Scaling Laws as a Tool of Materials Informatics

Patricio F. Mendez[1], Reinhard Furrer[2], Ryan Ford[3], Fernando Ordóñez[4]

[1] Assistant Professor, Department of Metallurgical and Materials Engineering, Colorado School of Mines, Golden, CO 80401
[2] Assistant Professor, Departments of Mathematics and Computer Science, Colorado School of Mines, Golden, CO 80401
[3] Graduate Student, Departments of Mathematics and Computer Science, Colorado School of Mines, Golden, CO 80401
[4] Assistant Professor, Department of Industrial and Systems Engineering, University of Southern California, CA 90089

This paper discusses the utility of scaling laws to Materials Informatics and presents algorithm SLAW (Scaling LAW) [1], useful to generate scaling laws from statistical data. These laws can be used to extrapolate known materials property data to untested materials by using other more readily available information. This technique is independent of a characteristic length or time scale, so it is useful for a broad diversity of problems. In some cases, SLAW can reproduce the mathematical expression that would have been obtained through an analytical treatment of the problem. This technique has been originally designed for mining statistical data in materials processing and materials behavior at a system level, and it shows promise for the study of the relationship between structure and properties in materials.

Introduction

Materials Informatics is defined as “the high speed robust acquisition, management, analysis, and dissemination of diverse materials data”[2]. This emerging field is a response to the needs for faster development times for new materials and to the unprecedented amount and complexity of materials information resulting from modern modeling and experimental techniques. In this context, traditional techniques of analysis fall short and new approaches are being developed[3]; the technique presented here is one of them.

Two main thrusts characterize the modeling efforts in Materials Informatics: “hard modeling” and “soft modeling”[4]. Hard modeling encompasses computational strategies involving advanced discretization, parallel algorithms, and a software architecture for distributed computing systems. Among these approaches are atomistic models and ab-initio calculations, thermodynamic modeling, phase field simulation, and finite element modeling at a microstructural level. Soft modeling has been originally introduced by the life sciences and organic chemistry community, and it relates to statistically based, model-independent approaches. Among these approaches are the uses of regressions, neural networks, genetic algorithms, classification algorithms, principal component analysis, partial least squares, and other data mining techniques. SLAW is a soft modeling approach that uses regressions.

One of the goals of Materials Informatics is to make sense of the vast amounts of materials data available. This "knowledge extraction" includes the identification of outliers in the data, the development of models, pattern recognition, and forward and reverse data mapping[5]. Disciplines unrelated to Materials Informatics or its predecessors also faced similar problems, and developed appropriate tools. One set of such tools includes dimensional analysis, scaling, and their related techniques, and has been applied with much success to the study of fluid mechanics and heat transfer problems, but has seen lesser use in the field of Materials in general.

The use of tools related to dimensional analysis is not completely absent from Materials Informatics. For example Cebon and Ashby [6] proposed the use of dimensionless groups for the correlations between materials properties and for data estimation. One of the advantages of using dimensionless groups is that their form is very convenient for inverse mapping. Geller *et al.* [7] proposed using a dimensionless group, the Rice-Thomson parameter, as a heuristic principle in their search of ductilizing elements in Mo alloys. Hard modeling applications also involve dimensional considerations in the formulation of the constitutive equations and their mathematical treatment.

A family of techniques

Scaling laws, similarity, dimensional analysis, and other tools constitute a family of techniques that share in common an attention to the units of variables and parameters, and a reliance on power-law expressions.

Fourier's 1822 treatise on heat transfer[8] is usually considered the first clear appearance of dimensional analysis. It was not until 1914 that Dimensional Analysis was formalized by Buckingham[9] with the renowned "Pi theorem." Currently dimensional analysis is a standard component of the curriculum on fluid mechanics and transport phenomena. In essence, dimensional analysis reduces the number of parameters in a problem by considering their units. For example, many fluid mechanics problems involve four parameters (viscosity, density, length, and, velocity), which can be reduced to a single dimensionless parameter (the Reynolds number) by using this technique. Dimensional analysis can then significantly reduce the number of experiments necessary to characterize a system, and this has been used to much benefit in the fluid mechanics and heat transfer fields.

When a problem involves many parameters, dimensional analysis typically yields many dimensionless groups. Two systems are then considered "similar" when all dimensionless groups have the same value. Complete similarity is very difficult to achieve in physical models; also, as the number of dimensionless groups increases, the benefit of reducing the original number of parameters in the problem decreases. Considering that real problems always involve more parameters than considered formally, complete similarity is impossible in practice. Modeling of materials, whether their processing or the structure-properties relationship involves many more parameters than what is typical of other disciplines such as fluid mechanics; thus, a straight application of dimensional analysis is not expected to simplify the problem significantly. This is consistent with the relatively small use of this technique in materials science and engineering.

"Partial similarity" is of great help in real situations. A system is considered to have "partial similarity" when not all dimensionless parameters are considered. The concept of partial similarity is very useful, and very challenging to implement. If the most relevant parameters are considered, complex problems can often be represented accurately in a simple and insightful form with data collapsing around a single curve. Of course, if partial similarity is based on parameters of secondary importance; the same data becomes very difficult to interpret. Partial similarity analysis can provide accurate results over a wide range of parameters, often spanning several orders of magnitude. The range in which the results are valid relates to the range in which the discarded dimensionless groups are indeed negligible. The subspace of problem parameters in which partial similarity is valid is called a "regime." Within a regime, it is often possible to represent a complex problem in relatively simple terms. At the limit between two regimes, no simple solution is valid. Complex materials problems involving many physical phenomena involve more regimes than simpler problems. Different crystal structures can also act as different regimes, and statistical approaches in this case have been proposed by Le Page[10] and Fischer *et al.*[11]. Coupled problems, when properly treated, can still yield relatively simple solutions in each regime.

