INSTITUTE OF PHYSICS PUBLISHING MODELLING AND SIMULATION IN MATERIALS SCIENCE AND ENGINEERING

Modelling Simul. Mater. Sci. Eng. **13** (2005) 633–644 doi:10.1088/0965-0393/13/4/013

Statistical modelling of composition and processing parameters for alloy development

Alexandr Golodnikov[1], **Yevgeny Macheret**[2], **A Alexandre Trindade**[3], **Stan Uryasev**[1,4] **and Grigoriy Zrazhevsky**[1]

[1] Department of Industrial and Systems Engineering, University of Florida, PO Box 116595, Gainesville, FL 32611, USA
[2] Institute for Defense Analysis, 4850 Mark Center Dr., Alexandria, VA 22311, USA
[3] Department of Statistics, University of Florida, PO Box 118545, Gainesville, FL 32611, USA

E-mail: uryasev@ufl.edu

Received 21 January 2005, in final form 18 April 2005
Published 17 May 2005
Online at stacks.iop.org/MSMSE/13/633

Abstract

We propose the use of regression models as a tool to reduce time and cost associated with the development and selection of new metallic alloys. A multiple regression model is developed which can accurately predict tensile yield strength of high strength low alloy steel based on its chemical composition and processing parameters. Quantile regression is used to model the fracture toughness response as measured by Charpy V-Notch (CVN) values, which exhibits substantial variability and is therefore not usefully modelled via standard regression with its focus on the mean. Using Monte-Carlo simulation, we determine that the three CVN values corresponding to each steel specimen can be plausibly modelled as observations from the 20th, 50th and 80th percentiles of the CVN distribution. Separate quantile regression models fitted at each of these percentile levels prove sufficiently accurate for ranking steels and selecting the best combinations of composition and processing parameters.

1. Introduction

High strength low alloy (HSLA 100) steel was developed as a replacement for HY 100 in order to reduce fabrication costs by eliminating pre-heating during welding operations. The main goal of the development was to produce a homogeneous, fine grain, low carbon martensite microstructure with finely dispersed transformation phases. Important results and data on

[4] Author to whom any correspondence should be addressed.

the effects of composition, heat treatment and plate gauge on mechanical properties and Charpy V-Notch (CVN) impact on toughness are summarized by Goldren and Cox (1986), and Czyryca *et al* (1990).

Stochastic modelling of these data by Metzbower and Czyryca (2002) using neural networks[5], indicated trends in the dependence of mechanical properties on the alloy composition and heat treatment. The correlation between model predictions and experiments was shown to be higher for elongation and tensile yield strength than for CVN values. The CVN impact test is designed to provide a measure of metal resistance to fast fracture in the presence of a flaw or notch. It has been used extensively in mechanical testing of steel products, in research, and in procurement specifications for over three decades. Detailed descriptions of the CVN testing procedures and data analysis can be found in Ireland *et al* (1975), Holt (1990), Schmitt *et al* (1990) and Liaw (2000).

Although it is a popular test for characterizing metal toughness, the CVN test has some drawbacks. The measured quantity, fracture energy, is an invariant material parameter and depends on the geometry and testing temperature of a specimen (Corowin and Houghland 1986, Lucon *et al* 1999). The lack of homogeneity in the microstructure near the notch can lead to markedly scattered impact toughness values. There is also uncertainty associated with the location of the ductile-to-brittle transition region. These and other factors result in the large variation commonly observed in the Charpy impact testing results (Todinov 2004), and is the main reason for the difficulty in predicting CVN values.

The primary motivation behind the present investigation is to explore ways to improve the aforementioned model prediction accuracy, and develop computational tools to aid in alloy design, leading to a reduction in the amount of experimental melts. The hope is that these computational tools can in turn lead to efficient ways of optimizing the chemical composition and heat treatment of alloys. The results described in this paper show that quantile regression analysis is an effective tool for predicting CVN values (section 3). In combination with linear regression for yield strength (section 2), it can be used to accurately predict HSLA 100 steel properties on the toughness diagram (section 4). This raises the possibility of using the models to improve alloy design by helping refine successive experiments, ultimately identifying optimal combinations of composition and processing parameters. Such a procedure requires careful mathematical formulation as a numerical optimization problem. We have succeeded in implementing this approach, building on the results in this paper. The findings, presented in Golodnikov *et al* (2005b), show this to be a viable option.

The data for our analysis, obtained by normalizing that reported in Goldren and Cox (1986), showed significant effects of composition (C, Mn, Si, Cr, Ni, Mo, Cu), plate thickness, solution treating and ageing temperatures on tensile yield strength, elongation and CVN properties. As described by Goldren and Cox (1986), the alloy development included three rounds of laboratory melting, processing and testing of candidate steels. Round 1 involved the screening of 16 exploratory compositions. Round 2 consisted of additional laboratory testing of new heats for the four best compositions from Round 1. The most promising steel from Round 2 (the prime candidate steel) was the subject of more extensive testing in Round 3, in which nine heats were tested, including high side, low side and midrange chemistries. The ranking and selection of best alloys were based on combinations of yield strength and CVN at -84°C. Triplicate tests were run to obtain CVN values.

The actual data we analysed consisted of results for a total of 78 specimens of steels of 30 distinct types. The types (and number of specimens of each type) are as follows: 1–16 (1),

[5] Neural networks are flexible stochastic models for predicting an output (dependent) variable as a function of several input (independent) variables.

