Variance Stabilizing Regression Ensembles for Environmental Models

Anthony Bagnall, Ian Whittley, Matthew Studley, Mike Pettipher, Firat Tekiner and Larry Bull

Abstract—This paper describes linear regression models fitted for the 2006 predictive uncertainty in environmental modelling competition hosted at the WCCI 2006 conference. Entries into this competition are required to produce models of up to four non-linear regression problems. Rather than adopt a complex non-linear modelling technique, our approach is to fit linear models to transformed data, with adaptive methods used for setting parameters and estimating error. This paper describes several techniques popular with statisticians which are less well known in the computational intelligence community, then proposes new ways of using these statistics. We describe standard statistical transformation techniques, Yeo-Johnson and Box-Tidwell, and present stepwise algorithms for using these transformations on large data sets. These stepwise algorithms utilise the Ancombe procedure, runs tests on residuals, the Goldfeld-Quandt procedure and the Kolomogorov-Smirnoff test for normality. We combine these statistics with the transformation procedures to form a piecewise linear approach to environmental modelling.

I. INTRODUCTION

As part of the EPSRC sponsored grant super computer data mining (SCDM) we have developed classification and regression tools for use on High Performance Computing facilities (such as CSAR at the University of Manchester). Our main interest involves classification and regression problems on data sets with a large number of attributes and many cases. Regression involves forming a statistical model that maps some set of regressors onto a continuous response variable. The objective of regression modelling is usually two-fold: firstly, to be able to predict the response value of new cases based on observed regressor values; and secondly to explain the variation in the response in terms of the values of the regressors. Linear regression is the oldest and the simplest modelling approach. It is based on the three core assumptions of: linearity in the relationship between regressors and response mean; normality; and constant variance (homoscedasticity). When these assumptions hold, the linear model is optimal. However in most cases these strong assumptions do not hold. Throughout the history of regression research, a succession of complex modelling techniques have been proposed to overcome the shortcomings of linear regression. Statisticians have developed parametric methods such as generalised linear modelling and generalised least squares, and non parametric methods such as kernel smoothing. Concurrently, the AI community has proposed numerous algorithms using, for example, neural networks, regression trees and support vector machines. There is no doubt that the more complex algorithms can fit much better models. However, the greater predictive power sometimes comes at the cost of lack of interpretability; explaining the relationship between the regressors and the response is non-trivial with a neural network or a kernel technique. We believe than any case study of a regression problem should start with linear modelling, if necessary used in conjunction with transformations and a sensible analysis of the residuals. Residual analysis can identify violations of the regression assumptions. These violations can often be minimized through the transformation of the regressors and/or the response variable. There are many reasons to transform data as part of a regression analysis: to achieve linearity; to achieve homogeneity of variance, that is, constant variance about the regression equation; and to achieve normality (or at least symmetry) about the regression equation. One of the aims of this paper is to present to the neural network community how to best use linear regression in conjunction with standard transforms from the statistical literature.

The majority of statistical research focuses on using transformations on small problems with few regressors. The decision of which transformation to perform is usually made based on expert evaluation involving graphical techniques. We propose an automated method of constructing a model by greedily selecting variables for transformation. The algorithm could use many statistical tests to determine the quality of the fit. Since the objective of the competition is to accurately estimate the variability of predictions, we concentrate on producing models with constant error variance rather than focusing only on minimizing mean squared error.

The SCDM project involves implementing ensembles of regressors for large data sets on computer clusters of up to 512 processors. There are three main benefits of using linear models in ensembles. Firstly, they are fast to fit. The individual processors of HPC facilities are commonly no faster than desktop computers, and the most efficient way of ensembling is with an individual regressor on each processor. Hence, for large data, a fast model building process is desirable. Secondly, there is little variation in the time it takes to fit the model. Simple ensembles are only as last as their slowest member, and algorithms with a large variation in training time will slow down the ensemble. Thirdly they are easy to interpret. A common criticism of ensembles is that it is hard to understand the relationship between the regressors and the response. Using ensembles of linear regressors can make it easier to interpret the importance of regressors. Interpreting the importance of regressors in a model is of particular importance in environmental data sets. In this domain, the objective is usually a better understanding of the underlying processes and prediction is commonly of secondary interest.
The aim of this paper is to produce models that are comprehensible in terms of the problem domain with stable variance. We do not necessarily expect to produce the best models. However, we do hope to produce small, easily understood models that go some way to explaining the variation in the response in terms of functions of the residuals. Our algorithm design choices are driven by a desire for clarity and parsimony.