The challenge of partial similarity is in the determination of the most relevant set of parameters. Dimensional analysis provides no guidelines regarding the construction or choice of dimensionless groups. Techniques aiming at determining the most relevant dimensionless groups fall roughly within the "hard modeling" and "soft modeling" classification. The hard modeling approach is based on the governing equations resulting from fundamental principles. Prominent among these techniques is "inspectional analysis" or "ordering." Soft modeling approaches are based on statistical data, and involve iterative and artificial intelligence methods. SLAW belongs to this second category.

Inspectional analysis involves the construction of dimensionless groups from the governing equations. This approach was briefly presented by Bridgman[12], made explicit by Ruark[13], and is included in classic textbooks on Transport Phenomena such as Geankopolis[14], Bird, Stewart, and Lightfoot[15], Bejan[16], and Szekely and Themelis[17]. More recently, authors devoted entire chapters or whole books to exploring deeper aspects of this technique, among them Denn[18] and Deen[19] devoted a whole chapter to scaling, Kline[20] devoted a whole book, and Dantzig and Tucker[21] put emphasis on scaling throughout their book on modeling of materials processing. Heuristic approaches were investigated by Sides[22], Chen[23], and Astarita[24], and Yip[25] proposed an artificial intelligence approach. More recent additions to the field are the book by Krantz[26] and the computational implementation of[27].

Statistical approaches to the automatic detection of mathematical laws from data have been an active area of research in the Artificial Intelligence community. Important examples of this research are the algorithm BACON by Bradshaw *et al.*[28], algorithms ABACUS[29] and COPER[30], and recent work by Washio and Motoda[31, 32]. Efforts to use dimensional analysis to reduce the number of adjustable parameters in regressions were pioneered by Li and Lee[33], Dovi *et al.*[34], and Vignaux [35-37]. The limitations of using regressions based on dimensionless groups have been discussed by Hicks[38] and Kenney[39]. SLAW differs from statistical approaches in that the dimensionless groups employed are generated by the algorithm instead of being postulated *a priori*, in that it does not require integer exponents in the scaling laws, in that it allows for datasets in which variables change value simultaneously, and in that it explicitly searches for the simplest predictive formulation.

Dimensional analysis, similarity, and other techniques in this family generate results of the form a power-laws. For this reason it is relevant to discuss them in depth.

Scaling Laws, Power Laws, and Characteristic Values

In the context of SLAW, scaling is a procedure to obtain the characteristic values for the unknown dependent variables. The dependent variables depend on the independent variables and the problem parameters. The independent variables vary within a given system, and are typically the time-space coordinates x, y, z, t, but can also include other magnitudes; for example, temperature in systems with temperature-dependent properties. The parameters are constant for a given system, but vary across systems, such as viscosity in Newtonian fluids. The characteristic value of a variable corresponds to the value unrelated to the independent variables, for example an integral over the domain, or the value at a particular point in the domain. The characteristic values do not vary within a given system, and behave much like the problem parameters. A typical choice of characteristic value is the maximum value in the domain. Within this framework, the problem parameters act as "manifest variables," and combinations of parameters can act as "latent variables." In physics and applied mathematics, scaling approaches occasionally relate the evolution of the dependent variables to the independent variables without using the concept of characteristic values[40, 41].

Scaling approaches typically result in power-law expressions. In the case of SLAW and other engineering approaches, the power law yields an estimation of a characteristic value as a function of the problem parameters raised to constant exponents. If a is a numerical constant, and P, is a parameter, P^a is a scaling law, but a^P is not. If L is a characteristic value of length in the x direction, L^a is the type of scaling law within the scope of this paper, while x^a is not. In this work, scaling laws are the power laws resulting from the scaling process.

Scaling laws are ubiquitous in the sciences. They have been used in physics[42], biology[43-46], human behavior[47, 48], geophysics[49-51], networks and internet traffic [52, 53], economics[54-56] and more. Some reasons for the wide applicability of power law models are: 1) The combination of units has the form of a power law[12], 2) the expressions of many physical phenomena have the form of power laws, 3) the representation of empirical regressions of data in log-log plots often results in a straight line, which corresponds to a power law, and 4) typically, physical magnitudes are distributed exponentially instead of linearly[57, 58]. Scaling laws clearly indicate trends and can yield accurate predictions over several orders of magnitude. They also convey much intuitive meaning: the sensitivity of a power law to a given parameter is directly proportional to the exponent of the parameter; and dimensionless groups can be interpreted as a ratio between dominant forces.

Scaling laws are of enormous utility during the early stages of design, when the freedom in defining a system is largest. These laws are also useful for control systems and for decision-making algorithms, predicting the behavior of a system much faster than computationally intensive models. The design and interpretation of physical models such as surrogate materials, are based on scaling laws. When experimental databases or numerical models exist, scaling laws can be used to generalize and extrapolate the results obtained. Scaling laws provide design rules that can be used immediately by design engineers.

An appealing feature of scaling laws for Materials Informatics is that they provide estimations for a whole family of systems. A scaling law developed for a particular materials system is also applicable to all other materials within the same regime. In this framework, outliers can be readily identified, and they indicate either errors, or physical phenomena that had been disregarded but that should be accounted for the outlier. Another convenient feature of scaling laws is the simplicity of reverse mapping. Because of their mathematical form, scaling laws are specially good for capturing non-linear phenomena.