17–20 (6), 21 (1), 23 (1), 24 (8), 28 (8), 30 (4), 31 (2), 32 (8), 34 (1), 35 (4), 37 (1). The chemical compositions of the 30 types are given in appendix A. Apart from these compositions, there corresponds also to each specimen a single value of yield strength, and three values of CVN at −84°C. About 5% of N and Cb compositions in the original database were missing. In the ensuing analyses, we substituted median values for those missing data. The yield strength, chemical composition and temperature data were normalized by their average values. The plate thickness was used in absolute length dimensions, millimetres. In order to avoid negative values, the CVN data were modelled on the logarithm scale, and normalized by the average of the log CVN values. The CVN values appearing throughout this paper were obtained by exponentiating the normalized log CVN values.

2. A linear regression model for tensile yield strength

Linear regression is a widely used statistical procedure for modelling the relationship between several independent (or explanatory) variables, $X_1, \ldots, X_k$, and a particular dependent (or response) variable of interest, Y. The aim is that observations from the dependent variable, $Y_1 \ldots, Y_n$, can be modelled as a linear combination of observations from the corresponding independent variables,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i, \qquad i = 1, \ldots, n$$

for some unknown coefficients $\beta_0, \beta_1, \ldots, \beta_k$, to be estimated from the available data. The ϵ_i are zero-mean random terms that account for the surplus variability or scatter in Y that cannot be explained by $X_1, \ldots, X_k$. One can measure the proportion of variability in Y that is in fact explained by the Xs through a quantity known as R^2; e.g. $R^2 = 90\%$ means that only 10% of the variation in Y cannot be accounted for by the Xs. Many phenomena in science can be usefully modelled in this simple way, the cumulative effects of unmeasured and/or unforeseen variables being lumped into the stochastic ϵ term.

One of the main issues in regression model building when many independent variables are involved, is how to choose a subset so that a large value of R^2 is obtained with few X variables. It is desirable to exclude from selection variables that contribute little to the explanation of Y. Several *model selection* tools are available, and we refer the interested reader to Neter *et al* (1996), where an extensive treatment of all the facets of regression analysis can be found.

In an attempt to build regression models for yield strength and CVN at −84°C as the dependent variables with the data in appendix A, the available independent variables were divided into two groups: important and non-important. The important variables, found to have an appreciable effect on yield strength ('Yield') and CVN at −84°C, consisted of: C, Mn, Cr, Ni, Mo, Cu, N, V, plate thickness ('Thick'), solution treating ('Aust') and ageing temperatures ('Ageing'). The important variables comprised the pool of potential explanatory variables in the regression model building, except for Ni, which was highly correlated with these. (Collinear variables do not provide new information when considered simultaneously in a regression model and may be excluded without substantial loss of information.) The non-important variables, Si, Al, P, Cb and S, were not used in regression model building; merely maintained at some appropriately low level. A small, fixed amount of Cb was added to each alloy for possible improved precipitation strengthening. (In experimental melts the concentration of Cb typically shows some variability and we thought it important to capture this effect.)

To build a linear regression model for Yield, we used the method of *stepwise* variable selection in our search for a set of explanatory variables that might parsimoniously

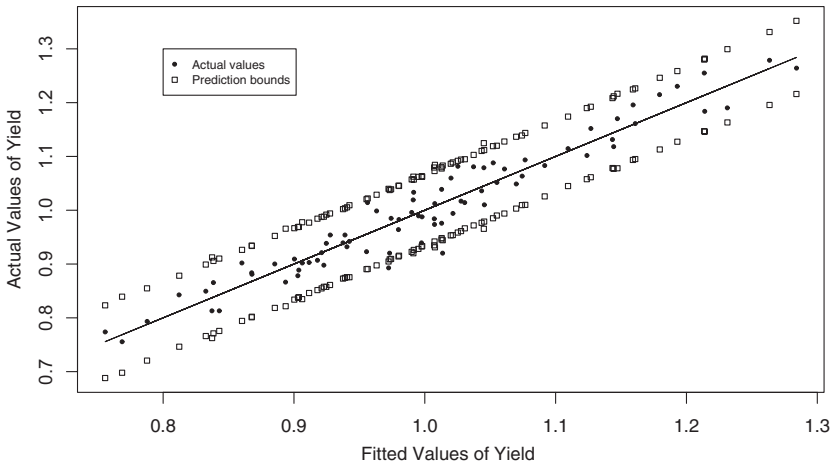

Figure 1. Actual versus linear regression model fitted values for Yield.

account for its variability (Neter *et al* 1996). This automated procedure led to the model

$$\text{Yield} = 0.7 + 0.173\,\text{C} + 0.045\,\text{Mn} + 0.13\,\text{Cr} + 0.123\,\text{Mo} + 0.358\,\text{Cu} + 0.004\,\text{V} - 0.002\,\text{Thick} + 0.285\,\text{Aust} - 0.759\,\text{Ageing},$$

which was also arrived at independently using other standard regression model selection criteria. Figure 1 shows the observed data plotted as a function of the fitted values from the model. The solid line is the estimated mean value of Yield. A 95% confidence band for this mean is also displayed. Although there are some influential points (observations that have a substantial effect on the fit of the model), additional diagnostic plots suggest that the regression model is an adequate fit to the data. All estimated coefficients are statistically significant, a strong indicator that each of the variables in the model is an important predictor of Yield. The model is able to account for about 94% of the variability in tensile strength ($R^2 = 0.94$). If the ith specimen has observed and (model) predicted Yield values of Y_i and P_i, respectively, we define the absolute prediction error as $E_i = 100|Y_i - P_i|/Yi$. For this dataset, the mean and maximum absolute prediction errors are 2.4% and 10.2% (steel # 34), respectively.

