Materials Science and Technology, 31 (11), 1370--1375.

Prediction of martensite start temperature

Mathew Peet

September 23, 2014

Abstract

Methods have been evaluated for the prediction of the martensite–start temperature as a function of composition. Linear regression models have been improved by applying the concept of a committee borrowed from more sophisticated empirical techniques. Neural networks and thermodynamic models are tested, and a hybrid neural network model is developed using the thermodynamic model. The performance of the models is compared by different methods of assessment. The thermodynamic model performance was the best when tested within a typical range of the input–space. Bayesian neural network possess the advantage that the predictions are naturally accompanied by a measure of the uncertainty. It is demonstrated that combining the thermodynamic model with neural network can combine the advantages of the two methods.

1 Introduction

Due to the importance of phase transformations and heat treatments on the mechanical properties of steels, and due to the apparent simplicity of the concept and relative ease of measurement, a large number of studies have been made of the influence of elemental composition on the martensite–start temperature (M_S). Prediction of M_S enables the design of new alloys or heat treatments, and can give helpful insights in the context of understanding mechanical properties when retained austenite is present.

Many authors have used multiple linear regression equations to summarise the influence of the various alloying elements on the martensite–start temperature (M_S) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. The form is;

$$M_S = k_0 + \sum k_i w_i \tag{1}$$

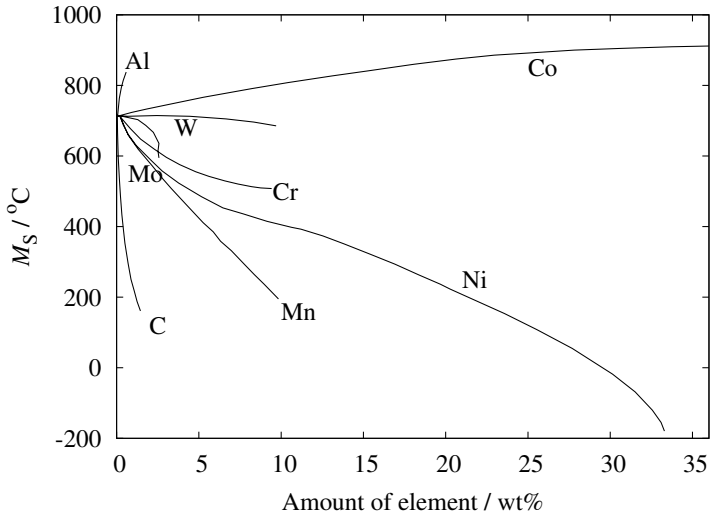

Figure 1: M_S temperature of Fe-X binary alloys (adapted from Izumiyami *et al.* [14]).

where k_0 is the 'M_S' of pure iron, w_i is the concentration of element i usually in weight percent, k_i is a parameter relating the concentration each element to the change in the M_S temperature.

Such equations are usually limited to a particular range of compositions, but the bounds of applicability may not have been determined or stated. There is often a temptation to use the equations outside the range of compositions used to determine the parameters.

One objection to such an approach is that the M_S temperature should not in general be expected to have a simple dependence on elemental composition. As shown in figure 1 [14]), even in binary solid solutions there is a non–linear dependence on solute concentration. Thermodynamic calculations of M_S using the Calphad approach therefore have an advantage over linear models, because they consider the change in chemical driving force due to binary and tertiary interactions, which is a more physically relevant parameter. Neural networks can also cope with the interdependencies necessary to estimate M_S [15], for example this approach was used by Vermeulen *et al.* [16].

Sourmail and Garcia–Mateo [17] have assessed various models for calculation of martensite–start temperature including thermodynamic calculation [18, 19] and Bayesian neural network methods [20], and have also made available an extensive database of steels [21]. They developed a model using a Bayesian neural network as developed by

Mackay that performed at least equally well as the thermodyanamic model approach.

Even the thermodynamic approach is eventually 'empirical' because it is necessary to calculate the critical value of the driving force necessary to trigger transformation. The estimation of the driving force itself is dependent on the empirical database and thermodynamic data.

Sourmail and Garcia–Mateo bounded the output of the neural network to prevent non–physical predictions. This approach was introduced by Yescas [22] in the context of phase fractions, but later results have shown that this can lead to poor performance by introducing an artificial bias in the cases were the bounding function is not supported by the underlying physics [23]. Therefore in this work a neural network was trained without using the bounding function for comparison.