Algorithm SLAW

SLAW is an algorithm designed to generate power laws from statistical data. The power law is based on the problem parameters (which can be "manifest" or "latent" variables), and estimates the characteristic value of an unknown dependent variable. Its ultimate goal is to generate a mathematical expression that reproduces the data, has predictive capabilities, and can yield insight into the problem being analyzed. SLAW focuses on problems where many parameters are present and a dominant subset of parameters must be identified. To obtain the resulting power law, SLAW performs a sequence of multivariate linear regressions based on the logarithm of the parameters and target quantity. A detailed description of the mathematical implementation of SLAW is included in[1], and a prototype algorithm can be downloaded from[59].

Iterations based on heuristic considerations are at the core of SLAW. The first heuristic consideration is the hypothesis that the target quantity can be captured by a power law. As discussed above, this is generally a good hypothesis. This heuristic also requires that the data being analyzed belong to a single regime. Data spanning more than one regime responds to more than one power law, and needs special considerations. The implementation of SLAW is best explained through a simple example that can be followed step by step, although SLAW is capable of handling much larger data sets.

As an example, consider the joining of a ceramic to a metal discussed in[60] and [1]. Figure 1 shows the geometry of the problem, which consists of two long cylinders; one made of ceramic and the other of metal. These two cylinders are joined at their circular bases at high temperature. The temperature variation between the hot joining temperature and the cooler room temperature causes the ceramic and the metallic cylinder to decrease in size. Typically, the metallic cylinder shrinks more than the ceramic cylinder, causing large stresses in the ceramic near the joint. For failure modes associated to the ceramic, these stresses weaken the joint, so their estimation is essential.

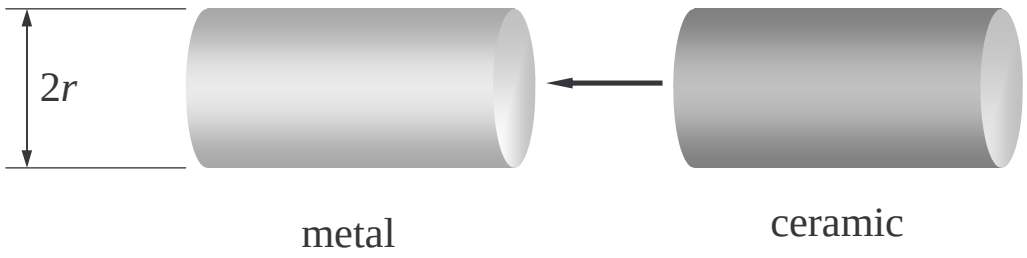

Figure 1: Schematic of ceramic to metal joint used to illustrate the implementation of SLAW

The numerical modeling of this problem was performed assuming an axisymmetric steady state problem at room temperature. The independent variables are r and z (for cylindrical coordinates). The dependent variable is the strain energy $u(r, z)$. The problem parameters are known because they were identified during the construction of the numerical model; they are listed in Table 1. These parameters include all suspected to play a role in the problem in the regime corresponding to the set of observations. Some parameters were not considered, such as strength of the ceramic or elastic limit of the metal. Indeed, there is an unlimited number of parameters that could be included, and careful judgment must be employed in not leaving out a critical parameter for the regime in question. The metric for the stresses in the ceramics is the "total elastic strain energy" U, which is the integral of the volumetric strain energy over the total volume of the ceramic cylinder. This characteristic value does not depend on the independent variables, and it is the target magnitude to be estimated.

Table 1: Parameters involved in the ceramic to metal joining example

Symbol	**Units**	**Description**
U	Pa m^3	total elastic strain energy in the ceramic
E_c	Pa	elastic modulus of the ceramic
E_m	Pa	elastic modulus of the metal
σ_Y	Pa	yield stress of metal
r	m	radius of cylinders
ε_T	-	differential thermal shrinkage

In this example, all relevant parameters are included; thus, consideration of the units (listed in Table 1) can help reduce the problem following the principles of dimensional analysis. Dimensional considerations should not be implemented when the dataset is missing critical parameters; if so, they can lead to erroneous conclusions in the building of dimensionless groups. Most soft modeling approaches do not take advantage of dimensional considerations. The need for information about the units is unique to SLAW. Table 2 summarizes the numerical simulations performed using ABAQUS, and it is all that is necessary to run SLAW. An interesting aspect of this dataset is that the radius r was not varied in the different numerical experiments.

Table 2: Dataset for the ceramic to metal joining example

ceramic	metal	E_c Pa	E_m Pa	σ_Y Pa	r m	ε_T -	U Pa m^3
Si_3N_4	Cu	3.04E+11	1.28E+11	7.58E+07	6.25E-03	6.85E-03	4.23E-03
Si_3N_4	Ni	3.04E+11	2.08E+11	1.48E+08	6.25E-03	5.15E-03	1.52E-02
Si_3N_4	Nb	3.04E+11	1.03E+11	2.40E+08	6.25E-03	2.10E-03	2.80E-02
Si_3N_4	Inco600	3.04E+11	2.06E+11	2.50E+08	6.25E-03	5.15E-03	3.78E-02
Si_3N_4	AISI304	3.04E+11	2.06E+11	2.56E+08	6.25E-03	7.10E-03	3.88E-02
Si_3N_4	AISI316	3.04E+11	1.94E+11	2.90E+08	6.25E-03	7.00E-03	4.91E-02
Al_2O_3	Ti	3.58E+11	1.20E+11	1.72E+08	6.25E-03	5.05E-04	1.04E-02
Al_2O_3	Inco600	3.58E+11	2.06E+11	2.50E+08	6.25E-03	2.95E-03	3.00E-02
Al_2O_3	AISI304	3.58E+11	2.00E+11	2.56E+08	6.25E-03	4.90E-03	3.16E-02