Along the lines of Metzbower and Czyryca (2002), we also fitted neural network models to these data. To check a regression model's predictive capability in a way that would enable a direct comparison to be made with these more flexible models, we formed sequences of in-sample datasets $B_1, B_2, \ldots, B_{78}$ and corresponding out-of-sample datasets $C_1, C_2, \ldots, C_{78}$. Each in-sample set B_i consists of all 78 data values except for the ith data value, $\{y_i\}$, which comprises exactly the corresponding out-of-sample set C_i. For each set B_i, we then fitted the multiple regression model, and used it to predict the yield value for the single data value in the out-of-sample set C_i. The mean absolute prediction error thus obtained was 2.8%. However, since the mean absolute prediction errors attained for the regression model (2.4% in sample and 2.8% out-of-sample) were in general superior to those attained for the neural networks, the regression model was selected for all subsequent work. Further statistical details on both the regression and neural network analyses can be found in Golodnikov *et al* (2005a).

3. Quantile regression models for CVN at −84°C

We encountered difficulty in attempting to build a linear regression model for CVN at −84°C. Whereas each specimen is characterized by a single value of yield strength, three values (as a rule distinct) are always observed for CVN. This and the fact that there is substantial variability in these values complicates the process of screening specimens with respect to CVN. As specified by Goldren and Cox (1986), an acceptable specimen must satisfy both a strength requirement (yield strength in the range 0.899–1.079), and a toughness requirement. The latter requires that the average of the three values of CVN be greater than 2.568, with no single value below 2.568 by more than 0.313. In other words, if for some specimen the smallest of its three values of CVN is less than 2.14, it should be disqualified. It therefore follows that the mean value of CVN cannot be used as a criterion for steel selection. Consider, for example, the following two specimens:

(a) Steel # 30 with Thick = 51 mm, Aust = 0.988, Ageing = 0.888, had the following three values of CVN: 3.307, 3.179, 1.955 (mean value of CVN = 2.813, minimum value of CVN = 1.955).
(b) Steel # 8 had the following three values of CVN: 2.599, 2.658, 2.728 (mean value of CVN = 2.662, minimum value of CVN = 2.599).

If the mean value of CVN is used as a criterion, the first specimen should be classified as better than the second. On the other hand, the minimal value of CVN for the first specimen, 1.955, is less than 2.14, which therefore disqualifies it. The second specimen, however, qualifies, since both its mean and minimum values of CVN exceed 2.568.

We therefore need a more flexible approach which can predict, for example, any given quantile (or percentile) of the distribution of CVN, rather than a single value (the mean) as is the case with standard regression. *Quantile regression*, a nonparametric generalization of regression introduced by Koenker and Bassett (1978), permits one to do just that. Its nonparametric nature means that minimal distributional assumptions are imposed on the data, i.e. the response is not required to be normal. It also allows for heteroscedasticity in the data, that is the variability in the response can vary for different values of the explanatory variables.

The concept of a *quantile* (or *percentile*) of the empirical distribution of some quantitative characteristic, such as CVN, can be explained as follows. If n specimens are ranked according to CVN, with the 1st being the smallest and the nth the largest, then, for any $0 \leqslant \theta \leqslant 1$, the θth quantile (or $100\,\theta$th percentile) of this distribution would be the CVN value of the $n\theta$th specimen. If, as is usually the case, $n\theta$ is not an integer, the quantile is obtained by linear interpolation between the two adjacent CVN values. In quantile regression, the quantile of a dependent variable is generalized to a regression setting where it is modelled as a function of the independent variables. A more technical description of quantile regression can be found in Golodnikov *et al* (2005a), and is reproduced here in appendix B for the interested reader.

In order to determine which quantile might plausibly correspond to the smallest of the three values of CVN, we successively drew three random values from a standard normal distribution, each time selecting only the smallest value. This was repeated 10 000 times, producing an average value of 0.19. The interpretation is that the smallest of three such CVN values can be approximately equated with the 20th percentile of the CVN distribution. By symmetry, the middle and largest values can therefore be associated with the 50th and 80th percentiles, respectively.

We used the method of stepwise variable selection in building a quantile regression model for the 20th percentile of log CVN. As in its standard regression counterpart, this procedure adds and removes explanatory variables alternately from the model using significance tests,

Table 1. In-sample and out-of-sample performance of the fitted 20th, 50th and 80th percentile regression models for log CVN.

	Number and percentage of actual values of log CVN		
Falling	20% surface	50% surface	80% surface
Strictly below surface (in-sample)	38 (16.2%)	111 (47.4%)	189 (80.8%)
Below or on surface (in-sample)	45 (19.2%)	127 (54.3%)	198 (84.6%)
Strictly below surface (out-of-sample)	48 (20.5%)	115 (49.2%)	184 (78.6%)
Below or on surface (out-of-sample)	55 (23.5%)	119 (50.9%)	186 (79.5%)
Total number of points	234	234	234

eventually converging on a candidate model(s). A quantile regression model fitting library, *quantreg*, is available in the statistical software R (http://lib.stat.cmu.edu/R/CRAN/). It does not, however, include an automated tool for performing stepwise variable selection; thus we wrote our own program, which arrived at two possible models. These were combined (by averaging coefficients) giving the model:

$$\begin{aligned}20\%\ \log \mathrm{CVN} = {} & 0.000 - 0.1\ \mathrm{Mn} + 0.04\ \mathrm{Cr} - 0.419\ \mathrm{Mo} + 0.608\ \mathrm{Cu} - 0.144\ \mathrm{N} \\ & -0.035\ \mathrm{V} - 0.004\ \mathrm{Thick} - 0.693\ \mathrm{Aust} + 1.692\ \mathrm{Ageing}.\end{aligned}$$