In this work various models were compared. The concept of a committee of models was applied to multiple linear regression models. In addition to the Bayesian neural network, attempts were also made to combine the neural network method with thermodynamic modelling.

2 Method

A database (data set A) of 69 points collated by Sourmail and Garcia-Mateo but unused in the training of their model were used to assess the performance of various models by Pearson's correlation coefficient, the root mean squared error, test error and the log predictive error. The database provided by Sourmail contained data also used by Capdevilla *et. al.* [24], Vermuelen *et. al.* [16] and data used by Ghosh and Olson to derive the critical driving force.

The test error (TE) minimisation or sum squared error

$$\sum \left(t^{(m)} - y^{(m)} \right)^2 \tag{2}$$

is commonly used in assessment of regression models. The log predictive error (LPE) is an alternative, that has the advantage of penalising wild predictions less when they are accompanied by large uncertainties. Assuming that for each example m the model gives a prediction according to the normal distribution with average $y^{(m)}$ and variance ${\sigma^{(m)}}^2$;

$$\mathrm{LPE} = \sum_m \left[\frac{1}{2} \left(t^{(m)} - y^{(m)} \right)^2 / {\sigma^{(m)}}^2 + \log(\sqrt{2}\sqrt{\pi}\sigma^{(m)}) \right] \tag{3}$$

Pearson's correlation coefficient, r and it's square r^2 the coefficient of determination are also popularly used to compare model perfor-

mance. The sample correlation coefficient is defined in terms of the sample standard deviations s_x and s_y and the sample covariance s_xy.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{4}$$

Positive correlation is indicated when r is close to 1, and negative correlation when close to -1.

The residual sum of squares RSS is another measure of the discrepancy between the data and an estimation model,

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

where y_i is the ith value of the variable to be predicted, and $\hat{y}_i$ is the predicted value i.e. $f(x_i)$

There is scope to introduce alternate objective functions depending on the purpose of the modelling. Two functions were introduced to minimise both TE and LPE, the simplest being F_1=LPE×TE. Due to the sensitivity to scaling, if the function $\sqrt{(}\mathrm{TE}^2 + \mathrm{LPE}^2)$ were used the result would be dominated by the TE. The function F_2=$\sqrt{(}\mathrm{TE}_N^2 + \mathrm{LPE}_N^2)$ gives a similar ranking to F_1, where TE_N and LPE_N^2 are the values of TE and LPE normalised by dividing by the maximum value of each i.e. $\mathrm{TE}_N = \frac{\mathrm{TE}}{\mathrm{TE}_{\max}}$.

2.1 Regression equations

The MLR equations were ranked by their performance on data set A and combined to make committee models of the best n models. The average and standard deviations were calculated for these combinations. Equations due to Tamura, Pickering and Finkler and Sehirra [8, 11, 12] relevant to specific composition ranges where excluded at this stage since the predictions were usually outliers with respect to the other functions – although it is expected that they be preferred in their appropriate ranges.

Combining the equations in this simple way, is equivalent to allowing each parameter to be represented by a distribution of values rather than a single value, providing an assessment of the uncertainty as a function of position in the input space. Uncertainty of the parameters can be estimated in multi–linear regression, but these values are seldom reported or used. The approach espoused here benefits from the large number of independent assessments made by previous researchers to provide an assessment of predictive ability.

Similar concepts are used in Bayesian approaches to neural network modelling. For example the use of a committee of models combined to

produce lowest combined test error [20], and Markov–Chain Monte–Carlo techniques were the distribution of weights is explicitly assessed during training and incorporated into the predictions [25].

2.2 Neural Network models

Two neural network models were successfully trained using the training data used by Sourmail and Garcia–Mateo. The first simply used the database provided but without normalisation to constrain the output to the range 0-1000 K. The second used predictions of the thermodynamic model as implemented by Sourmail as an input for training to produce a hybrid model. Using the Neuromat model manager software a number of sub-models were trained and then ranked by log predictive error (LPE) and the best models combined to produce a committee with the lowest combined test error.

