DETERMINATION OF THE MOST RELIABLE GLASS PROPERTY VALUES BY THE SCIGLASS INFORMATION SYSTEM

Mazurin Oleg*
Thermex Company, 86 Sredniy Prospect, St-Petersburg, 199106, Russia
mazurin@itcwin.com

Gankin Yuriy
MDL Information Systems, 200 Wheeler Road, Burlington, MA, USA

The latest versions of commercially available glass property databases are powerful tools of predicting glass properties. It is possible to predict properties by using two kinds of models: statistical and structural ones. Most of the known statistical models for prediction of glass properties are polynomials obtained by approximation of a set of data points belonging to a selected composition area. Structural models are algorithms permitting calculation of glass properties from their compositions on the basis of the knowledge or suppositions on a glass structure and chemical interaction between components. A scientist can produce his/her own statistical model for prediction of properties of a given glass composition on the base of selected literature data. As to structural models, a selection of data taken from a database should be used to choose the most reliable model among the existing ones and determine a model error.

(Key words: glass property database, glass property prediction, reliability of property data)

1. Introduction

The importance of determination of glass properties with maximum accuracy is obvious. As is shown in Ref. [1], in the last few decades the number of publications with new experimental data has steadily decreased and the average reliability of those data has also decreased. This means that accurate property predictions for any given glass composition have become increasingly important. Over a long period of time scientists have been developing various methods of glass property calculation (see, for example, Ref. [2]). New opportunities for correct predictions of glass properties have been opened with the advent of universal glass property databases, namely INTERGLAD [3] and SciGlass [4]. Their latest versions contain most of the existing experimental data on glass properties.

At the same time, an analysis of up-to-date scientific literature on glass shows that proper attention to this area of application of glass property databases has not been given by most of glass scientists. In our opinion this makes solution of some problems of glass science and technology considerably less effective.

The main objective of this paper is to attract attention of the glass community to the benefits of this particular usage of glass property databases and present some ideas on the corresponding procedures. It is impossible to cover all the pertinent information in a short presentation. Much more detail can be found in the papers referred to below.

2. General information

At present it is possible to use two main ways of property predictions. One of them is traditional, i.e. using the existing methods of property calculations (see surveys in Ref. [2, 5]). The other deals with development of a statistical model for solving any new problem and using this model to predict properties of a given glass composition or a group of similar compositions.

We will begin with the use of the existing methods of property calculations. In the SciGlass Information System several dozens of the most well known methods of calculations are compiled and can be easily applied. Some of these methods are included in the INTERGLAD database, as well. However, results of property predictions by various methods for the same glass composition may be in many cases quite different. It is important to know which particular method is the most reliable one. Unfortunately, it is not possible to select a method that will be the best for all glass compositions. For

various composition areas various methods of property calculations appear to be the most reliable. Some examples of this are shown in the paper by Priven and Mazurin [5]. Accordingly, model errors (standard deviation "s" of the differences: (experimental values) - (calculated values) = residuals) also depend on selected composition areas.

To select the most reliable method of property prediction for a given glass composition one has to compile the existing data for a composition area to which a given glass or glasses belong and compare different model predictions. Statistical processing of the residuals permits selecting the requisite method.

It is to be noted that the existing methods of property calculations can be divided into two main groups. The methods belonging to the first group may be called statistical models. Usually these models are described by polynomials with coefficients determined by least-squares fitting. The methods belonging to the second group may be called structural-chemical models, or, simply, structural models. The authors of structural models have tried to combine the statistical approach to processing of the existing experimental data with application of their understanding of various aspects of the glassy state. They made allowances for either the direct knowledge, or more or less plausible suppositions on the influence of a composition on a glass structure and interaction of specific components with each other, as well as the influence of these factors on glass properties. From our point of view, among many structural models of this kind there are six that combine reasonable universality with reasonable reliability, namely, the models by Huggins and Sun [6], Appen [7], Demkina [8], Gan Fuxi [9], and Priven [4, 10]. All these models are included in SciGlass.

It is clear that any model, either statistical or structural, is based on processing of a certain amount of experimental data. To develop their models some authors used mainly their own experimental data. There was an obvious advantage in this approach: all data were obtained by the same technique and usually the quality of these data was sufficiently high. At the same time there were serious disadvantages in this approach, as well. The most important one was the requirement of a large amount of experimental data. For example, to develop models by Appen or by Demkina, teams of qualified specialists had to work for about a decade. This was possible in the middle of the 20th century, which is out of question nowadays. At present, development of any more or less universal model for predicting glass properties can be based only on compilation of as many existing data as possible, i.e., it should be based on the data compiled in glass property databases. The advantages of this particular approach were first demonstrated by Priven [10]. This does not exclude the availability of development of new structural or statistical models by using a limited amount of experimental data of particularly high quality. Examples of the efficient use of such sets of data for development of statistical models have been recently given by Fluegel et al. [11, 12]. However, due to a comparatively small number of data points (190-300) such models could have only a limited application inside rather narrow areas of the studied compositions.