This procedure was repeated for the 50th and 80th percentile surfaces of log CVN, the respective percentiles for the middle and largest of the three values of CVN observed for each specimen. In the former just one model was identified, while the latter surface is again an average of two competing models.

$$\begin{aligned}50\%\ \log \mathrm{CVN} = {} & 0.018 - 0.27\ \mathrm{C} - 0.143\ \mathrm{Mn} - 0.34\ \mathrm{Mo} + 0.388\ \mathrm{Cu} - 0.227\ \mathrm{N} \\ & -0.035\ \mathrm{V} + 1.538\ \mathrm{Ageing} \\ 80\%\ \log \mathrm{CVN} = {} & 1.05 - 0.103\ \mathrm{C} - 0.258\ \mathrm{Mn} + 0.05\ \mathrm{Cr} - 0.073\ \mathrm{Mo} - 0.105\ \mathrm{Cu} \\ & -0.170\ \mathrm{N} - 0.02\ \mathrm{V} - 0.001\ \mathrm{Thick} - 0.462\ \mathrm{Aust} + 1.23\ \mathrm{Ageing}.\end{aligned}$$

To assess the goodness of fit, Koenker and Machado (1999) introduced a quantile regression equivalent of the coefficient of multiple determination, $R^1(\theta)$, which measures the proportion of the 'variability' in the response that is accounted for by the fitted θth quantile surface. The variability measured here though is that assessed by the weighted combination of absolute residuals in criterion (B.1), and thus $R^1(\theta)$ constitutes a local measure of goodness of fit for the θth quantile rather than a global measure over the whole distribution like R^2. The three models fitted above had the following values: $R^1(0.2) = 0.52$, $R^1(0.5) = 0.51$, $R^1(0.8) = 0.42$. While not as dramatic as the $R^2 = 0.94$ value attained for Yield, capturing about 50% of the variability in the quantiles does allow for the useful utilization of these models as classifiers in ranking steels, as will be seen in section 4.

A few other obvious diagnostics can also be checked. First, the in-sample fit of the models can be assessed by counting the percentage of actual log CVN values that fall strictly below and at or below the fitted regression surfaces. For example, if the 20% regression model is a good fit, the first percentage should be less than or equal to 20, while the second should be greater than or equal to 20. Analogous results hold for the 50% and 80% models. The observed values are presented in the top portion of table 1.

To check the out-of-sample fit, we formed sequences of in-sample datasets $B_1, \ldots, B_{234}$ and corresponding out-of-sample datasets $C_1, \ldots, C_{234}$, as in section 2. Each in-sample set B_i consists of all 234 data values except the ith data value, $\{y_i\}$, which comprises exactly

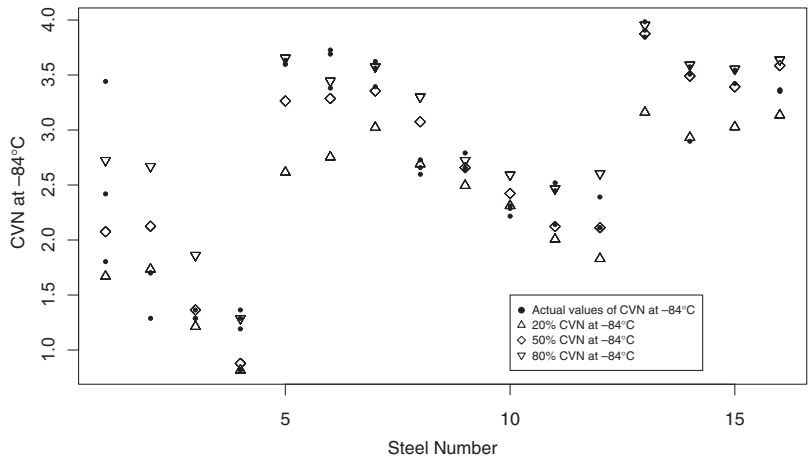

Figure 2. Actual values of CVN at −84° C and 20th, 50th and 80th percentile predictions from the fitted quantile regression models. (Steels 1–16 only.)

the corresponding out-of-sample set C_i. For each set B_i, we then fit the percentile regression model, and use it to predict the CVN value for the single data value in the out-of-sample set C_i. These predicted and actual data values are again compared in terms of how often the actual values fall below and at or below the predicted values. The results appear in the bottom portion of table 1. In both the in-sample and out-of-sample cases the attained percentages of points falling below and at or below the predicted values are close to the nominal levels of 20%, 50% and 80%. This suggests that the proposed quantile regression models for log CVN are a reasonably good fit to the data.

Finally, figure 2 shows both the fitted percentiles (20%, 50%, 80%) as well as the three observed values of CVN for each of steels 1–16 in the dataset. As can be seen, the fitted values provide plausible estimates of the percentiles in question. The results for the remaining steels, 17–37, are similar.