2.2.1 Attempt to model error

A neural network model was trained with the target being the difference between the value predicted using the Ghosh and Olsen thermodynamic model and the experimental database, but no systematic trend could be extracted. Such an approach seems to have the potential to improve performance if there were a systematic error in the model.

2.2.2 Hybrid model

Thermodynamic prediction of the martensite–start temperature was performed using the model of Ghosh and Olsen as implemented and provided by Sourmail (GOS). The thermodynamic prediction was added into the input data of the training database, to produce a dataset for the hybrid thermodynamic–Bayesian neural network model (Hybrid GOS/BNN).

3 Results

3.1 Performance of committee of linear models on 'data set A'

The committee model of the best 3 performing MLR equations had the lowest test error of the combined MLR models, performing better than any of the individual models. Considering the LPE then the committee model of the 11 MLR equations had the best performance

and performed reasonably well when measured by TE, therefore it was also the best performing MLR model when considering simultaneous minimisation of LPE and TE (see table 2). Comparison against the unseen data of 'test A' data with the 11 member committee model can be seen in figure 3. The error bars represent one standard deviation of the predictions of the various equations.

The combined empirical models work reasonably well for general steels included in 'test A' which have no elements added significantly beyond the austenite solubility limit. There is a systematic error in the calculation of high chromium steels in the test data set, these data can be expected to have formed an amount of ferrite. An improved calculation may be expected if the change in composition of the austenite can be accounted for, for example using thermodynamic calculation of equilibrium composition at high temperature. In the committee models these wild predictions are in regions where the various models disagree and are accompanied by larger uncertainties.

Further inspection of the predictions and the associated errors shows that the error bars are reasonably appropriate to the performance of the model, from 69 test cases 71% were within one standard deviation and 86% are within 2 standard deviations from the prediction.

Figure 2 is an example of how the predictions can change as a function of the input space, note with the MLR equations only linear trends can be captured and error can only vary linearly as a function of composition for each element.

3.2 Performance of thermodynamic and neural network models on 'test A' data

On the 'test A' database the model of Ghosh and Olson performed very well when measured by the test error. The RSS of 20 (standard error assuming no bias) is comparable to the standard deviation of 12 in experimental measurements of a single alloy with M_S 370°C reported by by Yang and Bhadeshia [26]. Sourmail and Garcia–Mateo previously demonstrated that the thermodynamic model is not able to perform as accurately when elements are added beyond their solubility limit in the austenite phase (i.e. addition of strong carbide formers such as Nb, V).

The unbounded Bayesian neural network performed best when measured by LPE but the worst when considering the test error. Comparing figure 7 and 8 it seems that a more general trend was extracted in comparison to when the bounding function was applied.

The performance of the hybrid thermodynamic-Bayesian neural

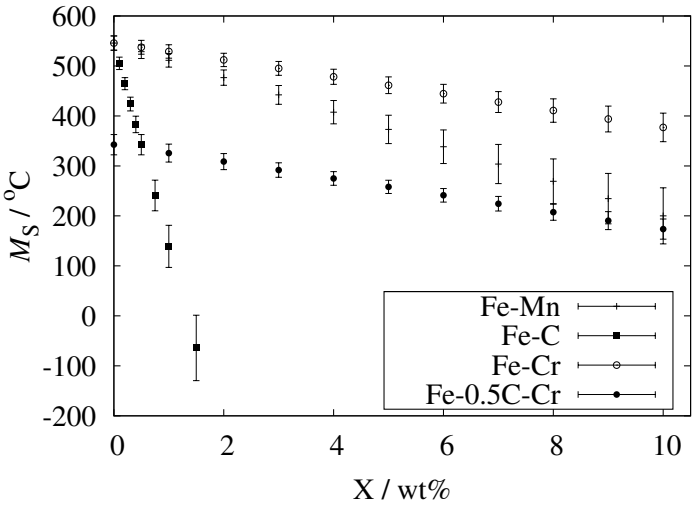

Figure 2: Calculated trends using committee of best 9 multi-linear models.

network model had a slightly better TE and a much improved LPE compared to the bounded BNN. Compared to the unbounded BNN the TE is very much better and the LPE performance is worse.

The performance of the various modelling techniques is summarised in table 1 and figure 4, figures 5–8 show the experimental versus predicted results for each of the models.