The first step of the computer implementation of SLAW is to perform a linear regression of the data in logarithmic space using all parameters. This corresponds to Iteration 1 in Table 3. This iteration considers all five parameters plus a numerical constant. The exponents of each parameter are listed in the corresponding column. The power law resulting from this iteration is then

$$U_1 = e^{-0.43} E_c^{-1.15} E_m^{0.19} \sigma_Y^{1.78} r^{2.19} \varepsilon_T^{0.13} \quad (1)$$

where U_1 is the power law for estimating U resulting from the first iteration. In this expression, the exponent of r was chosen arbitrarily by the software, for there was insufficient information to determine it. The error considered in Table 3 is the square root of the average of the residuals squared; for small errors, it is equivalent to an average relative error. The evolution of error and number of parameters considered is illustrated in Figure 2. The power law from Iteration 1 has the minimum error; however, it does not result in the correct units, and its exponents do not seem to convey generality.

Table 3: Iterations performed by SLAW in identifying a candidate model for the ceramic to metal joining example

Iteration	#parameters	constant	E_c	E_m	σ_Y	r	ε_T	error
1	5	-0.43	-1.15	0.19	1.78	2.19	0.13	2.6%
2	5	-1.05	-0.95	0.17	1.78	3.00	0.14	2.8%
3	4	-0.64	-0.82	0.00	1.82	3.00	0.18	4.0%
4	4	0.00	-0.90	0.00	1.90	3.00	0.19	5.2%
5	3	0.00	-1.04	0.00	2.04	3.00	0.00	16.9%
6	2	0.00	0.00	0.00	1.00	3.00	0.00	772%

Iteration 2 still considers all five parameters, but adds the constraint of dimensional homogeneity. Because of this additional constraint, the error of this iteration is slightly larger than for the previous one (2.8% vs. 2.6%). This iteration illustrates a remarkable aspect of SLAW, which is the determination of the exponent of r as *exactly* 3, despite that r was not included in the sensitivity study. This determination is possible because this value of the exponent is the only possibility for satisfying dimensional homogeneity.

Iteration 3 is the first step of SLAW towards simplification of the power law, and it considers only four of the five parameters. All possible eliminations of one parameter were considered, and the elimination of E_m was the one that involved the smallest error (4%). This approach of eliminating one parameter at a time is called “backward elimination.”

Iteration 4 also explored all possible eliminations of one of the remaining parameters or the constant. Because this iteration discarded the numerical constant the model still involves four parameters. The elimination of the numerical constant is controversial, and future implementations of SLAW might no perform it. While a mathematical expression without a constant is simpler, keeping the constant might enable the use of better error metrics. Iteration 5 eliminates ε_T and considers only three parameters with a relatively small error of 16.9%.

How far can the power law be simplified at a small cost in accuracy? Iteration 6 provides the answer. This iteration considers only two parameters, but now the error jumped to 772%. Clearly, this simplification should not be adopted. This significant jump in error can be used as a heuristic for determining when to stop the backward elimination process.

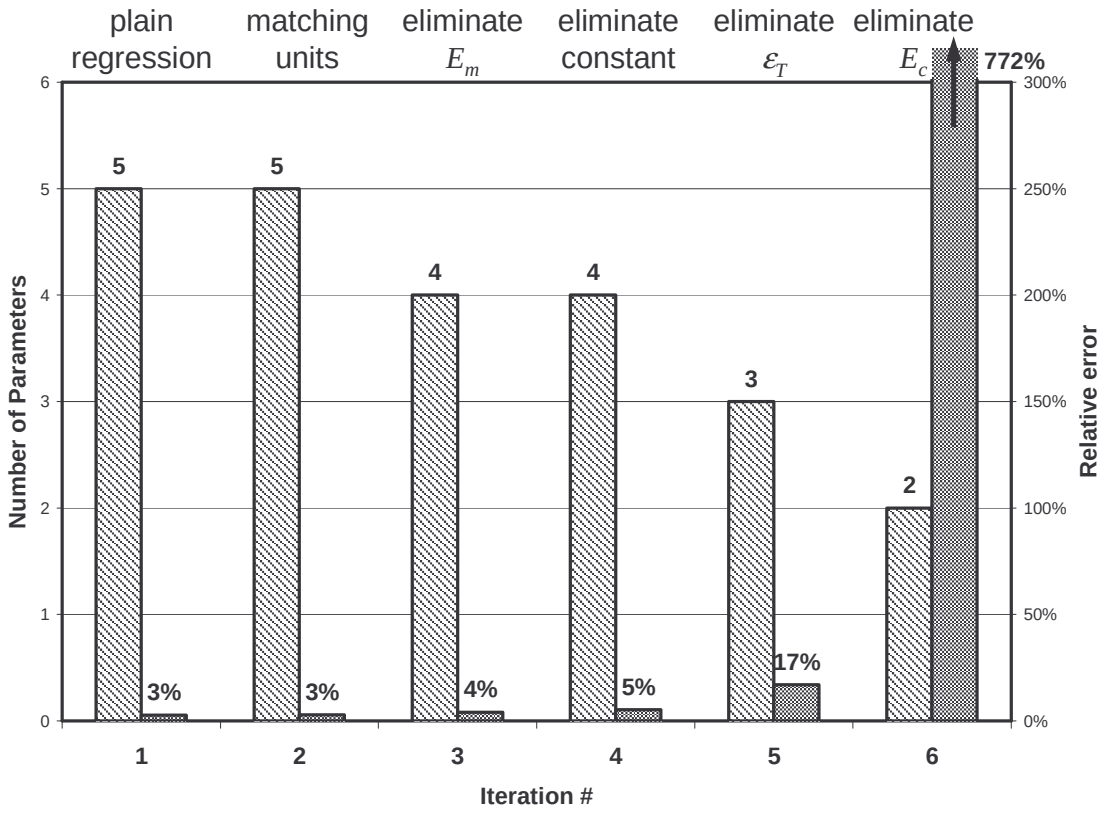

Figure 2: Evolution of number of parameters considered and corresponding error for backwards elimination in the ceramic to metal joining example