The rest of this paper is structured as follows. In Section II we provide background into regression modelling and describe common diagnostics. In Section III we outline popular transformation schemes to remove multi-collinearity and heteroscedasticity. In Section IV we detail two ways of using these transformation schemes for problems with a large number of regressors, the first being a sequential greedy algorithm and the second a distributed ensemble approach. In Section V we present the results of these algorithms on the competition data sets and in Section VI we conclude and highlight the next steps.

II. LINEAR REGRESSION ALGORITHM

In this Section we briefly review the salient features of model fitting with linear regression. Further information can be found, for example, [1]. Assuming we have $n$ observations of $m$ regressor variables $x_1, \ldots, x_m$ and response variable $Y$, then the linear regression model is

$$Y = X\beta + \epsilon,$$

where $Y$ is the $n \times 1$ vector of response observations, $X$ is the $(n \times m + 1)$ regressor observations (including a constant term), $\beta$ is the unknown parameters and $\epsilon$ is the error terms. The least squares estimators of the parameters $\beta$,

$$B = (X^T X)^{-1} X^T Y,$$

are optimal (in the sense that they are identical to the maximum likelihood estimators) if the $n$ random variables $\epsilon^T = (\epsilon_1, \ldots, \epsilon_n)$ are independent normal random variables with constant variance. However, if any of these assumptions are violated then the linear model is often a very poor choice of model. Violation of these assumptions can be caused by many factors. The underlying relationships may be intrinsically non-linear, or there may be mixture model effects overlapping. In this case a non-linear model will be more appropriate. Alternatively, there may be a linear relationship in an alternative feature space. In this situation a transformation of the variables may yield a linear model where the regression assumptions hold. This transformation may be performed implicitly through a kernel function with a support vector machine, or explicitly through the derivation of new variables. In this paper we concentrate on methods for automatically transforming variables to improve the fit while maintaining as parsimonious model as possible. Another issue effecting the model is the selection of regressors. Correlation between regressors, commonly called multi-collinearity, can cause instability in the inverse matrix $(X^T X)^{-1}$ and the inclusion of redundant terms can inflate the error estimate.

The selection of transformations and regressors requires a measure of quality of a model. The simplest headline statistic to assess model quality is the sum of squared error, SSE,

$$SSE = (Y - \hat{Y})^T (Y - \hat{Y})$$

where $\hat{Y} = XB$. However, the SSE does not help us determine whether the assumptions of normality, independence and constant variance have been violated.

A. Diagnostic Techniques

In this Section we briefly describe some common means of assessing linear models. Regression diagnostics are based on the residuals $\epsilon_i = y_i - \hat{y}_i$. If the regression assumptions hold, the error terms will be Gaussian. However, the residuals are not normally distributed, nor are they independent. It is usual to standardize the residuals. Let $h_{ii}$ be the diagonal terms of the $(n \times n)$ Hat matrix

$$H = X(X^T X)^{-1} X^T$$

the standardised or studentized residuals are then defined as

$$t_i = \frac{\epsilon_i}{s \sqrt{(1 - h_{ii})}}$$

The studentised residuals have zero mean and unit variance, but are not normally distributed. Further transformations can be performed to approximate normality if necessary (for example, Theil’s BLUS procedure [2]). However, the studentized distribution will be regular, and significant significant deviation from the standard normal distribution provides evidence of assumption violation. Broadly speaking, there are three types of deviation we are attempting to detect with residual plots, which we demonstrate with the residuals of the full linear model for the four competition data sets shown in Figure 1.

