Estimates for material properties
II. The method of multiple correlations

By D. Bassetti¹, Y. Brechet¹ and M. F. Ashby²

¹Laboratoire de Thermodynamique et Physico-Chimie Métallurgiques,
ENSEEG–1130 rue de la Piscine, Domaine Universitaire,
BP 75–38402, Saint Martin d’Hères Cedex, France

²Engineering Design Centre, Department of Engineering, University of Cambridge,
Trumpington Street, Cambridge CB2 1PZ, UK

Received 7 May 1997; accepted 31 July 1997

The value of a database of material properties is enhanced if the data which it
contains have been subjected to a range of checks and if missing data can, with
stated accuracy, be estimated. Two ways of doing this were described in an earlier
paper. We here extend the range of methods by developing two new, more general,
approaches.

Keywords: material data estimation; material data validation; property correlations

1. Introduction

It is a not-uncommon frustration of applied scientists that, to develop their ideas, or
models or designs, they require data for material properties, and find that the ones
they want are missing. The properties of the prime materials of engineering—steels
and non-ferrous metals, common polymers, certain ceramics and glasses—are well
documented, though even here there can be gaps. The properties of newer or more
exotic candidates are another matter. An engineer seeking a data-profile for many
polymers and for most ceramics will find ‘holes’—missing data, or data so suspect
as to be unusable. A geophysicist or an expert in biomechanics has more severe
problems; here the material-diversity is greater, the difficulties in measurement more
formidable, and gaps in the data are the rule rather than the exception. But in all
these fields, some data exist and are trusted. Could they be exploited to make some
sort of estimates, enough to be going on with, for the missing information?

Our particular interest, material selection for engineering design, is an example of a
data-limited activity. Selection procedures depend, by their nature, on the existence
of complete data, in a uniform format, of known reliability and precision. A database
containing holes will, in a selection exercise, result in the mistaken rejection of poten-
tial candidates; and data of unknown accuracy compromises the selection itself.

A simple approach to these problems was developed in a companion paper (Ashby
1998, hereafter called Paper I). Here we develop a more sophisticated and general
strategy. The ideas are these. First, if the several known properties of a material
closely resemble those of a group of others (its ‘siblings’) then it is plausible, at
least, that its remaining not-known properties might resemble them too. Second, if
reliable data are available for most of the properties of most of the materials in a
database, then those which are missing can be estimated by establishing, by appropriate search routines, correlations between those which are there and combining these in an appropriate manner. The resulting estimate is based not on one correlation but instead makes use of weighted combinations of all of them to extract a best estimate for the missing property. The first idea can be thought of as a sophistication of the range checks of Paper I; the second as an extension and generalization of the correlation-based method described there.

We develop the combined-correlation method first, then describe the proximity (or ‘sibling-likeness’) method, and finally explore an optimum combination of them both.

2. Estimation methods

(a) Correlations

Suppose that one knows the \( p \) properties \( (x_1, x_2, \ldots, x_p) \) of a material \( M \) in a class, \( \varepsilon \), and that one seeks to estimate the properties \( (x_{p+1}, \ldots, x_N) \). We define a correlation coefficient, \( \alpha_{pq} \), between the logarithms \( \log(x_p) \) and \( \log(x_q) \) of a material as follows. Let the function

\[
\log(X_q) = \alpha_{pq} \log(X_p) + \gamma_{pq},
\]

describe the best linear correlation between the data for \( x_q \) and \( x_p \) (and thus describes a power-law relationship between \( x_q \) and \( x_p \)). Then the function,

\[
\phi(\alpha_{pq}, \gamma_{pq}) = \langle (\log(x_q) - \alpha_{pq} \log(x_p) - \gamma_{pq})^2 \rangle,
\]

measures the sum of the squares of the deviations of the data from this line. The minimum of this function is found by setting \( \partial \phi / \partial \alpha = 0 \) and \( \partial \phi / \partial \gamma = 0 \), giving

\[
\alpha_{pq} = \frac{\langle \log(x_p) \log(x_q) \rangle - \langle \log(x_p) \rangle \langle \log(x_q) \rangle}{\langle (\log(x_p))^2 \rangle - \langle \log(x_p) \rangle^2},
\]

and

\[
\gamma_{pq} = \frac{\langle \log(x_q) \rangle \langle (\log(x_p))^2 \rangle - \langle \log(x_p) \rangle \langle \log(x_p) \log(x_q) \rangle}{\langle (\log(x_p))^2 \rangle - \langle \log(x_p) \rangle^2},
\]

where the notation \( \langle \cdot \cdot \cdot \rangle \) means ‘the average over the entire class \( \varepsilon \) of materials’.