The sudden jump in error is not a coincidence of this example, but something to be expected. As discussed above, if all the observations corresponded to the same regime, the system can typically be represented by a power law based on the dominant parameters. As long as those dominant parameters are included in the iterations, the error associated to the power law should be small. This error will be associated to measurement or calculation errors and to the small effect of secondary parameters. After all secondary parameters have been eliminated, the next iteration will eliminate a dominant parameter, making the estimation meaningless, and significantly increasing the error. In essence, SLAW follows the maxim often attributed to Einstein that "everything should be made as simple as possible, but not simpler." Taking advantage of this sudden increase in error is a feature unique to SLAW. When a dataset spans more than one regime, the error increase is not as sharp, and the determination of the point where to stop the backward elimination process becomes uncertain. The power law obtained from Iteration 5 is the simplest accurate estimation, and its expression is

$$U_5 = E_c^{-1.04} \sigma_Y^{2.04} r^3 \tag{2}$$

The exponents of this expression are very close to round numbers, suggesting that the obtained power law is very close to what would have been obtained through an analytical approach to the problem. Typically in this case, researchers assume that the exponent should actually be a round number, and that the difference is due to statistical error, a recent example of this approach is in [61]. The rounded expression of Iteration 5 has an error of 37% and has the following expression:

$$U_{5\text{rounded}} = E_c^{-1} \sigma_Y^{2} r^3 \tag{3}$$

This is the same expression obtained by mathematical analysis of the problem in [60], and has an intuitive interpretation: the strain energy in the ceramic is due to the elastic loading induced by the plastic deformation of the metal at its yield stress. It makes sense, then, that the elastic modulus of the metal or the amount of thermal strain are unimportant in this case. Although SLAW performed its iterations blindly, the final result is insightful, in addition of predictive.

The error of the rounded expression can be reduced by considering a ranked set of dimensionless groups. Algorithm SLAW can be run again, but this time using $U/U_{5\text{rounded}}$ as the target variable. If all parameters and a constant are considered in this case, the error is the same of Iteration 1, and if no parameters or constant are considered, the error is the same as that of the rounded expression. Each elimination of a parameter in this second round of SLAW defines a dimensionless group. Because the effect on error of the first dimensionless groups is small, these groups are secondary. The dimensionless groups eliminated in the later stages convey larger errors and are more relevant. In this case, the most relevant group is the numerical constant, yielding the final expression

$$U_{\text{estimated}} = e^{0.33} E_c^{-1} \sigma_Y^{2} r^{3} \tag{4}$$

Figure 3 illustrates a comparison between Equation 4 and the values calculated using finite elements. The good correspondence is remarkable, considering the broad range of metals and ceramic properties involved.

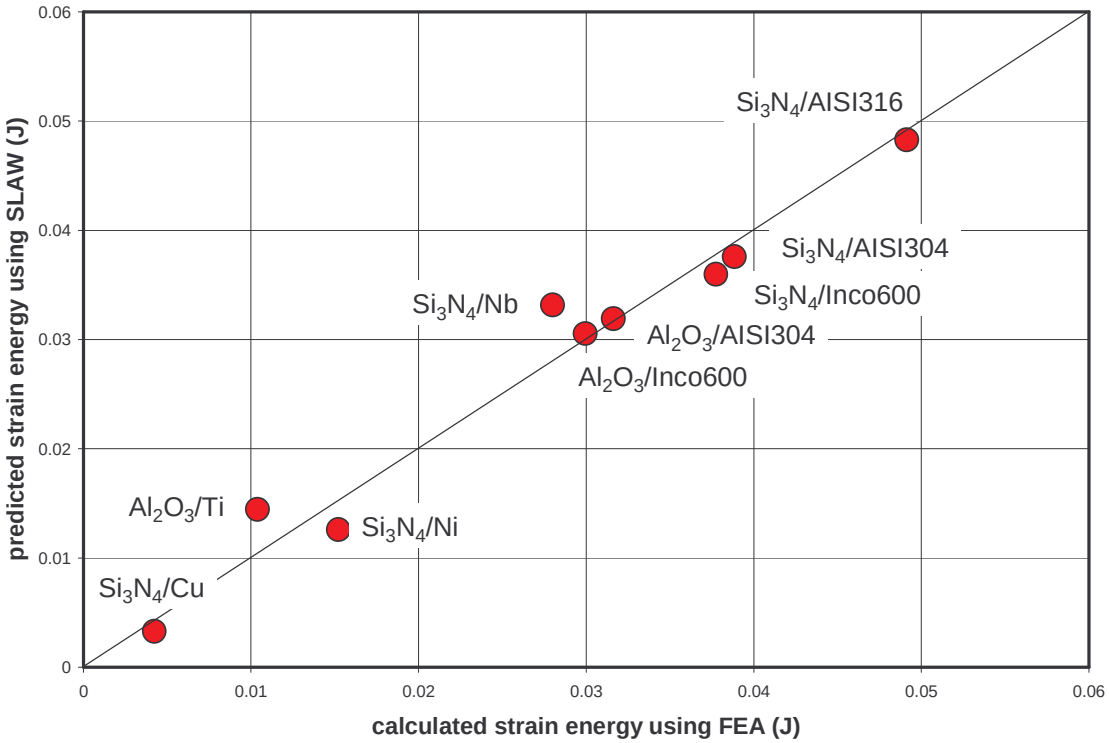

Figure 3: Comparison between calculated strain energy using finite element analysis (horizontal axis) and estimated values using Equation 4 (vertical axis). The points represent the ceramic to metal joints listed in Table 2.