4. Use of the regression models for ranking steels

The regression models developed in the previous section for Yield and the 20th, 50th and 80th percentiles of CVN at −84°C, can now be used to select the best steels, paralleling the approach of Goldren and Cox (1986). In their alloy development programme Goldren and Cox (1986) conducted a total of three rounds of laboratory melting, processing and testing of candidate steels. Round 1 involved the screening of 16 steels. Round 2 consisted of additional laboratory testing of new heats for the four best steels from Round 1. The most promising steel from Round 2 was the subject of more extensive testing in Round 3, resulting in the identification of the best steel. The screening of specimens used actual values of yield strength and the CVN.

In section 3 we showed that the average of the smallest of the three values of CVN corresponds to approximately 20% of the distribution of the CVN. Therefore, we may use the quantile regression model for the 20% of CVN in the screening phase to certify that a specimen satisfies the toughness requirement, if its model-predicted 20% value exceeds 2.568.

As a check on the validity of this approach, we will now compare the decisions made in Rounds 1–3, which are based on actual test data, with the analogous decisions which would have been reached using the regression models, which are based on predicted data. For this

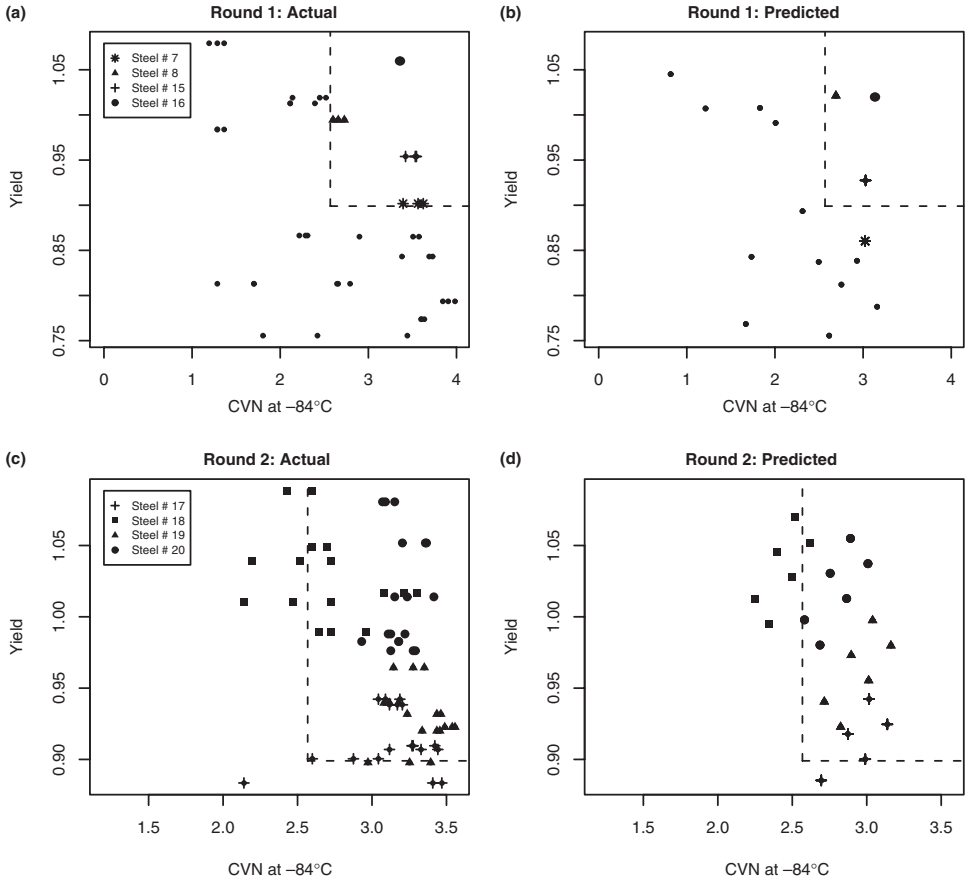

Figure 3. Round 1 (steels 1–16) and Round 2 (steels 17–20) results. Actual and predicted values of Yield plotted versus actual and 20% predicted values of CVN at −84°C.

purpose we constructed two graphs for each Round (figures 3 and 4). The first graph shows actual values of yield and CVN, and the second predicted values from the models. Note also that the first contains three times the number of points as the second, since the second constitutes predictions for just the lowest of the three values in the first. The dashed lines in the toughness diagram designate lower acceptance bounds for strength (0.899) and toughness (2.568).

From the toughness diagram based on the actual data of figure 3(*a*), steels 7, 8, 15 and 16 were the only ones that satisfied both the strength and the toughness requirements. Consequently, these four steels were selected for further study in Round 2. The diagram based on predictive data in figure 3(*b*) certifies these same steels as meeting both the strength and the toughness requirements, with the exception of steel 7 which does not satisfy the strength requirement (and should therefore be disqualified). The use of predicted data thus results in the rejection of one additional steel in Round 1, as compared with actual test data.

From figure 3(*c*) we can conclude that three specimens of steel 18 do not satisfy the toughness requirement, and one specimen of steel 17 does not satisfy the strength requirement. Of the remaining specimens that meet both the strength and the toughness requirements, steels 19 and 20 exhibit the best toughness combinations. On the basis of the toughness diagram based on actual data in figure 3(*c*), steel 20 would be deemed to be superior to steel 19.