The strength of the correlation between two properties \( x_p \) and \( x_q \), described by equation (2.2) is measured by the quantity, \( \beta_{pq} \):

\[
\beta_{pq} = \frac{1}{\phi(\alpha_{pq}, \gamma_{pq})}.
\]

If the correlation approaches perfection, i.e. \( x_p = Ax_q^\alpha \) for all materials, with \( A \) and \( \alpha \) as simple constants, the value of \( \beta \) tends to infinity. The worse the correlation, the lower is the value of \( \beta \).

An estimate for the missing property \( x_q \) of a material \( M \) can now be made via a formula taking account of the correlations \( \alpha_{pq} \) between its known properties and the missing one, \( x_q \). These estimates are weighted by the reliability, \( \beta_{pq} \), of the

† As explained in Paper I, it is generally found that properties are logarithmically distributed (rather than linearly) so that a logarithmic average gives a more realistic estimate.

The method of multiple correlations

\[ \log(x_q) \text{ material property to be estimated} \]

\[ \alpha \text{ neighbour-material} \]

\[ \log(x_i(M)) \]

\[ \log(x_i(M_v)) - \log(x_i(M)) \]

Figure 1. A property \( x_q \) plotted against another, \( x_i \), showing a correlation.

correlation. The strengthening (or attenuating) effect of the weighting is adjusted by raising \( \beta_{pq} \) to a power \( n \):

\[ x_q^{(1)}(M) = \exp \left\{ \sum_{i=1}^{p} \left[ (\beta_{iq})^n (\alpha_{iq} \log(x_i(M)) - \gamma_{iq}) \right] / \sum_{i=1}^{p} (\beta_{iq})^n \right\} \]. (2.6)

The value of \( n \) is subsequently chosen to minimize the error in the estimate, as described below. Note the effect of the \( \beta \)-weighted summation: it is to form an estimate based on all the \( p \) correlations between the properties \( x_p \) and the missing one \( x_q \), weighting each by the function \( \beta_{iq}^n \) of their strengths.

(b) The proximity method

A second independent approach estimates the unknown properties of a material from the properties of its neighbours according to the principle: ‘the nearer a material lies to its neighbours in property space, the greater the chance that it will resemble them’. The key definition is that of ‘distance in property space’.

Consider a material \( M \) in the class \( \varepsilon \) for which \( p \) properties are known. The normalized distance, \( d(M_v) \) of a neighbouring material \( M_v \) from \( M \), in the hyperspace of dimensions \( M \times p \), is defined (figure 1) by

\[ d(M_v) = \left( \sum_{i=1}^{p} \frac{(x_i(M_v) - x_i(M))^2}{x_i(M)} \right)^{1/2}. \] (2.7)

An estimate for a missing property \( x_q \) of \( M \) is found by taking the values of that property for all the neighbours and weighting these by their ‘nearness’, \( 1/d(M_v) \), in property space:

\[ x_q^{(2)}(M) = \frac{\sum_{M_v} x_q(M_v)}{\sum_{M_v} \frac{1}{d(M_v)}}. \] (2.8)

The summations are over the \( M_v \) materials of the class \( \varepsilon \).

This second method has value when there are no good correlations between properties. And, unlike the correlation method, this estimate does not depend on the
number of unknown properties, so it remains usable even when the number of missing properties is large. As with the first method, the weighting can be adjusted by applying a power to the terms, but we have found that the simple expression (2.8) works best.