![Residual plots](image)

Fig. 1. Residuals plotted against fitted values for the standard linear model

All four residual plots show deviation from normality. The temperature data is the most regular and the deviation is hard to detect visually. There appears to be equal variation about
the mean, but with several residual large values and a possible indication of increasing variance. This type of deviation is the classic case of cone-shaped heteroscedasticity.

Since both SO2 and Precipitation are strictly positive the residuals are bounded below, hence the linear minimum in the residual plot. This feature of the data often indicates that a transformation of the response variable may improve the regression. The two plots also both exhibit heteroscedasticity, although with different functional forms: Precipitation shows a steadily increasing variance, whereas SO2 shows variance at a maximum in the mid range of the fitted values. Finally, the synthetic residuals show decreasing variance and a clear pattern. This pattern of residuals often indicates model misspecification, which can sometimes be mitigated against by transformation of response and/or regressors.

This informal analysis highlights why these data sets present an interesting challenge. There are three types of deviation we are attempting to detect. Firstly, increasing or decreasing variance as demonstrated by precipitation and temperature. Secondly, patterned variation shown by SO2 and synthetic. Finally, we also want to detect non-normality, which is apparent in SO2, precipitation and synthetic. Our main concern is in identifying whether the distribution of the residuals is changing with different response values. Several methods for detecting patterns in these graphs have been proposed. We describe four test statistics commonly used to determine whether there is significant deviation of this kind. An excellent description of the type of tests we are looking at can be found in [3].

1) The Anscombe procedure for Increasing/Decreasing Variance (AP): The Anscombe procedure [4] (AP) involves measuring the correlation between the response variable and the absolute or square of the residuals. If $\hat{y}_i$ are the fitted values, $|e_i|$ the absolute residuals, and $\bar{y}$ and $\bar{e}$ the mean fitted value and absolute response respectively, then the AP statistic we use is defined as

$$
\rho_{e} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(|e_i| - \bar{e})}{\sqrt{(\hat{y}_i - \bar{y})^2(|e_i| - \bar{e})^2}}
$$

$\rho_{e}$ can detect increasing or decreasing variance as shown by Temperature and Precipitation, although it may not detect heteroscedasticity of the form shown by SO2, since the overall correlation may be low. When the assumption of homoscedasticity is true, $\rho_{e}$ can be transformed into a variable with a student T distribution.

2) Runs Test for Pattern in Residuals (RT): Patterns in the residual plots (which may not be detected by AP) may be detected by a test of independence. Parametric tests such as Durban-Watson or other autocorrelation tests can be used. However, simpler non-parametric tests can also be useful. We use a statistic derived from the runs above and below the mean. Let $u$ be the binary series of above and below 0 for the residuals, where

$$
u_i = \begin{cases} 
1 & \text{if } e_i > 0 \\
0 & \text{if } e_i \leq 0 
\end{cases}
$$

The runs count $r$ is the number of times the series $u$ changes value, i.e.,

$$
r = \sum_{i=2}^{n} r_i
$$

where

$$
r_i = \begin{cases} 
1 & \text{if } u_i \neq u_{i-1} \\
0 & \text{if } u_i = u_{i-1} 
\end{cases}
$$

When the assumption of independence is true, a statistic of $r$ follows an approximately normal distribution.

3) Goldfeld-Quandt Procedure for measuring effect of regressors (GQ): The Goldfeld-Quandt [5] (GQ) involves removing sections of the data then comparing the SSE on regressions of different sections of the regressor space. Goldfeld-Quandt can be used to assess the influence of individual regressors on heteroscedasticity. Given a set of $p$ attributes denoted $S_k$ to assess whether attribute $x$ is causing the variability, the data set of size $n$ is first sorted by $x$, then split into three by order, $k_1$, $k_2$ and $k_3$ of size $n_1$, $n_2$ and $n_3$. Generally $n_1 = n_3$ and $n_2 = n/4$. Two regressions are performed on $k_1$ and $k_3$ and sum of square residuals $s_1$ and $s_3$ are found. Under the null hypothesis of homoscedasticity, the statistic given by

$$
q = \frac{(n_3 - p - 1)s_1}{(n_1 - 1)s_3}
$$

follows an $F(n_1 - p - 1, n_3 - p - 1)$ distribution. Hence $q$ can be used to quantify how much a single variable is contributing to heteroscedasticity, and the $q$ values for several variables can be used to rank the variables influence on the inflating variance.

