Statistical approach for automated weighting of the experimental data and outlier detection

Setareh Zomorodpoosh[1], Irina Roslyakova[1], Abdulmonem Obaied[1], Richard Otis[2], Brandon Bocklund[3], Ingo Steinbach[1]

1 - ICAMS, Ruhr-University Bochum, Universitätstr. 150, 44801, Bochum, Germany

2 - Engineering and Science Directorate, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

3 - Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA

E-mail: setareh.zomorodpoosh@rub.de

Traditionally every CALPHAD assessment starts with collection of available experimental data and their critical review. Then different weights will be assigned to considered data sets for parameters estimation in order to achieve the best possible agreement with the experimental information. Therefore, weighting of data sets involved in calculations becomes one of the critical and most important steps during thermodynamic calculation using CALPHAD method. Usually, such assignment of weights for experimental / DFT data is done manually based on knowledge and experience of researcher performing calculations. Therefore, in order to avoid human factors, we propose an automated assessment criterion to determine the weight of data sets using data mining methods. Our strategy for automated weighting of experimental data sets is based on the k-fold cross validation method [1], modified under the condition that each data set contains unequal number of observations, which is a typical situation during CALPHAD assessment.

Application of this approach allows us to determine the importance of each data sets involved in assessment and show the impact of weighting in statistical analysis results of each model. Our goal is to find such weights that will increase an accuracy of applied model by minimization of the residual standard error (RSE). To demonstrate the proposed approach, we applied it for the fitting of heat capacity data of pure elements such as Al, Cr, Fe, Ni, and Mg using segmented regression [2] and Chen-Sundman models [3]. Currently, these models are intensively used for the development of the third generation CALPHAD databases. Additionally, to the newly proposed procedure of automated weighting, we developed the method for outlier detection,

which is based on combination of data mining tools and thermo-physical properties such as enthalpy.

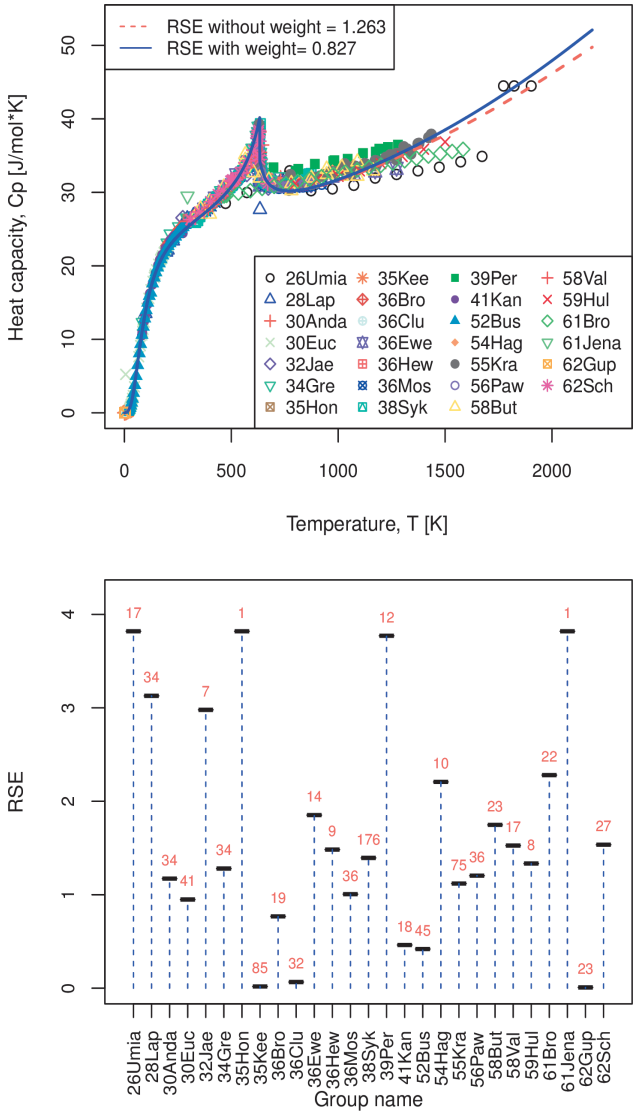

Figure 1. Fitted heat capacity of pure Ni using unweighted (red dashed line) and weighted (blue line) segmented regression [2] in comparison with experimental data (top), and RSE values (blue dashed lines) including the number of observations for each data set (red text) (bottom).

References:

[1] Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press New York, NY, USA,2014.

[2] I. Roslyakova, B. Sundman, H. Dette, L. Zhang, and I. Steinbach. Modeling of gibbs energies of pure elements down to 0 k using segmented regression. CALPHAD Journal, 55, 2016.

[3] Q.Chen and B.Sundman. Modeling of thermodynamic properties for bcc, fcc, liquid, and amorphous iron. Journal of Phase Equilibria, 22(6): 631-644, 2001