At present several structural models enumerated above are much more universal than the existing statistical ones. Therefore, when the most reliable model for calculating a property of glass belonging to a certain composition area is to be chosen one has usually to compare model errors of these structural models. The other possibility to predict a glass property is to develop a specific statistical model, based on a particular set of data points, that is also compiled with the help of a glass property database. The larger the number of data points, the higher is the reliability of the results of such comparison. At present the SciGlass [4] database contains more data than INTERGLAD [3]. Therefore SciGlass has been selected as a basis in this paper.

Probably, the most crucial problem of the approaches to property prediction discussed here is the following. The quality of the measurements reported by various authors differs. Sometimes we can find results obtained by obviously erroneous techniques; sometimes one can encounter obvious misprints. One can wonder whether it is safe to use all the existing experimental data as a base for statistical analysis. In principle, a selection of only the most reliable data or weighted regression depending on the data quality could be considered useful. We think that such approach could be used in some special cases (especially, in the case when in a certain composition area the number of data points is very small). However, from our point of view, in general, such approach could not be recommended. In many cases it is very difficult to compare the quality of different studies by the texts of their publications. The time period of a publication could not be used as an indication of data reliability. Even a publication in a highly ranked scientific journal is not a safe sign of a high quality of

data reported. It seems that there is only one indication of the high enough probability of data precision. It is the fact of application of chemical analysis to determine studied glass compositions. Thus selection of only chemically analyzed glasses by Fluegel et al. [12] for their statistical model could be considered as a reasonable decision. However, in most datasets the percentage of chemically analyzed glasses is too small for development of a statistical model. Besides, there are cases when even results for analyzed compositions prove to be erroneous (see examples in [12] and [13]). Thus, generally, it seems safer to use all the existing data.

Studies of a large number of various kinds of dataset selections have led us to the following conclusions. When a dataset contains more than 20-30 data points taken from more than 4-5 sources of information, the distribution of residuals can be more or less correctly described by a Gaussian curve. Examples of such distributions specific for datasets with large and comparatively small numbers of data points are shown in Fig.1. Some other examples can be found in Ref. [14, 15]. In our opinion, this is a sufficient reason for application of statistical processing of all selected data points. The procedure of it is as follows. After the initial determination of polynomial coefficients on the base of all data, the specific residuals were compared with the model error. All data points with residuals three times higher than the model error were considered as outliers and were not included in determination of a new polynomial. The calculations were repeated until no new outliers were detected.