4) Kolmogorov-Smirnoff Test for Normality (KS): If we consider the standardised residuals as a univariate series of independent observations, we wish to test whether there is significant deviation from the expected pattern of values one would expect from observations from the standard normal. There are many tests for normality, most of which are based on regression versus expected quantiles (e.g. Shapiro-Wilk, Filiben and D’Agostino statistics) or goodness of fit tests (e.g. Chi-square, Kolmogorov-Smirnoff and Durbin tests). We use the Kolmogorov-Smirnoff (KS) test statistic as a measure of normality. KS is a suitable statistic to use when the test variance is known. To measure the KS statistic for residuals $r$, first sort $r$ to obtain the order statistics $d$. The find the maximum deviation between the observed cumulative distribution function and the theoretical cumulative distribution function under the assumption of normality.

One of the contributions of this paper is to use the Kolmogorov-Smirnoff, Anscombe, Goldfeld-Quandt and runs test in conjunction with the MSE to determine whether to transform variables and select variables for inclusion.
The transformations we use are described in the following Section.

III. SINGLE VARIABLE TRANSFORMATIONS

We consider the family of power transforms of both the response and the regressors.

A. Response Transformations

Heteroscedasticity can often be removed through transformation of the response variable. The Box-Cox transform [6] described in Equation 1 is usually adopted.

\[
y^\lambda = \begin{cases} 
  \log(y) & \text{if } \lambda = 0 \\
  \frac{y^{\lambda} - 1}{\lambda} & \text{otherwise }
\end{cases}
\]

(1)

Lambda is normally set as the value that minimises the MSE for a given set of regressors, usually found by an interval search on a reasonable range, for example $[-3, 3]$. The MSE must be calculated after inversion back into the original units for the comparison to be meaningful. Box-Cox is only applicable to strictly positive series. A recently proposed alternative is the Yeo-Johnson transform [7], given in Equation 2.

\[
\phi(\lambda, y) = \begin{cases} 
  \frac{\log(y + 1)}{(y+1)^{\lambda-1}} & \text{if } \lambda = 0, y \geq 0 \\
  \frac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\
  -\frac{\log(1-y)}{(1-y)^{\lambda-1}} & \text{if } \lambda = 2, y < 0 \\
  -\frac{1}{2}\frac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 2, y < 0
\end{cases}
\]

(2)

It is our objective to keep the final model as interpretable as possible. To that end, we always round any transformation we perform to the nearest 0.5 under the assumption that $y^{0.5}$ has a clearer interpretation than, for example, $y^{0.432}$.

B. Regressor Transformations

We consider three types of transformation in the regressor/attribute space. Attribute selection, i.e. deciding on whether to retain a regressor in the model, feature creation, i.e. combining several attributes to form a new regressor and transformation, i.e. taking a function of an attribute.

We perform attribute selection using a simple stepwise regression procedure to minimize the Bayesian Information Criterion (BIC). Variables are added to the model using a greedy algorithm that selects the attribute that most decreases the SSE. A variable is only added if it significantly improves the BIC. After each variable is added, tests are performed to see if variables already in the model can be removed.

Principle component analysis is used as the feature creation mechanism. This serves to remove the problem of colinearity and reduce the dimensionality of the problem. PCA is a popular transformation in environmental sciences so it can efficiently summarising interacting factors.

We consider the family of Power functions to transform response variables. The Box-Tidwell algorithm [8] is based on the following observations. Consider the model

\[
y = \beta_0 + \beta_1 x_1^\alpha + \epsilon
\]

the problem then is to estimate $\alpha$. If we consider an estimate of $\alpha$, say $\alpha_0$, the Taylor expansion of $x^\alpha$ around $\alpha_0$ is

\[
x^\alpha = x^{\alpha_0} + (\alpha - \alpha_0)x^{\alpha_0} \log(x) + O((\alpha - \alpha_0)^2).
\]

as $\alpha_0$ approaches $\alpha$, the term $O((\alpha - \alpha_0)^2)$ gets vanishingly small. Thus if we consider the model

\[
y = \beta_0 + \beta_1 x_1^{\alpha_0} + b_2 x_2^\alpha \log(x) + \epsilon
\]

where $b_2 = \beta_1 (\alpha - \alpha_0)$, then an updating equation is given by $\alpha = \frac{b_2}{\beta_1} + \alpha_0$. This leads naturally to the iterative algorithm described by Algorithm III-B. The Box-Tidwell algorithm can only be applied to strictly positive variables. Also note that it does not always converge, and is sensitive to initial conditions.