(c) Coupling of the correlation and the proximity method

The two methods can be coupled, giving an estimate which draws on the best of both. This is achieved by correcting the correlation-based estimate by a weighting based on neighbour distance. The elementary term in the estimation is

$$\log(x_q(M_v)) - \alpha_{iq}[\log(x_i(M_v)) - \log(x_i(M))]$$

(2.9)

This estimate is weighted: first by the reliability of the correlation; and second by the normalized distance of the material from its neighbours. One obtains the final equation for the corrected estimate:

$$x_q^{(3)}(M) = \exp\left\{\sum_{M_v}\left[\frac{1}{d(M_v)}\sum_{M_v}^{}\frac{1}{d(M_v)}\sum_{i=1}^p\left\{\frac{(\beta_{iq})^n}{\sum_{j=1}^p(\beta_{jq})^n}\right\}\times[\log(x_q(M_v)) - \alpha_{iq}[\log(x_i(M_v)) - \log(x_i(M))]]\right\}\right\}.$$  

(2.10)

(d) The uncertainty in the estimates

A measure of the uncertainty or ‘probable error’ can be found by systematically removing each value for a property in a database of class $\varepsilon$, and using each of the three procedures to estimate it, giving $x_q^{(*)}(M)$; the * means the estimate method (1 = correlations, 2 = proximity, 3 = coupling of both). The standard deviation, $\sigma_q^{(*)}$, in the estimate of the property is then calculated from

$$\sigma_q^{(*)} = \sqrt{\frac{1}{n_\varepsilon} \sum_M \left(\frac{x_q^{(*)}(M) - x_q(M)}{x_q(M)}\right)^2},$$

(2.11)

where $n_\varepsilon$ is the number of materials in the class.

3. Implementation and tests of the methods

How good are the estimates? There is more than one question here. The word ‘good’, deconstructed, suggests an exploration of the following.

1. Which properties, within a given class, show strong correlations?
2. How accurately can the methods predict one missing property?
3. How does this accuracy change when several properties are missing?
4. Which method (correlation, proximity, or the combination) gives the most accurate estimates?

We address these in turn. The tests draw on a set of databases which form part of the Cambridge Materials Selector (CMS 1995). A subset of the properties on which the test were conducted is listed in table 1. Some of the databases deal with a broad family of materials: metals and alloys, ceramics, polymers and composites; others deal with specific sub-classes, aluminium alloys, for example. All, as presently constituted, contain no holes.

Table 1. Properties on which the correlation method was tested

<table>
<thead>
<tr>
<th>property</th>
</tr>
</thead>
<tbody>
<tr>
<td>general physical properties</td>
</tr>
<tr>
<td>atomic volume, $V_m$</td>
</tr>
<tr>
<td>density, $\rho$</td>
</tr>
<tr>
<td>mechanical properties</td>
</tr>
<tr>
<td>bulk modulus, $K$</td>
</tr>
<tr>
<td>Young’s modulus, $E$</td>
</tr>
<tr>
<td>shear modulus, $G$</td>
</tr>
<tr>
<td>Poisson’s ratio, $\nu$</td>
</tr>
<tr>
<td>compressive strength, $\sigma_c$</td>
</tr>
<tr>
<td>elastic limit, $\sigma_y$</td>
</tr>
<tr>
<td>modulus of rupture, $\sigma_{\text{MOR}}$</td>
</tr>
<tr>
<td>tensile strength, $\sigma_t$</td>
</tr>
<tr>
<td>endurance limit (fatigue), $\sigma_e$</td>
</tr>
<tr>
<td>hardness, $H$</td>
</tr>
<tr>
<td>ductility (elongation), $\varepsilon_f$</td>
</tr>
<tr>
<td>fracture toughness, $K_{\text{Ic}}$</td>
</tr>
<tr>
<td>loss coefficient, $\eta$</td>
</tr>
<tr>
<td>thermal properties</td>
</tr>
<tr>
<td>melting point, $T_m$</td>
</tr>
<tr>
<td>maximum use temperature, $T_{\text{max}}$</td>
</tr>
<tr>
<td>minimum use temperature, $T_{\text{min}}$</td>
</tr>
<tr>
<td>specific heat, $C_p$</td>
</tr>
<tr>
<td>latent heat of fusion, $L_m$</td>
</tr>
<tr>
<td>thermal conductivity, $\lambda$</td>
</tr>
<tr>
<td>expansion coefficient, $\alpha$</td>
</tr>
<tr>
<td>electrical properties</td>
</tr>
<tr>
<td>electrical resistivity, $R$</td>
</tr>
</tbody>
</table>

(a) Property groups with strong correlations: metals and alloys database

The coefficient $\beta_{pq}$ measures the strength of the correlation between two properties. A large value of $\beta$ ($\beta > 100$) means a strong correlation. The quantity,

$$Q = \frac{\beta_{pq}}{\langle N(\beta_{pq}) \rangle},$$  \hspace{1cm} (3.1)

measures the fractional contribution of one contribution to the overall estimate. Values for $Q$ for the ‘metals and alloys’ database (containing records for 39 materials) are shown in figure 9.