Discussion

The example presented here, and additional examples presented in[1] justifies the heuristic reasoning embodied in SLAW. The systematic reduction of degrees of freedom in SLAW also helps avoid the problem of "overfitting" in which the regressions reproduce the input data, but have a poor performance with new data[4, 62, 63].

SLAW shares some similarities with Principal Component Analysis (PCA)[64] in its goal of discriminating dominant variables from secondary ones. A significant difference between SLAW and PCA is that the former simplifies the system by eliminating *physical* parameters, while in PCA each component can potentially involve all parameters. For this reason, the simplifications from SLAW are not orthogonal in the sense of PCA. The advantage of the approach followed by SLAW is that by eliminating physical parameters, less information is necessary to make estimation. Once the dominant set of parameters has been established, estimations can be performed with very coarse knowledge about the secondary parameters, enabling the use of incomplete datasets. The use of incomplete datasets is also a goal of other algorithms such as classification and regression trees[63]. A minimum of knowledge of the secondary parameters is still necessary in SLAW to ensure that the intended estimations correspond to the proper regime.

Because SLAW is based on constrained regressions, special attention must be paid to proper optimization issues and statistical issues. From the point of view of optimization, the backward elimination procedure does not generate all possible linear regressions, since it sequentially removes one parameter at a time. The procedure might actually not obtain the linear regression with smallest error; it is possible that a parameter that was eliminated in a previous iteration becomes dominant as other parameters are eliminated. Alternative methods of scanning could help overcome this potential problem. For instance, we could conduct an exhaustive search by considering all possible regressions, or use forward selection or stepwise regression. In this case, each iteration would see a decreased error, with a sudden decrease when all dominant parameters are considered for the first time. One advantage of this approach is that once an acceptable estimation is obtained, it is not necessary to keep iterating until all parameters are considered. We are currently exploring this type of exhaustive algorithms.

From the point of view of statistics, relevant considerations involve the choice of a metric for the quality of the estimations, the selection of a particular iteration to provide a model, and the rounding of the exponents. The metric currently used by SLAW relates to the average relative error of the proposed power law. This metric, however, can be misleading when the dataset involves small variations of some parameters. There are indications that R^2 or related statistical quantities could be more meaningful metrics. In this case, the elimination of the numerical constant should not be allowed at any iteration. The selection of a particular iteration to provide the chosen power law is more complex. For problems with small error of measurement or calculation and within a single regime, a clear jump in the error makes the selection fairly obvious. However, when a clear jump in the error is not evident, it cannot be discriminated whether the original data was of poor quality (measurement error) or if the measurements spanned multiple regimes. Of course, there is also a possibility that the heuristic of imposing a power law form is not valid. There is no reliable way yet of discriminating which of these possibilities might be occurring. The rounding of the exponents is also related to statistical error; currently, SLAW rounds the exponents to the nearest third or quarter, regardless the quality of the input data. Fortunately, analytical solutions seldom involve exponents with fractions finer than quarters or thirds. It should be possible, however, to optimize the coarsening of the rounding of the exponents based on statistical considerations. Another relevant statistical consideration is the use of regressions in a logarithmic space. One of the requirements of regressions is that the error be distributed symmetrically around the mean. This is not a problem if the symmetry occurs in the logarithmic space, which seems to be the case for most physical magnitudes[57, 58], or if the errors are small compared to the mean.

SLAW takes advantage of dimensional homogeneity to reduce the number of degrees of freedom in the regressions. Dimensional homogeneity requires that all relevant parameters be included, even those that are constant throughout the whole dataset. Dimensional analysis also has the same stringent requirement. When a "hard model" is available, it can be used to determine the parameters that should be included in the dataset. When there is uncertainty about having *all* relevant parameters included in the dataset, it is also possible to employ SLAW without concern for units. In this case, the reduction of degrees of freedom based on dimensional considerations is not possible, and the estimations obtained are valid only as long as the unknown dominant parameters remain unchanged.

The current implementation of SLAW requires that all observations in the dataset belong to a single regime. This requirement can be difficult to fulfill. A similar approach in which multiple scaling laws are pursued simultaneously might relax this requirement to systems spanning multiple regimes; such an approach has been pursued in the field of artificial intelligence by Kokar[65].

Summary

Scaling laws have a proven track record of success in many scientific disciplines, and show much promise in the area of Materials Informatics. Algorithm SLAW based on linear regressions constrained by dimensional homogeneity can provide scaling laws that are accurate, simple, and insightful. In some cases, this approach can anticipate the results from analytical modeling. SLAW's features have been valuable for the study of materials at a system level, and it could be an important tool in the study of the structure-property relationship in materials informatics.

References

1. Mendez, P.F. and Ordonez, F., *Scaling Laws from Statistical Data and Dimensional Analysis.* Journal of Applied Mechanics, 2005. **72**(5): p. 648-657.