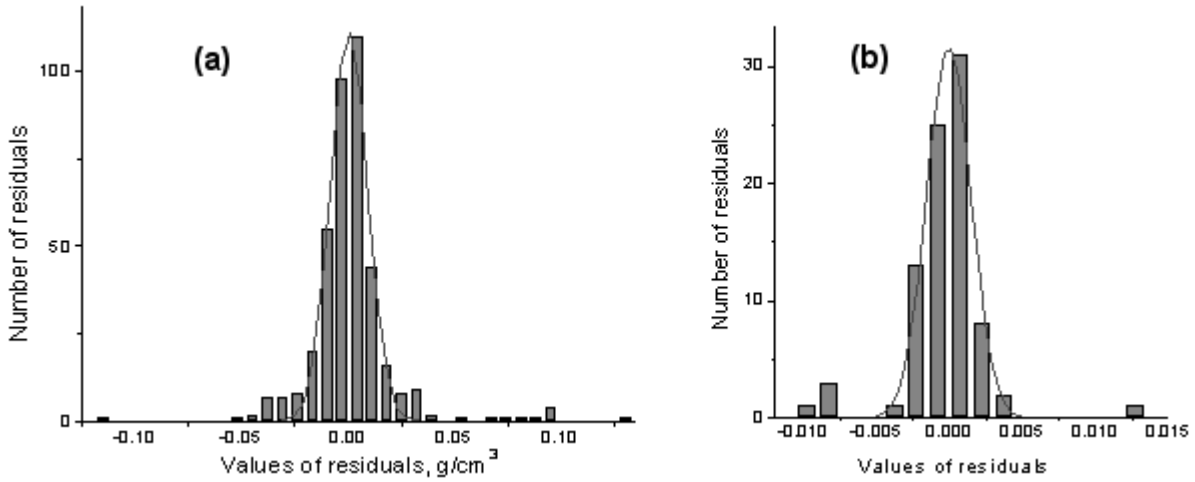

Fig.1.Two examples of residual distribution after approximation by quadratic polynomials. Solid lines represent Gaussian curves. (a) Density of ternary Na_2O-CaO-SiO_2 glasses (396 data points, 11 outliers); s = 0.0138 g/cm^3. (b) Refractive index (n_d) of ternary Na_2O-BaO-SiO_2 glasses (85 data points, 5 outliers); s = 0.0164.

3. Predictions of properties of binary glasses

For the most direct demonstration of some main features of statistical models some binary glass systems seem to be particularly convenient. As an example of the use of such model the composition dependence of the value T_3 (temperature at viscosity equal to 10^3 Poise) is presented in Fig.2. An option of SciGlass Information System was used for statistical processing of the data. Outliers were selected by the procedure described above. Standard deviations s for polynomials of the first, second, and third orders were compared. The value of s for a cubic polynomial appeared to be minimal. This polynomial should be considered as the most reliable. It is to be noted that residuals for the data published by the scientists who are known as the best in this area of studies (such as Lillie [16]), are small in comparison with the model error, which is an additional confirmation of the reliability of such system of property predictions.

It is possible to show [17] that for binary glasses the reliability of the property predictions by statistical models is in most cases better than that obtained by structural models. However, when using statistical models it is well to bear in mind two important limitations. One limitation is quite obvious: a reliable property prediction can be obtained only inside the composition range of experimental data. It is easy to demonstrate this limitation by division of a dataset into two parts shown in Fig. 2. Let us suppose that only compositions within the range from 12 to 24% Na_2O are studied experimentally. If we try to use the polynomials obtained on the base of these data points to predict T_3 values for glasses

with a higher concentration of Na_2O, we will obtain highly erroneous results, as is demonstrated by Fig. 3. If we select the composition range from 12 to 26% Na_2O, the result will also be erroneous, but quite different.

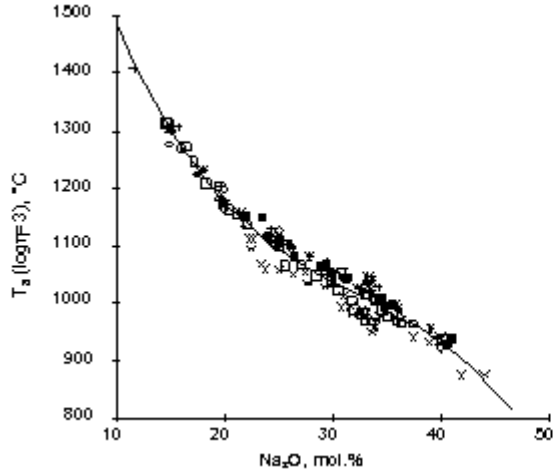

Fig.2. Composition dependence of T_3 (temperature at 10^3 P) for sodium silicate melts. The approximating curve is described by a cubic polynomial.

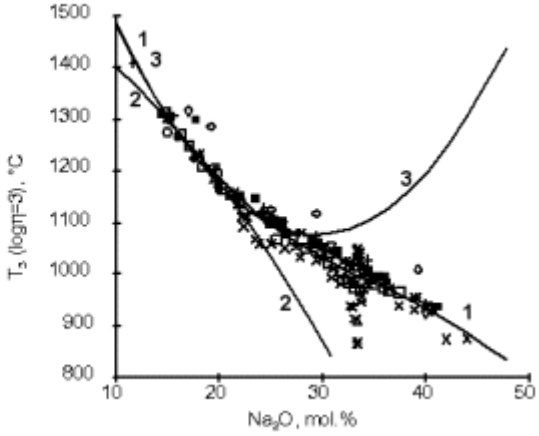

Fig.3. The same as in Fig.2. 1 is the approximating curve for the whole dataset; 2 is the curve for a dataset in the range 12-24% of Na_2O; 3 is the curve for a dataset in the range 12-26% of Na_2O. 2 and 3 – quadratic polynomials.