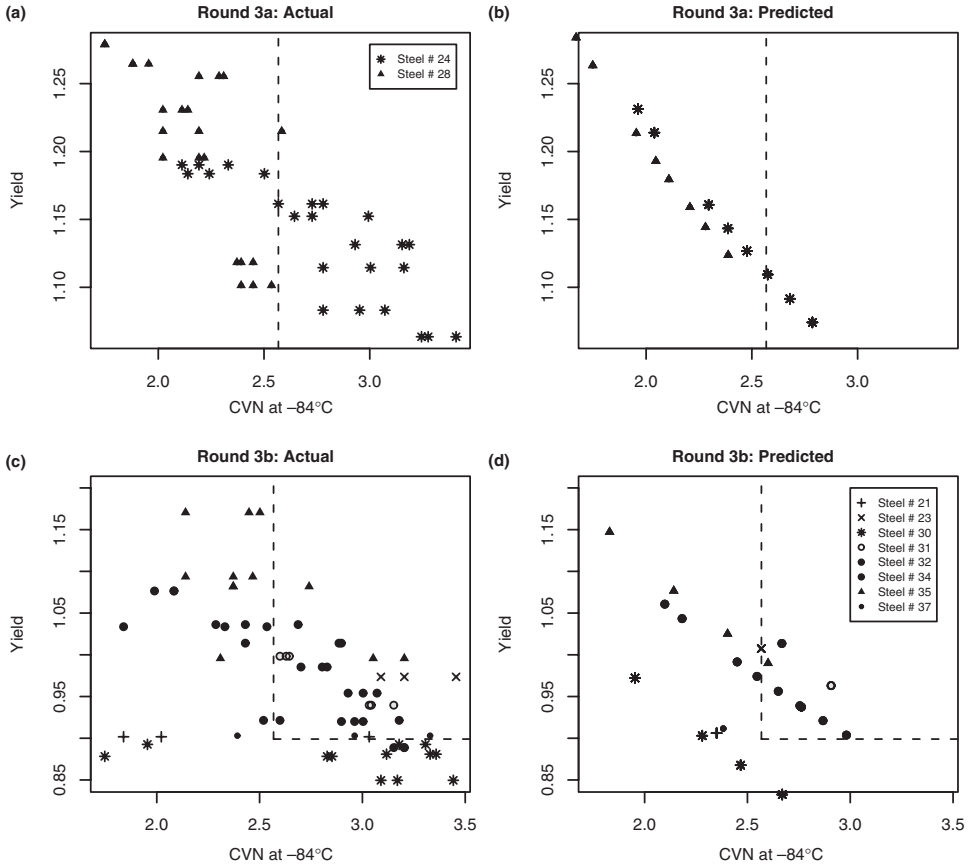

Figure 4. Round 3a (steels 24,28) and Round 3b (steels 21, 23, 30–37) results. Actual and predicted values of Yield plotted versus actual and 20% predicted values of CVN at −84˚C.

From the corresponding predicted data of figure 3(*d*), five specimens of steel 18 do not meet the toughness requirement, and two of steel 17 do not meet the strength requirement. As in the case of actual data, use of the predicted data also results in the conclusion that of the remaining specimens that meet both requirements, steels 19 and 20 exhibit the best toughness combinations. Likewise, steel 20 would be classified as superior to steel 19. The use of predicted versus actual data thus results in similar conclusions, although three additional specimens would be eliminated in Round 2.

Figures 4(*a*) and (*b*) exhibit actual and predicted data for steels 24 and 28. These steels distinguish themselves from the others in that their yield strengths generally exceed the 1.079 maximum allowed. The toughness diagram based on actual data in figure 4(*a*) allows us to conclude that all specimens of steel 28 and two specimens of steel 24 do not satisfy the toughness requirement. From figure 4(*b*), on the other hand, all specimens of steel 28 and five of steel 24 do not satisfy the toughness requirement. The use of the predicted in lieu of the actual data thus results in similar conclusions, but once again three additional specimens would be screened out in this case. In both cases (actual and predicted) there is one specimen of steel 24, with a yield strength value below the 1.079 maximum allowed. This specimen could be considered the best steel, but its thickness (19 mm) is too small for practical applications.

Consider now the remaining steels from Round 3. From figure 4(*c*), we can conclude that steels 21, 37, all specimens of 35, five of 32 and two specimens of steel 30, do not satisfy the toughness requirement. Additionally, one specimen of steel 32, and two of 30 do not meet the strength requirement. Steels 23, 34, all specimens of 31 and two of 32 satisfy both the strength and toughness requirements. Figure 4(*c*) also allows us to conclude that steel 23 is the best in Round 3. From the predicted data of figure 4(*d*) we can see that steels 23, 34, 31, four specimens of 32 and one specimen of steel 35 satisfy both the strength and toughness requirements. The most attractive steels (having the largest values of Yield) are 23 and 34. (The discrepancy in the predicted values of Yield for these steels is 0.6%, a statistically insignificant difference.)

The conclusions reached in Round 3 on the basis of model-predicted data are therefore similar to those reached on the basis of actual data. A very attractive feature of these statistical models is their ability to screen out more undesirable specimens than would be possible by conducting physical laboratory experiments. In addition, these models permit us to optimize the chemical composition and processing parameters, leading to a very efficient approach for alloy development that integrates experimental data with statistical modelling. The optimization simulations based on these regression models are presented in Golodnikov *et al* (2005b). They demonstrate that although steel 23 is very competitive, an improvement of the order of 20% for Yield or 30% for the lowest value of CVN can still (theoretically) be attained by successive refinements of the experimental process.