Algorithm 1 Box-Tidwell algorithm for a single attribute

1) initialise $\alpha_0$
2) initialise $i = 0$
3) while not(finshed) do
   a) Fit Line $E(y_i) = b_0 + b_1 x_1^{\alpha_i}$
   b) Fit Line $E(y_i) = b_0 + b_1 + b_2 + x_1^{\alpha_i} + x_2^\alpha \log(x)$
   c) Update $\alpha_{i+1} = \alpha_i + \frac{b_2}{b_1}$
   d) if $|\frac{b_2}{b_1}| < TOL$ finished =true

IV. DATASET TRANSFORMATIONS

In section II we described four measures of fit to assess the suitability of a particular model (MSE, KS, GQ and AP). In section III we described how individual variables could be transformed to improve that fit. Two observations need to be made about these tests and procedures. Firstly, the tests described in Section II are by no means infallible and can lead to different conclusions. For example, a test built on GQ may reject the null hypothesis of homoscedasticity when an AP based test will not. Secondly, the transformations described in Section III are for single variables only. Assessing combinations of transformations simultaneously is much more complex. Several approaches to this problem have been proposed. Generalized additive models (GAM) [9] and the alternating conditional expectation (ACE) technique [10] are non-parametric iterative transformations that minimize the error variance by single-function minimizations. However, GAM and ACE are not suitable for our objectives, as it does not produce an explicit (i.e. parameterised) transformation.

Our objective is to produce simple, comprehensible models that can give insights into the problem or at least provide benchmark results for more complex techniques. We propose two algorithms for finding the minimal, best, transformed linear regression. We use both algorithms with the Principle Components, retaining only the components that explain the top 95% of the variation, although this is not essential. In a genuine analysis, we would attempt to gain expert confirmation of the components prior to modelling with them. However, it is common in environmental modelling.
to combine attributes in this way, and it helps the modelling by removing co-linearity.

We denote the independent variables $x_i$ and the response $y$. The Box-Tidwell or Yeo-Johnson transformation in the context of a given model is denoted $\tilde{x}_i$ and $\tilde{y}$.

The greedy sequential algorithm described in by Algorithm IV forms the core fitting technique used in the model building described in Section V. As with all stepwise approaches, we iteratively add in attributes until no improvement is found, at each step attempting to remove attributes that have become redundant. We start with an empty model (in the interests of parsimony).

**Algorithm 2 Stepwise Regression Procedure with Iterated Transformations**

1. input candidate attributes $X$, transformation techniques and significance test.
2. initialise selected attributes $S = \emptyset$
3. while not(finished)
   a. Choose regressor in $x \in X$ that most improves the model $S$ with either $y$ or $\tilde{y}$ after transformation to $\tilde{x}$
   b. If model $S \cup \tilde{c}$ is significantly better than model with $S$
      i. $S = S \cup \tilde{x}$
   else
      i. finished = true
   d. Removing any $\tilde{x}' \in S$ that does not significantly decrease the quality of the model

The key features in defining this algorithm are determining how we perform the transformations (including a metric for selecting a parameter) and the related issue of how we rank the transformed attributes. The traditional approach is to select attributes based solely on the reduction in the SSE. However this often leads to the inclusion of variance inflating regressors. Our novel approach is to use four criteria for variable selection. These are p-values derived from the statistics $s$ (estimate of standard deviation), $\rho_{\epsilon}$ (correlation between predicted and residuals), $r$ (run count of residuals), $q$ (Goldfeld-Quandt ratio) and $k$ (maximum CDF deviation). We only consider a variable for inclusion if a null hypothesis concerning the regression assumptions based on the distribution of these test statistics cannot be rejected at the 10% level. Thus we only include a variable in the model if it can be firstly shown to significantly improve the SSE, and secondly if there is no evidence that the inclusion increases the heteroscedasticity. Note that the level of the test is currently set roughly to illustrate the process. Properly, we should make allowances for the fact we are performing five tests on each attributes and we are testing multiple regressors.