Strong correlations (high values of $Q$) are found within certain property groups. The first group relates to mechanical strength: the compressive strength; the elas-
tic limit; the fatigue limit; the hardness; the modulus of rupture; and the tensile strength. The second group involves elastic and thermal properties. The bulk modulus, shear modulus, Young’s modulus, maximum service temperature, melting point, latent heat of fusion and specific heat all show one or more strong correlations. Other properties, though still showing some correlation, are much less tightly coupled to each other. These include the ductility, the toughness, the damping coefficient, Poisson’s ratio and the expansion coefficient. The electrical resistivity, for metals, is strongly correlated with the thermal conductivity, but with nothing else. This last group of properties poses the greatest problems of estimation.

In some cases (minimum service temperature, loss coefficient) the poor correlation is a direct result of the arbitrary or poor quality of the data in the database; in others (electrical and thermal conductivity) it reflects the great sensitivity of the property to small changes in composition or structure. Figure 2 shows the expected error in the measurement, $\sigma$, plotted against the correlation strength, $\beta$; it has the expected inverse form.

(b) Estimates for one missing property: metals and alloys database

To test the methods, we first removed one property of one material from a database, and used the procedures of the last section to estimate its value, comparing it with the real datum to give a measure of its accuracy. This was repeated for each material in turn, and then for each property.

The databases list a range for each property: $(x_p)_{\text{min}}$ to $(x_p)_{\text{max}}$. In applying the method, the property ranges of the databases were first averaged to give single-point properties by forming geometric means:

$$\bar{x}_p = \sqrt{(x_p)_{\text{min}}(x_p)_{\text{max}}}.$$  

(3.2)

These point-values were then inserted in the equations of the last section to obtain estimates, $x_p^{(*)}$, and standard deviations, $\sigma_p^{(*)}$, for the estimates, where, as before, the $(*)$ indicates the estimation method $(1 = \text{correlations}, 2 = \text{proximity}, 3 = \text{coupling of both})$.

The histograms of figure 3 show the precision with which each property can be estimated. The correlation method works best, followed closely by the coupled correlation–proximity method. The proximity method itself is considerably less accurate. The precision with which unknown values are estimated, however, varies greatly.
The method of multiple correlations

<table>
<thead>
<tr>
<th>Property</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic volume, $V_m$</td>
<td></td>
</tr>
<tr>
<td>Density, $\rho$</td>
<td></td>
</tr>
<tr>
<td>Bulk modulus, $K$</td>
<td></td>
</tr>
<tr>
<td>Compressive strength, $\sigma_c$</td>
<td></td>
</tr>
<tr>
<td>Ductility (elongation) $\varepsilon_f$</td>
<td></td>
</tr>
<tr>
<td>Elastic limit, $\sigma_y$</td>
<td></td>
</tr>
<tr>
<td>Endurance limit (fatigue), $\sigma_e$</td>
<td></td>
</tr>
<tr>
<td>Fracture toughness, $K_{IC}$</td>
<td></td>
</tr>
<tr>
<td>Hardness, $H$</td>
<td></td>
</tr>
<tr>
<td>Loss coefficient, $\eta$</td>
<td></td>
</tr>
<tr>
<td>Modulus of rupture, $\sigma_{MOR}$</td>
<td></td>
</tr>
<tr>
<td>Poisson's ratio, $\nu$</td>
<td></td>
</tr>
<tr>
<td>Shear modulus, $G$</td>
<td></td>
</tr>
<tr>
<td>Tensile strength, $\sigma_t$</td>
<td></td>
</tr>
<tr>
<td>Young's modulus, $E$</td>
<td></td>
</tr>
<tr>
<td>Latent heat of fusion, $L$</td>
<td></td>
</tr>
<tr>
<td>Maximum use temperature, $T_{\text{max}}$</td>
<td></td>
</tr>
<tr>
<td>Melting point, $T_m$</td>
<td></td>
</tr>
<tr>
<td>Minimum use temperature, $T_{\text{min}}$</td>
<td></td>
</tr>
<tr>
<td>Specific heat, $C_p$</td>
<td></td>
</tr>
<tr>
<td>Thermal conductivity, $\lambda$</td>
<td></td>
</tr>
<tr>
<td>Expansion coefficient, $\alpha$</td>
<td></td>
</tr>
<tr>
<td>Electrical resistivity, $R$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Bar chart of the accuracy, $\sigma_p^x$, of the three methods when applied to a database of generic metals and alloys. The correlation method (uppermost bar in each group of three) generally gives the best estimate.