5. Summary and conclusions

Statistical model predictions were investigated as an aid to alloy development and selection of optimal HSLA 100 steels, as determined by their strength and toughness properties. We developed a multiple regression model which accurately predicts tensile yield strength based on its composition, plate thickness, solution treating and ageing temperatures. Because the three values of CVN at -84°C observed for each steel exhibit substantial scatter, a standard regression model that attempts to predict only the mean value was found to be a poor classifier for this characteristic. Finding that the three CVN values can plausibly be modelled as observations from the 20th, 50th and 80th percentiles of the CVN distribution, we built instead three separate quantile regression models, one for each of these percentile response surfaces. Each of these models involved different explanatory variables, in essence targeting three separate regions of the CVN distribution function. Applying analogous measures of goodness of fit as for standard regression allowed us to conclude that the quantile regression models were sufficiently accurate to be useful in the selection of good steels.

Finally, we conducted discriminative analyses of steels using the fitted regression models and compared the results with conclusions made solely on the basis of actual experimental data. This showed that the proposed statistical regression modelling approach yields results similar to those obtained by experiments. Consequently, the regression models could be used to screen out inferior alloys and identify promising compositions and processing parameters for developing better steels. Such a capability can lead to a greater reduction in the cost of experimental programmes, shifting resources for the development of optimized alloys into more cost-effective computer modelling techniques. Furthermore, the proposed computer modelling approach has the potential to aid in the design of optimized alloys by conducting additional computer simulations, as demonstrated by Golodnikov *et al* (2005b). This has important implications for the design of physical experiments targeting the discovery of new and improved alloys. Its primary merit is the incorporation of information from prior experimental data into mathematical models, which are subsequently used in refining the fabrication process.

Appendix A. Chemical compositions of the alloys

Table A1. Chemical compositions of the steel alloys (weight %). Values in each column are normalized with respect to the column mean.

Steel	C	Mn	Si	Cr	Ni	Mo	Cu	Cb	Al	N	P	S	V
1	0.86	1.16	1.43	0.78	0.32	0.46	0.51	0.85	1.35	0.76	0.83	0.69	12.02
2	0.84	1.17	NA	0.8	0.32	0.62	0.67	NA	NA	0.72	NA	NA	12.02
3	0.84	1.47	NA	1.09	0.33	0.63	0.85	NA	NA	1.03	NA	NA	23.12
4	0.79	2.01	1.53	1.05	0.32	0.62	0.83	0.82	0.9	1.5	0.83	0.79	30.83
5	0.86	0.87	1.33	0.69	0.74	0.65	0.63	0.88	1.57	0.74	0.93	0.63	0
6	0.81	0.94	NA	0.81	0.84	0.63	0.76	NA	NA	NA	NA	NA	0
7	0.84	0.94	NA	1.09	0.9	0.63	0.83	NA	NA	NA	NA	NA	0
8	0.79	1.41	1.48	1.39	1.01	1.15	0.96	0.88	1.35	0.53	0.93	0.76	0
9	1.27	1.75	0.71	0.4	0.28	0.57	0.68	0.82	0.9	0.8	0.83	0.76	0
10	1.29	1.73	NA	0.4	0.31	0.8	0.75	NA	NA	NA	NA	NA	0
11	1.31	1.74	NA	0.4	0.35	1.28	0.85	NA	NA	NA	NA	NA	0
12	1.22	1.74	0.87	0.4	0.35	1.6	0.83	0.82	0.45	0.57	0.83	0.88	0
13	0.86	0.47	0.66	0.4	0.83	0.65	0.87	0.82	0.9	0.65	0.83	0.63	0
14	0.84	0.56	NA	0.4	0.98	0.82	0.96	NA	NA	NA	NA	NA	0
15	0.79	0.74	NA	0.81	1.06	0.97	1.06	NA	NA	NA	NA	NA	0
16	0.77	0.95	0.77	1.11	1.17	1.14	1.13	0.82	0.67	0.49	0.83	0.72	0
17	1.02	1.06	1.38	1.17	0.96	0.63	0.85	1.07	0.9	0.87	0.74	0.76	0
18	0.95	1.41	1.28	1.4	1.01	1.11	0.94	1.01	1.35	0.96	0.74	0.82	0
19	0.95	0.75	0.66	0.9	1.09	0.97	1.06	1.01	1.35	0.85	0.74	0.69	0
20	0.95	0.92	0.66	1.09	1.14	1.09	1.08	1.14	1.35	0.89	0.83	0.76	0
21	0.61	0.81	0.18	0.93	1.11	1.11	1.04	0.7	0.63	0.67	0.93	0.91	0
23	1.02	0.97	0.61	1.16	1.15	1.18	1.15	0.92	0.7	1.04	1.02	0.91	0
24	0.95	0.97	0.87	1.15	1.16	1.18	1.14	0.92	1.03	1.07	1.11	0.88	0
28	1.18	1.11	1.38	1.31	1.21	1.28	1.25	1.07	1.06	1.43	1.3	1.45	0
30	0.97	0.75	0.61	0.77	1.06	0.97	0.97	0.85	0.76	0.98	1.11	1.13	0
31	0.97	0.71	0.66	0.78	1.07	0.98	0.98	0.85	0.58	1.11	1.02	1.13	0
32	0.97	0.7	0.71	0.78	1.08	1	0.98	0.85	0.43	1.09	1.11	1.2	0
34	1.36	0.86	1.33	1.01	1.11	1.05	1.06	1.45	1.26	1.64	1.39	2.08	0
35	1.31	0.88	1.89	1.01	1.1	1.06	1.06	1.49	0.99	0.98	1.76	1.76	0
37	0.99	0.72	1.28	0.89	1.05	0.91	0.97	0.88	1.12	0.88	1.11	1.45	0