The thermodynamic model has the lowest test error when in the range of predictions, but these models are not accompanied by any indication of the accuracy. The test reported in table 1 can provide an error estimate of ± 20 based on RSS value. This results in a calculated LPE of 182, better than the Hybrid model, but should not properly be compared using the data from which the value of RSS was derived – additional data would be required to test this.

The neural network should be properly preferred because it offers to have the size of the error estimate to vary as a function of the input space, therefore the behaviour is that values far from the training data will be accompanied by large error estimates.

The Bayesian neural network trained without bounding the M_S had the best performance when measured using LPE but performed poorly when assessed by the test error.

Development of the hybrid model combining thermodynamic and BNN demonstrated that is can result in better performance when considering the error estimate accompanying the predictions.

Using the dataset applied the performance of the hybrid neural

Table 1: Performance of various models of M_{S} temperature.

Model	r	r^2	RSS	TE	LPE
Ghosh and Olson (Thermodynamic)	0.95	0.91	20	13466	∞
Sourmail (Bounded BNN)	0.85	0.72	38	50614	627.5
Present (BNN)	0.47	0.22	77	204224	166.9
Present (Hybrid BNN)	0.87	0.76	37	47998	214.2
Andrews (MLR)	0.55	0.30	35	42913	∞
Eldis (MLR)	0.92	0.85	36	45515	∞
Kung/Andrews (MLR)	0.91	0.83	36	44265	∞
Kung/Steven and Haynes (MLR)	0.91	0.83	47	77152	∞
Steven and Haynes (MLR)	0.91	0.83	47	77242	∞
Kunitake (MLR)	0.89	0.80	46	73515	∞
Kunitake and Ohtani (MLR)	0.86	0.75	46	73515	∞
Tamura[2] (MLR)	0.86	0.75	47	73409	∞
Monma (MLR)	0.86	0.73	48	80785	∞
Nehrenberg (MLR)	0.77	0.59	70	170665	∞
Finkler and Sehirra (MLR)	0.91	0.82	76	196716	∞
Tamura[1] (MLR)	0.70	0.49	90	277455	∞
Payson and Savage (MLR)	0.70	0.49	105	379927	∞
Rowland and Lyle	0.72	0.52	107	395527	∞
Andrews b (Polynomial)	0.55	0.30	109	410657	∞
Grange and Stewart (MLR)	0.63	0.40	157	850835	∞
Carapella	0.63	0.39	115	459241	∞
Pickering (MLR)	0.85	0.73	186	1190500	∞
Present (Best 3 MLR)	0.93	0.86	33	36619	789.3
Present (Best 5 MLR)	0.92	0.85	35	41742	254.1
Present (Best 9 MLR)	0.91	0.83	39	51594	268.7
Present (Best 10 MLR)	0.91	0.82	40	55337	230.4
Present (Best 11 MLR)	0.91	0.82	40	56025	182.6

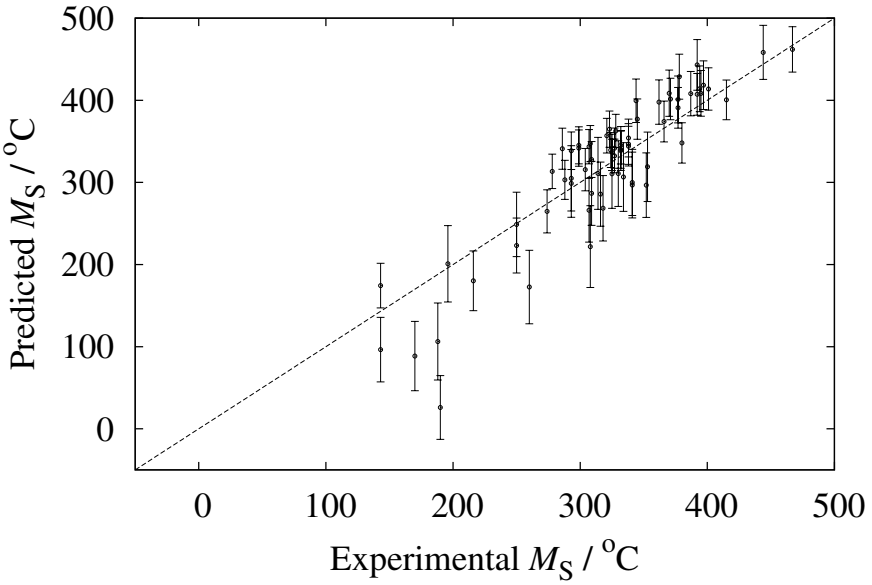

Figure 3: Combined prediction from best 11 linear regression models.