from property to property, so a global average is less interesting than one for a single property.

The elastic properties (Young’s modulus, shear modulus) and the mechanical strength (elastic limit, tensile strength, etc.) are estimated with a useful precision, because the correlations which link them are strong. The estimation of the expansion coefficient is a more severe test: for this broad database for all metals, the error is 35%; for a narrower class, such as light alloys (see below) it becomes much more accurate.

The estimates for the electrical resistivity, thermal conductivity, and above all the loss coefficient illustrate the limitations of the method when applied to a database containing a wide spectrum of materials. These, too, improve when the classes are made more specific.

The reliability of the method when many properties are unknown

The tests of §3b explored the estimation of one property of a material when all its others were known—a favourable hypothesis. In practice, more than one property may be missing. Here we investigate the performance of the algorithms when many properties within the same strongly correlated property group are unknown—a particularly unfavourable scenario. We take the group of mechanical properties (three moduli and three strengths, six in all) as an example. The effect of the number of missing properties on the uncertainty of the estimate is shown in figure 4. The results are obtained by scanning through the ensemble of classes for materials of the family ‘metals and alloys’, considering all possible multiple combinations of missing properties within the group.

The curves reveal the number of known properties which are necessary in order to evaluate the unknown ones in a satisfactory way. If more than four properties in the strongly correlated group of six are missing, the average uncertainty of the estimates exceeds 25% and beyond this rises steeply. Provided at least one of the properties within a strongly correlated group is known, it is possible to make a useful estimate for the others. When all members are unknown, the estimate becomes dependent on remoter and weaker correlations and is consequently much less accurate. This behaviour is characteristic of a strongly correlated group; the weakly correlated groups listed above are less sensitive to the number of unknowns.

Comparison and optimization of the three methods

Experience with the ‘metals and alloys’ database suggests that the correlation method works best, followed closely by the coupled correlation/proximity method. The proximity method itself is considerably less accurate. We anticipate that the proximity method may give more accurate estimates when correlations are weak or many properties are missing.

It was mentioned earlier that the possibility exists of adjusting the weighting factor $\beta$ in equation (2.6) by raising it to a power, $n$, and that a similar thing could be done to the weighting factor $1/d$ in equations (2.8) and (2.10), seeking the value that minimized the estimation error, $\sigma_p^{(n)}$. We find that a power $n = 2$ on the correlation
The method of multiple correlations

Atomic volume, \( V_m \)
Density, \( \rho \)
Bulk modulus, \( K \)
Compressive strength, \( \sigma_c \)
Ductility (elongation) \( \varepsilon_f \)
Elastic limit, \( \sigma_y \)
Endurance limit (fatigue), \( \sigma_e \)
Fracture toughness, \( K_{IC} \)
Hardness, \( H \)
Loss coefficient, \( \eta \)
Modulus of rupture, \( \sigma_{M \ OR} \)
Poisson's ratio, \( \nu \)
Shear modulus, \( G \)
Tensile strength, \( \sigma_t \)
Young's modulus, \( E \)
Maximum use temperature, \( T_{max} \)
Minimum use temperature, \( T_{min} \)
Specific heat, \( C_p \)
Thermal conductivity, \( \lambda \)
Expansion coefficient, \( \alpha \)

Figure 5. Bar chart of the accuracy, \( \sigma_p^{(e)} \), of the three methods when applied to a database of ceramics.

weighting factor \( \beta \), and a power of 1 for the weighting factor \( 1/d \) gave the lowest errors.