Appendix B. Technical details of the quantile regression methodology

For a random variable Y with distribution function $F_Y(y) = P(Y \leqslant y)$ and $0 \leqslant \theta \leqslant 1$, the θth quantile function of Y, $Q_Y(\theta)$, is defined to be

$$Q_Y(\theta) = F_Y^{-1}(\theta) = \inf\{y \mid F_Y(y) \geqslant \theta\}.$$

For a random sample $Y_1, \ldots, Y_n$ with empirical distribution function $\hat{F}_Y(y)$, we define the θth empirical quantile function as

$$\hat{Q}_Y(\theta) = \hat{F}_Y^{-1}(\theta) = \inf\{y \mid \hat{F}_Y(y) \geqslant \theta\},$$

which can be determined by solving the minimization problem

$$\hat{Q}_Y(\theta) = \arg\min_y \left\{ \sum_{i|Y_i \geqslant y} \theta|Y_i - y| + \sum_{i|Y_i \leqslant y} (1-\theta)|Y_i - y| \right\}.$$

The θth *quantile regression* function is a generalization of the θth quantile function to the case when Y is a linear function of a vector of $k+1$ explanatory variables $\boldsymbol{x}' = [1, x_1, \ldots, x_k]$ plus random error, $Y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta}' = [\beta_0, \beta_1, \ldots, \beta_k]$. It can be written as

$$Q_Y(\theta|\boldsymbol{x}) = \inf\{y \mid F_Y(y|\boldsymbol{x}) \geqslant \theta\} \equiv \boldsymbol{x}'\boldsymbol{\beta}_\theta.$$

Estimates of coefficients of the θth quantile regression function can be found by performing the minimization

$$\hat{\boldsymbol{\beta}}_\theta = \arg\min_{\boldsymbol{\beta}_\theta} \left\{ \sum_{i|Y_i \geqslant \boldsymbol{x}_i'\boldsymbol{\beta}_\theta} \theta|Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\theta| + \sum_{i|Y_i \leqslant \boldsymbol{x}_i'\boldsymbol{\beta}_\theta} (1-\theta)|Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\theta| \right\}, \tag{B.1}$$

which can be reduced to a linear programming problem and solved via standard optimization methods. The value $\boldsymbol{x}'\hat{\boldsymbol{\beta}}_\theta$ is then the estimated θth quantile (or $100\,\theta$th percentile) of the response variable Y at $\boldsymbol{x}$ (instead of the estimated mean value of Y at $\boldsymbol{x}$ as would be the case in standard least squares regression).

References

Corowin W R and Houghland A M 1986 *The Use of Small-Scale Specimens for Testing Irradiated Material* (*American Society for Testing and Materials* ASTM STP vol 888) (*Philadelphia, PA*) pp 325–38

Czyryca E J, Link R E, Wong R J, Aylor D A, Montemarano T W and Gudas J P 1990 *Naval Engineers J.* **102** 63–81

Goldren A P and Cox T B 1986 *AMAX Materials Research Center Report* CPR-2 (Ann Arbor, MI)

Golodnikov A, Macheret Y, Trindade A, Uryasev S and Zrazhevsky G 2005a *Dept. of Industr. and Systems Eng. Research Report* #2005-1 (Gainesville, FL: University of Florida)

Golodnikov A, Macheret Y, Trindade A, Uryasev S and Zrazhevsky G 2005b *Dept. of Indust. Systems Eng. Research Report* #2005-2 (Gainesville, FL: University of Florida)

Holt J M (ed) 1990 *Charpy Impact Test: Factors and Variables ASTM STP 1072, Standard Test Methods for Notched-Bar Impact Testing of Metallic Materials* (*ASTM Book of Standards* ASTM E23-04, vol 03.01)

Ireland D R, Server W L and Wullaert R A 1975 Procedures for Testing and Data Analysis *Effects Technology Inc. ETI Report* TR75-43 (Santa Barbara, CA)

Koenker R and Bassett G 1978 *Econometrica* **46** 33–50

Koenker R and Machado J A F 1999 *J. Amer. Statist. Assoc.* **94** 1296–310

Liaw P K 2000 *Impact Toughness Testing and Fracture Mechanics* (*American Society for Metals ASM Handbook* vol 8) (*Materials Park, OH*) p 563 et seq.(esp. *Impact Testing of Metals* ASTM STP, vol 466, pp 596–611)

Lucon E *et al* 1999 *Pendulum Impact Testing: A Century of Progress* (*West Conshohocken, PA*) (*American Society for Testing and Materials ASTM STP* vol 1380) ed T Siewert and M P Manahan, pp 146–63

Metzbower E A and Czyryca E J 2002 *Proc. TMS 2002 Fall Meeting: Modeling the Performance of Engineering Structural Materials (Columbus, Ohio)* vol 3, ed D R Lesuer *et al* pp 343–55

Neter J, Kutner M H, Nachtsheim C J and Wasserman W 1996 *Applied Linear Statistical Models* (Chicago: Irwin)

Schmitt W, Bohme W and Sun D-Z 1990 *Proc. 10th Biennial European Conf. on Fracture ECF-10 (Berlin)* vol 1, ed K Schwalbe and C Berger (Engineering Materials Advisory Service) pp 159–70

Todinov M T 2004 *Nucl. Eng. Des.* **231** 27–38