Table 2: Ranking of models using combined LPE×TE parameter.

Model	LPE	TE	$F1$ / 10^6	Rank	$F2$	Rank
Present MLR best 11	56025	183	10.23	1	0.3589	1
Present hybrid BNN	47998	214	10.28	2	0.3590	2
Present MLR best 5	41742	254	10.61	3	0.3813	3
Present MLR best 10	55337	230	12.75	4	0.3983	4
Present MLR best 9	51594	269	13.86	5	0.4239	5
Present MLR best 3	36619	789	28.90	6	1.0159	7
Sourmail bounded	50614	628	31.76	7	0.8327	6
Present BNN	204224	167	34.08	8	1.0221	8

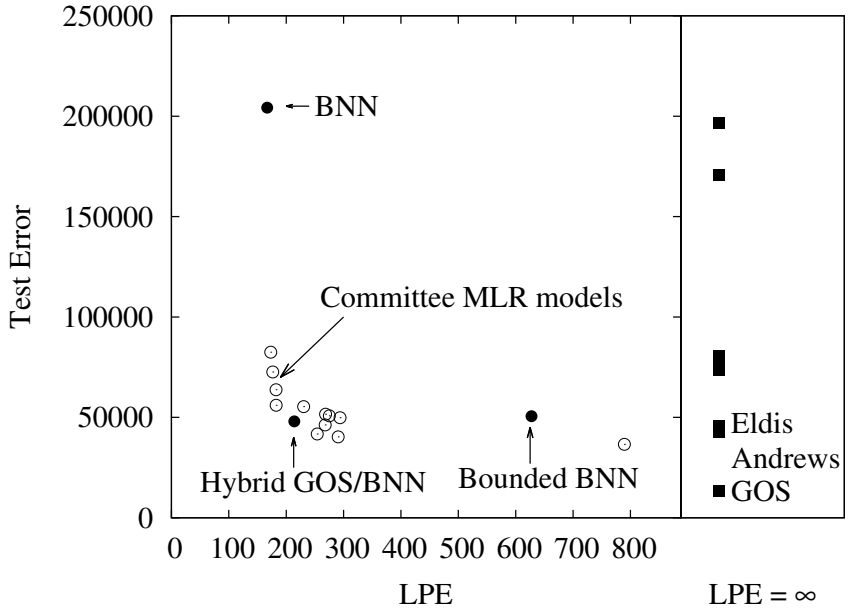

Figure 4: Performance of various models on 'test A' of M_{S}, the model with lowest test error is that by Ghosh and Olson (GOS). Bounded BNN is due to Sourmail and Garcia–Mateo, Models BNN, Hybrid GOS/BNN and combined linear models are as described in this work.

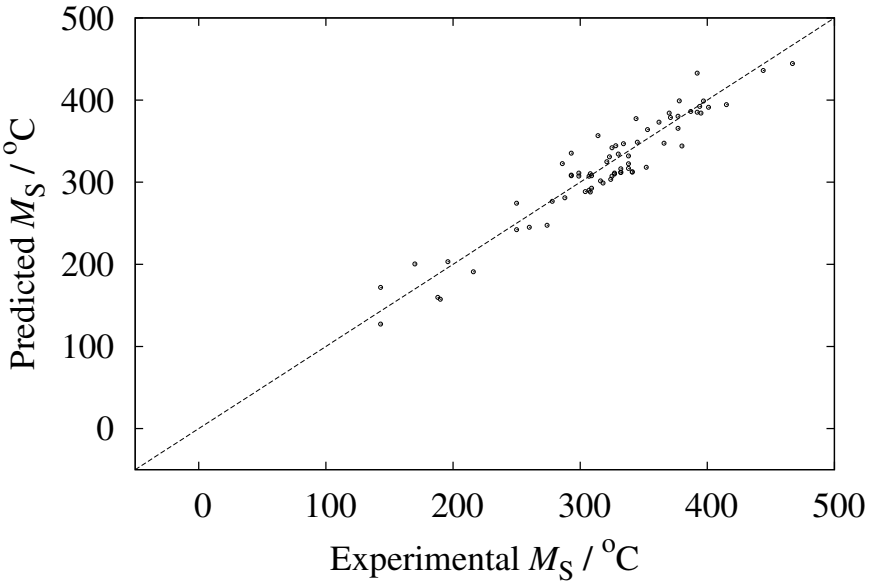

Figure 5: Ghosh and Olson model performance on 'test A', calculation following Sourmail.