2. Hunt, W.H., *Materials Informatics: Growing from the Bio World.* JOM, 2006. **58**(7): p. 88-88.
3. Ferris, K.F., Peurrung, L.M., and Marder, J., *Materials Informatics: Fast Track to New Materials.* Advanced Materials & Processes, 2007. **165**(1): p. 50-51.
4. Liu, Z.K., Chen, L.Q., and Rajan, K., *Linking Length Scales Via Materials Informatics.* Jom, 2006. **58**(11): p. 42-50.
5. Hull, C.W., *Method for Production of Three-Dimensional Objects by Stereolithography*. 1990, 3D Systems, Inc: United States.
6. Cebon, D. and Ashby, M.F., *Engineering Materials Informatics.* MRS Bulletin, 2006. **31**(12): p. 1004-1012.
7. Geller, C.B., et al., *A Computational Search for Ductilizing Additives to Mo.* Scripta Materialia, 2005. **52**(3): p. 205-210.
8. Fourier, J.B., *Théorie Analytique De La Chaleur*. 1822.
9. Buckingham, E., *On Physically Similar Systems; Illustrations of the Use of Dimensional Equations.* Physics Review, 1914. **4**(4): p. 345-376.
10. Le Page, Y., *Data Mining in and around Crystal Structure Databases.* MRS Bulletin, 2006. **31**: p. 991-994.
11. Fischer, C.C., et al., *Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics.* Nature Materials, 2006. **5** (8): p. 641-646.
12. Bridgman, P.W., *Dimensional Analysis*. First ed. 1922, New Haven: Yale University Press. 113.
13. Ruark, A.E., *Inspectional Analysis: A Method Which Supplements Dimensional Analysis.* Journal of the Mitchell Society, 1935. **51**: p. 127-133.
14. Geankoplis, C.J., *Transport Processes and Separation Process Principles: (Includes Unit Operations)*. 4th ed. 2003, Upper Saddle River, NJ: Prentice Hall Professional Technical Reference. xiii, 1026 p.
15. Bird, B.R., Stewart, W.E., and Lightfoot, E.N., *Transport Phenomena*. First ed. 1960: John Wiley & Sons. 780.
16. Bejan, A., *Convection Heat Transfer*. 3rd ed. 2004, Hoboken, N.J.: Wiley. xxxi, 694 p.
17. Szekely, J. and Themelis, N.J., *Chapter 16: Similarity Criteria and Dimensional Analysis*, in *Rate Phenomena in Process Metallurgy*. 1971, John Wiley & Sons. p. 557-597.
18. Denn, M.M., *Process Fluid Mechanics*. First ed. Prentice-Hall International Series in the Physical and Chemical Engineering Series, ed. Amundson, N.R. 1980, Englewood Cliffs, NJ: Prentice-Hall. 383.
19. Deen, W.M., *Analysis of Transport Phenomena*. 1998, New York: Oxford University Press. xix, 597 p.
20. Kline, S.J., *Similitude and Approximation Theory*. 1986, New York: Springer-Verlag. xix, 229.
21. Dantzig, J.A. and Tucker, C.L., *Modeling in Materials Processing*. 2001, Cambridge, England; New York: Cambridge University Press. xiii, 363 p.
22. Sides, P.J., *Scaling of Differential Equations. "Analysis of the Fourth Kind"*. Chemical Engineering Education, 2002(summer): p. 232-235.
23. Chen, M.M., *Scales, Similitude, and Asymptotic Considerations in Convective Heat Transfer*, in *Annual Review of Heat Transfer*, Tien, C.L., Editor. 1990, Hemisphere Pub. Corp.: New York. p. 233-291.
24. Astarita, G., *Dimensional Analysis, Scaling, and Orders of Magnitude.* Chemical Engineering Science, 1997. **52**(24): p. 4681-4698.

25. Yip, K.M.K., *Model Simplification by Asymptotic Order of Magnitude Reasoning.* Artificial Intelligence, 1996. **80**(2): p. 309-348.
26. Krantz, W.B., *Scaling Analysis in Modeling Transport and Reaction Processes: A Systematic Approach to Model Building and the Art of Approximation*. 2007, Hoboken, N.J.: John Wiley & Sons.
27. Mendez, P.F. *Advanced Scaling Techniques for the Modeling of Materials Processing*. in *Sohn Symposium*. 2006. San Diego, CA.
28. Bradshaw, G., Langley, P., and Simon, H.A. *Bacon 4: The Discovery of Intrinsic Properties*. in *Third Nat. Conf. of the Canadian Society for Computational Studies of Intelligence*. 1980. Victoria, BC, Canada.
29. Washio, T. and Motoda, H. *Extension of Dimensional Analysis for Scale-Types and Its Application to Discovery of Admissible Models of Complex Processes*. in *2nd International Workshop on Similarity Method*. 1999.
30. Kokar, M.M., *Determining Arguments of Invariant Functional Descriptions.* Machine Learning, 1986(1): p. 403-422.
31. Washio, T. and Motoda, H. *Extension of Dimensional Analysis for Scale-Types and Its Application to Discovery of Admissible Models of Complex Processes*. in *2nd Int. Workshop on Similarity Method*. 1999.
32. Washio, T., Motoda, M., and Niwa, Y. *Enhancing the Plausibility of Law Equation Discovery*. 2000.
33. Li, C.C. and Lee, Y.C., *A Statistical Procedure for Model-Building in Dimensional Analysis.* International Journal of Heat and Mass Transfer, 1990. **33**(7): p. 1566-1567.
34. Dovi, V.G., et al., *Improving the Statistical Accuracy of Dimensional Analysis Correlations for Precise Coefficient Estimation and Optimal- Design of Experiments.* International Communications in Heat and Mass Transfer, 1991. **18**(4): p. 581-590.
35. Vignaux, G.A., *Dimensional Analysis in Operations-Research.* New Zealand Operational Research, 1986. **14**(1): p. 81-92.
36. Vignaux, G.A. and Scott, J.L., *Simplifying Regression Models Using Dimensional Analysis.* Australian \& New Zealand Journal of Statistics, 1999. **41**(1): p. 31-41.
37. Vignaux, G.A. *Some Examples of Dimensional Analysis in Operations Research and Statistics*. in *4th International Workshop on Similarity Methods*. 2001. Stuttgart, Germany: University of Stuttgart.
38. Hicks, B.B., *Some Limitations of Dimensional Analysis and Power Laws.* Boundary-Layer Meteorology, 1978. **14**: p. 567-569.
39. Kenney, B.C., *On the Validity of Empirical Power Laws.* Stochastic Hydrology and Hydraulics, 1993. **7**: p. 179-194.
40. Barenblatt, G.I., *Scaling, Self-Similarity, and Intermediate Asymptotics*. 1st ed. Cambridge Texts in Applied Mathematics. 1996, New York: Cambridge University Press. xv, 386.
41. Barenblatt, G.I., *Scaling*. Cambridge Texts in Applied Mathematics. 2003, Cambridge: Cambridge University Press. xiv, 171 p.
42. Taylor, M., et al., *100 Years of Dimensional Analysis: New Steps toward Empirical Law Deduction.* Submitted to New Journal of Physics (IOP), 2007. **arXiv:0709.3584v3 [physics.class-ph]**.
43. Stahl, W.R., *Dimensional Analysis in Mathematical Biology I. General Discussion.* Bulletin of Mathematical Biology (Springer), 1961. **23**(4): p. 355-376.
44. Stahl, W.R., *Dimensional Analysis in Mathematical Biology. Ii.* Bulletin of Mathematical Biology (Springer), 1962. **24**(1): p. 81-108.