(e) Other databases: ceramics

The same tests have been applied to four other databases containing data for other classes of solid. The results are described briefly here and in the following subsections.

The test database of ceramics contains 33 records. The relationship between ceramics properties differs from that for metals in several ways. As examples: the compressive strength differs greatly from the tensile strength and modulus of rupture; the fracture toughness is uncorrelated with the tensile strength, and the thermal and electrical conductivities are uncoupled. The quality of the data in the database, too, differs from that for metals: it is less good. The data in the table are plotted in figure 5, indicating (as before) that, almost always, the correlation and coupling methods give the best results. Many properties are estimated with useful precision, but others carry such large uncertainties as to be of little help. In particular, the

toughness, the loss coefficient and the thermal conductivity are so poorly correlated with other properties that the method breaks down.

(f) Thermoplastics

The database contains 15 thermoplastics. The ability of the method to predict a missing property is illustrated by figure 6. Even with a relatively small number of materials, missing properties are estimated with useful accuracy. Two properties elude the method completely: ductility and the minimum service temperature. The elastic properties—Young’s modulus, bulk modulus and the shear modulus—are less well estimated than in the other classes of material.

(g) Polymer–matrix composites

The database contains 56 composites (some fibre reinforced, most particulate reinforced), based on 15 different polymer matrices, and a variety of reinforcements (car-
The method of multiple correlations

Atomic volume, $V_m$
Density, $\rho$
Bulk modulus, $K$
Compressive strength, $\sigma_c$
Ductility (elongation) $\varepsilon_f$
Elastic limit, $\sigma_y$
Endurance limit (fatigue), $\sigma_e$
Fracture toughness, $K_{IC}$
Hardness, $H$
Loss coefficient, $\eta$
Modulus of rupture, $\sigma_{MOR}$
Poisson's ratio, $\nu$
Shear modulus, $G$
Tensile strength, $\sigma_t$
Young's modulus, $E$
Glass transition temperature, $T_g$
Maximum use temperature, $T_{max}$
Minimum use temperature, $T_{min}$
Specific heat, $C_p$
Thermal conductivity, $\lambda$
Expansion coefficient, $\alpha$

Figure 7. Bar chart of the accuracy, $\sigma_p^{(\ast)}$, of the three methods when applied to a database of polymer–matrix composites.

bon, kevlar, glass fibres and various particulates). The effectiveness of the method is shown in figure 7. The elastic properties are not calculated in a very reliable way; other methods, based on composite theory, allow better estimates for unidirectional composites and laminates. The remaining properties show useful correlations, allowing an estimate of one to be made when the others are known.

(h) Aluminium alloys

This database differs from the others in describing materials of one narrow class: that of aluminium alloys. It contains records for 83 aluminium alloys, both cast and wrought, in various states of heat treatment, allowing a test of the estimation method on a single specialized sub-class of materials. The results are shown in figure 8. Here the coupling of correlation and proximity give the most accurate estimates. The errors are remarkably small. The only property which creates real problems is the ductility, i.e. the elongation-to-fracture.

4. Applications and perspectives

This paper describes a novel method for estimating missing properties for a material for which some properties are known, by exploiting correlations and analogies with its better-characterized siblings. The performance of the method depends, for obvious reasons, on the quality of the ‘known’ data and the strength of the correlations. Its power lies in its ability to combine information from all correlations and analogies, weak and strong, in a manner weighted to maximize the precision of the estimate. It offers, together with the methods of Paper I, a path for the estimation of material properties with calculable standard deviation.

The method described here reaches beyond that of Paper I in its more general applicability. Here we have used it to estimate properties (‘attributes’) of materials (‘objects’) which, because of their underlying dependence on bond type and structure, are inter-dependent. It could be applied to the analysis of any database of
The method of multiple correlations

Figure 9. A map of the correlation strength, $Q$ (expressed as a percentage), for data from the metals database. The number indicates the strength of the correlation between the two properties (identified by the numbers in the left-hand column). The numbers in each row sum to 100.
objects with attributes among which correlations or sibling-relationships are known or suspected.

We acknowledge the financial support of the US Advance Research Project Agency through the University Research Initiative under Office of Naval Research Contract no. N-00014092-J-1808, the Region Rhône-Alpes and the Alliance Program (British Council and French Ministry of Foreign Affairs).

References