network was similar to the committee model of linear regression models. Committee of 11 models and the hybrid BNN were ranked first and second by the parameters $F1$ and $F2$. This may be a result of the limited testing applied, the neural network contains knowledge of many more elements, but in principle these data could also be used to train new multi-linear models.

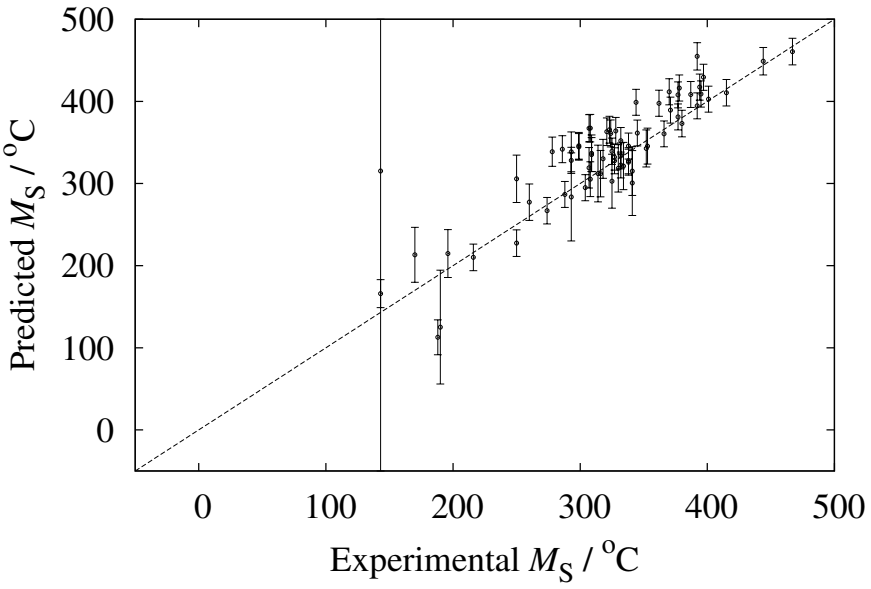

Figure 6: Hybrid model performance on 'test A', BNN with additional input from Ghosh and Olson model.

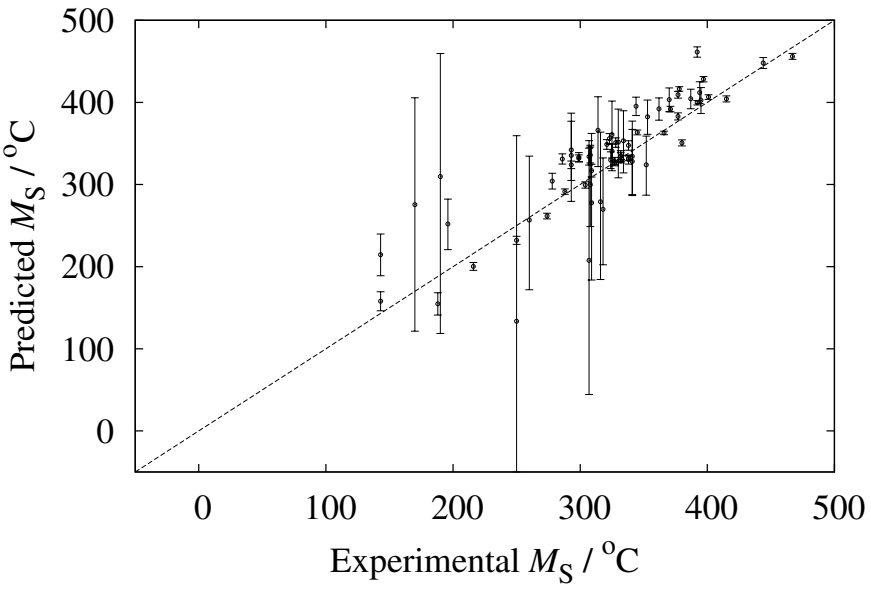

Figure 7: Bounded BNN of Sourmail and Garcia–Mateo, performance on 'test A'.