45. Kokshenev, V.B., *Observation of Mammalian Similarity through Allometric Scaling Laws.* Physica a-Statistical Mechanics and Its Applications, 2003. **322**(1-4): p. 491-505.
46. Azad, R.K., et al., *Segmentation of Genomic Dna through Entropic Divergence: Power Laws and Scaling.* Physical Review E, 2002. **65**(5): p. art. no.-051909.
47. Nakamura, T., et al., *Universal Scaling Law in Human Behavioral Organization.* Phys. Rev. Lett., 2007. **99**: p. 138103.
48. Brockmann, D., Hufnagel, L., and Geisel, T., *The Scaling Laws of Human Travel.* Nature, 2006. **439**: p. 462-465.
49. FAUG, T., et al., *Varying Dam Height to Shorten the Run-out of Dense Avalanche Flows: Developing a Scaling Law from Laboratory Experiments.* Surveys in Geophysics, 2003. **24**: p. 555-568.
50. Arunachalam, V.M. and Muggeridge, D.B., *Ice Pressures on Vertical and Sloping Structures through Dimensional Analysis and Similarity Theory.* Cold Regions Science and Technology, 2003. **21**(3): p. 231-245.
51. Housen, K.R., Schmidt, R.M., and Holsapple, K.A., *Crater Ejecta Scaling Laws - Fundamental Forms Based on Dimensional Analysis.* Journal of Geophysical Research, 1983. **88**(B3): p. 2485-2499.
52. Barabasi, A.-L. and Albert, R., *Emergence of Scaling in Random Networks.* Science. **286**: p. 509-5012.
53. Carlson, J.M. and Doyle, J., *Power Laws, Highly Optimized Tolerance and Generalized Source Coding.* Physical Review Letters, 2000. **84**(24): p. 56-59.
54. Jong, F.J.d. and Quade, W., *Dimensional Analysis for Economists*. Contributions to Economic Analysis; 50. 1967, Amsterdam: North Holland Pub. Co. xiii, 223.
55. Chave, J. and Levin, S., *Scale and Scaling in Ecological and Economic Systems.* Environmental and Resource Economics, 2003. **26**: p. 527-557.
56. Xu, Z. and Gencay, R., *Scaling, Self-Similarity and Multifractality in Fx Markets.* Physica A, 2003. **323**: p. 578 - 590.
57. Newcomb, S., *Note on the Frequency of Use of the Different Digits in Natural Numbers.* American Journal of Mathematics, 1881. **4**: p. 39-40.
58. Benford, F., *The Law of Anomalous Numbers.* Proceedings of the American Philosophical Society, 1938. **78**(4): p. 551-572.
59. *Scaling Laws. Slaw Homepage*, http://illposed.usc.edu/~pat/SLAW.
60. Park, J.-W., Mendez, P.F., and Eagar, T.W., *Strain Energy Distribution in Ceramic to Metal Joints.* Acta Materialia, 2002. **50**: p. 883-899.
61. Huang, J., et al., *Capillary Wrinkling of Floating Thin Polymer Films.* Science, 2007. **317**: p. 650.
62. Mazzatorta, P., et al., *The Importance of Scaling in Data Mining for Toxicity Prediction.* Journal of Chemical Information and Computer Sciences, 2002. **42**(5): p. 1250-1255.
63. Li, Y., *Predicting Materials Properties and Behavior Using Classification and Regression Trees.* Materials Science And Engineering A-Structural Materials Properties Microstructure And Processing, 2006. **433**(1-2): p. 261-268.
64. Rajan, K., *Materials Informatics.* Materials Today, 2005(October): p. 38-45.
65. Kokar, M.M., *A Procedure of Identification of Laws in Empirical Sciences.* Systems Science, 1981. **7**(1).