Fig. 3 also demonstrates another feature of statistical models which one has to be aware of. A random distribution of both more and less correct data within the studied composition area can lead to situations when one or two highly erroneous data points at the borders of this area can lead to considerable deviations of the predicted values from the most probable ones within the border layer of the experimentally determined areas. Thus the danger of committing considerable errors when predicting properties near the border of the dataset area ("border effect") should be always taken into account. The most dependable predictions are usually made in the central part of any dataset. It should be mentioned that, if one determines polynomials on the base of data points belonging to two distant areas (for example, for ranges from 12 to 24% and from 35 to 42% Na_2O), the obtained approximating curve will be practically the same as that shown in Fig.2. Thus the interpolating use of statistical models can be considered as a quite acceptable one.

4. Predictions of properties of ternary and more complex glasses.

All main characteristics of statistical models described and demonstrated in Section 3 for binary systems are the same for more complex systems. The main difference is that in many cases the studied composition areas cover considerably smaller parts of the whole areas of systems and sometimes the composition dependencies of properties within ternary or four-component systems are more complicated than in binary systems. As a result, in some cases structural models appear to be more reliable than statistical ones. If one wants to compare the applicability of structural and statistical models he/she should begin with selecting the most dependable structural model for a chosen dataset.

To demonstrate such kind of selection we will use a density dataset of the system Na_2O-CaO-MgO-Al_2O_3-B_2O_3-SiO_2 where minimal concentrations of CaO and MgO are zero. The total number of data points is 315. Using one of the options in SciGlass we obtain the results presented in Table 1.

It is seen from the table that the composition limits of some structural models do not permit calculating densities for all investigated glasses. Of the three universal (for the given dataset) models the most precise is a model by Priven developed in 1998 [10]. It is reasonable to compare it with the precision of statistical models. To do this the following procedure was used. The dataset similar to the one used for evaluation of structural models was exported from the SciGlass Information System to MS Excel. Then the program "Property Calculator" (developed by the Laboratory of Glass Properties)

for determination of approximating polynomials was used. For linear, quadratic, and cubic polynomials the following model errors were determined after removal of outliers: 0.0404, 0.0299, and 0.0227 g/cm^3 correspondingly. One can conclude from these data that the advantage of the statistical approach is obvious in the described case.

Table 1. Comparison of results of evaluation of structural model errors for densities of a 6-component system

Author of a model	Demkina [8]	Gan Fuxi [9]	Appen [7]	Huggins & Sun [6]	Priven98 [10]	Priven2000 [4]
Number of data points	315	307	181	233	315	315
Model error, g/cm^3	0.0623	0.0786	0.0411	0.110	0.0537	0.0717
Number of data points (*)	306	301	178	229	306	306
Model error, g/cm^3	0.0464	0.0706	0.0331	0.104	0.0378	0.0655

(*) After removal of outliers

At the same time it is necessary to bear in mind that this example demonstrates actually not the possibility of property prediction but only the efficiency of property description for a selected dataset. All data described by polynomials were known in advance. To determine prediction ability it is reasonable to use the cross validation procedure [18]. Let us select a group of data points in the central part of a given dataset. After that we determine approximating polynomials for all data points excluding the mentioned central part of the dataset. Then we use the obtained polynomials to predict properties for data points belonging to the central part of the dataset, determine residuals and calculate model errors for this particular composition area.

For the central part of the dataset we selected 48 compositions with the following contents of two components (in mol.%): $5<Al_2O_3<12$ and $5<B_2O_3<20$. These two components were chosen on the grounds of the dependence of coordination numbers of Al and B on composition and specific influence of this coordination on glass properties. The calculation results are shown in Table 2.

Table 2. Results of predictions of density values for a group of compositions (see details in the text)

Kind of a polynomial	Model error in the main part of a dataset, g/cm^3	Number of data points	Model error in the central part of a dataset, g/cm^3	Number of data points
Linear	0.0420	260	0.0490	48
Quadratic	0.0322	265	0.0377	48
Cubic	0.0236	262	0.0794	48

It is to be noted that the model error of the Priven98 model for 48 data points of the central part of the dataset is equal to 0.0308 g/cm3. Thus in this case to predict glass properties the use of a structural model should be recommended.

In connection with the presented data it is necessary to make some remarks. It is to be pointed out that for this example a glass system containing B_2O_3, i.e., the component with particularly complicated influence on glass properties, was selected. For glass systems that do not contain B_2O_3 or contain only small amounts of this component prediction of glass properties on the base of statistical models is much safer than for boron-containing systems. It should also be taken into consideration that the lower the number of components, the more efficient is the use of statistical models. Note also that in general it is viscosity that is characterized by the most complicated influence of a glass composition. In the case of predicting this property one has to be particularly cautious and, if reliability of prediction is of prime importance it is reasonable to test both kinds of models on the base of all pertinent experimental data that could be found in an applied glass property database. Note also, that, as follows from Table 2, although a cubic polynomial is often considerably more precise than a quadratic one for correct description of the known data, a quadratic polynomial may prove to be safer than a cubic one for predictions of property data.