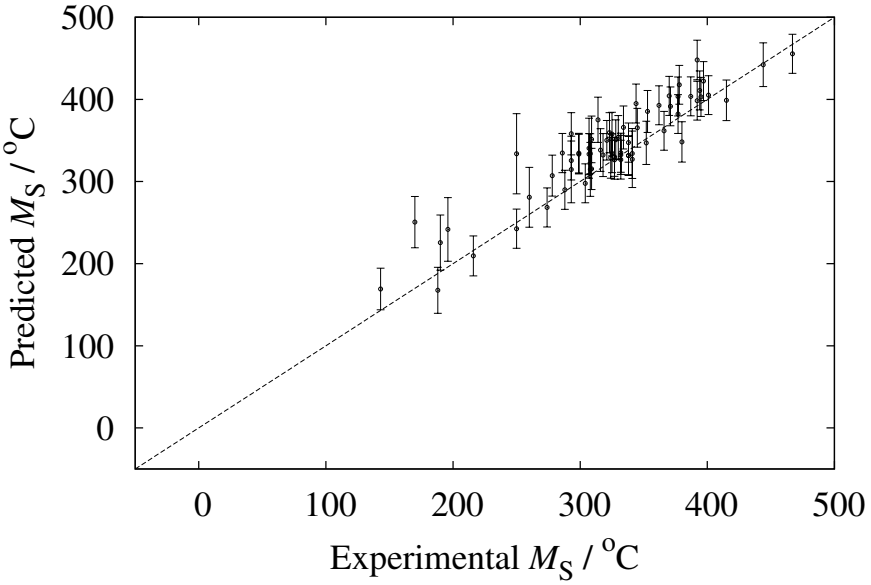

Figure 8: Performance on 'test A', Unbounded model using data collated by Sourmail and Garcia–Mateo

4 Conclusions

The performance of multiple linear regression models was improved by combining them into committee models. This has the advantage of introducing an estimate of the accuracy as well as allowing the results from various studies to be combined. The model with the best 11 performing linear models in the committee performed best when considering both test error and log predictive error, and the committee 3 performed the best considering only the test error.

The purely thermodynamic model is expected to give the best predictions as measured by TE. An error estimate of 20 was calculated for the thermodynamic model. Use of this thermodynamic model requires access commercially available thermodynamic databases and software.

It may not be safe to assume that the thermodynamic model will always give superior performance, the neural network schemes offer the advantage that large error estimates can provide an appropriate warning of extrapolation out of the data range.

Combining both thermodynamic modelling and Bayesian neural network, a hybrid model was produced. This allowed the strengths of both methods to be combined.

5 Acknowledgements

I am grateful to Prof. H.K.D.H. Bhadeshia and Dr M. C. Tsai for their useful feedback on this manuscript.

References

[1] P. Payson and C. H. Savage. Martensite reactions in alloy steels. *Trans. A.S.M.*, 33:261–281, 1944.

[2] L. A. Carapella. "Computing A" or m_s (transformation temperature on quenching) from analysis. *Metals Progress*, 46:108–108, 1944.

[3] E. S. Rowland and S. R. Lyle. The application of m_s points case depth measurement. *Trans. A.S.M.*, 37:27–47, 1946.

[4] R. A. Grange and H. M. Stewart. The temperature range of martensite formation. *Trans. A.I.M.E.*, 167:467–490, 1946.

[5] A. E. Nehrenberg. Discussion. *Trans. A.I.M.E.*, 167:494–498, 1956.

[6] W. Steven and A. G. Haynes. The temperature formation of martensite and bainite in low–alloy steels, some effects of chemical composition. *J. Iron Steel Inst.*, 183:349–359, 1956.

[7] K. W. Andrews. Empirical formulae for calculation of some transformation temperatures. *J. Iron Steel Inst.*, 203:721–727, 1965.

[8] I. Tamura. *Steel Material Study on the Strength*. Nikkan Kogyo Shinbun Ltd., Tokyo, 1970. in Japanese.

