. Snee, R. D. Test Statistics for Mixture Models, *Technometrics*, 1974, **16** (4), 533–537.
88. Casella, G. Leverage and regression through the origin, *Am. Statistician*, 1983, **37** (2), 147–152.
89. Hahn, G. J. *J. Qual. Technol.*, 1977, **9**, 56.
90. Scheffé, H. A method of judging all contrasts in the analysis of variance, *Ann. Math. Stat.*, 1953, **40**, 87–104. Scheffé, H. *The analysis of variance*, Wiley, New York, 1959, p 68.
91. Piepel, G. F. & Cornell, J. A. Mixture Experiment Approaches: Examples, Discussions, and Recommendations, *J. Qual. Technol.*, 1994, **26** (3), 177–196.