5. On testing the existing experimental data

As mentioned above, obviously erroneous results can be found in a considerable proportion of recent publications. It is clear that such errors can be revealed only in cases when authors report their data on glass compositions belonging to systems already studied by many other authors. Even in cases when errors can be detected easily (by using glass property databases) these errors are very harmful because

they increase the so-called information noise. However, they are particularly dangerous when they appear in papers concerning glass systems that have not been studied so far or have been studied only by one or two authors. Obviously, to prevent publication of erroneous experimental data, preliminary tests of experimental data are in the common interests of all members of the international glass community, including authors of papers, editors and reviewers of scientific journals, project leaders etc. Therefore, if the outliers found in the course of statistical processing of experimental data (cf. Fig.1) correspond to the data points taken from some of the recently published papers, they should attract particular attention of glass scientists involved in production and publication of such data. We believe it is an appropriate time for authors and reviewers to pay greater attention to inspection of the quality of experimental results before they are published and not afterwards. This problem is considered in greater detail in Ref. [1] and [13].

6. Conclusion

About 15 years ago the release of INTERGLAD opened a new era in the efficient use of the existing information on glass properties. Within a few years after that the SciGlass Information System appeared on the market. Since then great progress in this area has been achieved. In our opinion the use of glass property databases should be a compulsory element of any well-organized glass study. Among various directions of the use of these databases the possibility of reliable predictions of values of glass properties, as well as testing the existing data is very important. Not all problems related to the optimal use of both statistical and structural models for glass property predictions are completely solved at the moment. Nevertheless, even now the proper use of this useful instrument should lead to a decrease in cost and increase in quality of various kinds of projects connected with production and/or use of glass property data.

7. Acknowledgement

The authors are thankful to A.Fluegel and A.Priven for valuable discussions of many problems considered in this paper.

8. References

[1] Mazurin O.V., Streltsina M.V., Shvaiko-Shvaikovskaya T.P., and Mazurina A.O. Glass Phys.Chem., Vol. 28, 384-400(2002).
[2] Volf M.V. Mathematical Approach to Glass, Elsevier, Amsterdam, 1988.
[3] INTERGLAD International Glass Database, Version 5, New Glass Forum, 2001.
[4] MDL®SciGlass-6.0, San Leandro, CA: MDL Information Systems 2003.
[5] Priven A.I and Mazurin O.V., Glass Technol., Vol.44, 156-66(2003).
[6] Huggins M.L.and Sun K.-H., J. Amer. Ceram. Soc., Vol.36, 6-11(1943).
[7] Appen, A.A., Chemistry of Glass, 1970, Khimiya, Leningrad (In Russian).
[8] Demkina L.I. In: Physical and Chemical Bases of the Manufacture of Optical Glass, 1976, Khimiya, Leningrad, 78–116 (In Russian).
[9] Gan Fuxi, Scientia Sinica, Vol.17, 1974, 533-51(1974).
[10] Priven A.I., Glass Phys. Chem., Vol.24, 19-30(1998).
[11] Fluegel A, Varshneya A.K., Seward T.P., and Earl D.A., in: Proceedings 7th International Conference, Advances in Fusion and Processing of Glass III (Ceramic Transactions, Vol. 141; Eds.: J. R. Varner, T. P. Seward, H. Schaeffer), Rochester, New York, 2003.
[12] Fluegel A, Varshneya A.K., Earl D.A., Seward T.P. and Oksoy D., in: Proceedings 106th Annual Meeting & Exposition of the American Ceramic Society, Indianapolis, 2004.
[13] Mazurin O.V. and Gankin Y.V., to be published in J. Non-Cryst. Solids.
[14] Mazurin O.V., Proc. CONCIM conf., Bonn, 2003, to be published in J. Non-Cryst. Solids.
[15] Mazurin O.V., Streltsina M.V., Shvaiko-Shvaikovskaya T.P., and Mazurina A.O. Glass Phys.Chem., Vol. 29, 554-69(2003).
[16] Lillie H.R., J.Am.Ceram.Soc., Vol.22, 367-74(1939).
[17] Mazurin O.V., Streltsina M.V., Shvaiko-Shvaikovskaya T.P., and Leko V.K., to be published in Glass Phys. Chem.
[18] Esbensen K.H. Multivariate Data Analysis, 4th edition. CAMO ASA, 2000.