[9] K. Monma. *Tekko-zairyo-gaku*, chapter Martensite. Jikkyo-Syuppan Corp., 1972. in Japanese.

[10] C. Y. Kung and J. J. Rayment. An examination of the validity of existing empirical formulae for the calculation of m_s temperature. *Metallurgical Transactions A*, 1982.

[11] F. B. Pickering. Physical metallurgical development of stainless steels. In *Stainless steels '84*, pages 2–28. The Institute of Metals, London, 1985.

[12] H. Finkler and M. Schirra. Transformation behaviour of the high temperature martensitic steels with 8%–14% chromium. *Steel Research*, 67(8):328–342, 1996.

[13] T. Kunitake. Prediction of Ac1, Ac3, and MS temperatures of steel by empirical formulas. *J. Jpn. Soc. Heat Treat.*, 41:164–169, 2001.

[14] M. Izumiyama, M. Tsuchiya, and Y. Imai. Effects of alloying element on supercooled a3 transformation. *J. Japan Inst. Met*, 34(3):291–295, 1970.

[15] H.K.D.H. Bhadeshia. Neural networks in materials science. *ISIJ*, 39:966–979, 1999.

[16] W.G. Vermeulen, P.F. Morris, A.P. de Weijer, and S. van der Zwaag. Prediction of martensite start temperature using artificial neural networks. *Ironmaking and Steelmaking*, 23(5):433–437, 1996.

[17] T. Sourmail and C. Garcia-Mateo. Critical assessment of models for predicting the m_s temperature of steels. *Computational Materials Science*, 34:323–334, 2005.

[18] G. Ghosh and G. B. Olson. Kinetics of FCC→BCC heterogeneous martensitic nucleation, part 1: The critical driving force for athermal nucleation. *Acta Mat.*, 42:3361–3370, 1994.

[19] G. Ghosh and G. B. Olson. Kinetics of FCC→BCC heterogeneous martensitic nucleation, part ii: Thermal activation. *Acta Mat.*, 42:3371–3379, 1994.

[20] T. Sourmail and C. Garcia-Mateo. A model for predicting the m_s temperatures of steels. *Computational Materials Science*, 34:213–218, 2005.

[21] http://www.msm.cam.ac.uk/map/.

[22] M. Yescas, H.K.D.H. Bhadeshia, and D.J.C. MacKay. Estimation of the amount of retained austenite in austempered ductile irons. *Materials Science Engineering A*, A311:162–173, 2001.

[23] H.K.D.H. Bhadeshia. Neural networks in materials science. *Statistical Analysis and Data Mining*, 1:296–305, 2009.

[24] C. Capdevilla, F.G. Caballero, and C. Garcia de Andrés. Determination of ms temperature in steels: A bayesian neural network model. *ISIJ*, 42:894–902, 2002.

[25] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

[26] Hong-Seok Yang and H.K.D.H. Bhadeshia. Uncertainties in dilatometric determination of martensite start temperature. *Materials science and technology*, 23(5):556–560, 2007.

6 Appendix

Table 3: Ranges of applicability of equations where available.

	Number		C	Mn	Si	S	P	Ni	Cr	Mo	
Steven	59	min	0.11	0.2	0.11	0.004	0.005	0.00	0.00	0.0	
and		mean	0.33	0.702	0.241	0.019	0.019	1.918	0.93	0.247	
Haynes		max	0.55	1.67	1.74	0.014	0.038	5.04	3.34	1.0	
Andrews	184	min	0.11	0.04	0.11	0.000	0.000	0.00	0.00	0.0	
		mean	0.35	0.759	0.314	0.014	0.015	1.289	0.838	0.238	
		max	0.60	4.87	1.89	0.046	0.046	5.04	4.61	5.4	
Kunitake	54	min	0.13	0.39	0.15	–	–	0.00	0.00	0.0	
		max	0.73	2.04	1.62	–	–	5.00	1.95	0.54	
Sourmail	1091	min	0.0	0.0	0.0	–	–	0.00	0.00	0.0	V,Co,Al,
			0.36	0.74	0.31			3.70	1.01	0.24	W,Cu,Nb,
		max	2.25	10.24	3.8	–	–	31.54	17.98	8.0	Ti,B,N