V. RESULTS

We use the competition data to illustrate the benefit of using these statistics, and to develop a new approach to piecewise regression based on minimizing the violations from the assumed distribution rather than simply minimizing the SSE. Table I shows the regression results using linear stepwise regression (no transformation), stepwise transformation algorithm IV using just SSE as the entry criteria and stepwise transformation using combined test described in Section IV (stepwise combined). Firstly, Table I demonstrates that the transformations result in more accurate models. The validation set SSE is lower with all four data sets. However, this is not as a useful conclusion as may be initially thought. The test set residual analysis summarised by the descriptive statistics indicate changing variance and probable model specification. This is also indicated by the stepwise approach using the more stringent entry condition. With SO2 and Temp no variables are fitted at all, and only four enter the model with Precipitation.

**Table I**

<table>
<thead>
<tr>
<th>Temperature (PCA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>Stepwise Transform</td>
</tr>
<tr>
<td>Combined</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Precipitation (PCA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>Stepwise Transform</td>
</tr>
<tr>
<td>Combined</td>
</tr>
<tr>
<td>SO2</td>
</tr>
</tbody>
</table>

| Stepwise Transform | 0.855 | -0.055 | 0.195 | -4.364 | 1.0326 | 12   |
| Combined           | 0.84 | -0.053 | 0.196 | -2.68 | 1.0333 | 11   |

It is possible that a more sophisticated transformation search technique could yield better models for this data. However, particularly for SO2 and Precipitation, we would argue that a global linear model is not appropriate. Indeed, the significant deviation from the normality assumptions could be used as a justification for a non-linear modelling algorithm. Our focus instead remains with using transformations with linear models and demonstrating how the diagnostics could be used to guide this process. We wish to combine different linear models formed on alternative subsets of the data to form more accurate, stable predictions. To construct an ensemble it is necessary to diversify the output of each model. To do this we can alter the model fitting technique and/or alter the data set each model is formed on. To diversify between members of the ensemble we give each element an alternative criteria stopping criteria in the Stepwise procedure given by Algorithm IV. The best selected attribute is added only if it significantly reduces the SSE and does not increase heteroscedasticity as measured by either AP, KS, Runs or GQ. There are two ways we could split the data.

1) **Splits on the attribute space.** We could partition the attributes space and form a regression tree such as CART [11]. The GQ statistic could easily be adapted to decide on regressors to branch on. This transformation would be most beneficial when an overall mis-
specification is indicated by the diagnostic statistic. However, attribute splits are by definition mutually exclusive and hence not particularly suitable for ensembles, which usually built on overlapping attribute spaces.

2) **Splits on the response space.** The diagnostics that indicate dependency between residuals could be indicative that different regions of the response space are best fit by different linear models. This is more suitable for use with ensembles, as it will produce diverse models that may also combine to produce better predictions.

We perform splits on the response space with the objective of producing intervals where none of our regression tests fail. First we sort the training data by response value. We then split the data set into subsets containing an 5% of cases. Starting with the subset with the lowest response values, we fit a model with Algorithm IV. If this model passes our residual tests, we merge the current subset with the next subset and refit. This continues until the addition of an extra data set causes one of the tests to fail. In this way we hope to create regions of the response space where the transformation technique can linearise the problem.

To form a prediction, we obtain a mean estimate from each of our models. For the majority of the models, these estimates will be extrapolations, i.e. will fall out of the range of the test data. We average the mean estimates of the interpolating models, or, if there are no interpolating estimates, we weight the predictions by the distance from the centre of the interval.

The preliminary results demonstrate the piecewise approach can produce a more accurate model, and give insights into the problem. Table II shows the splits produced on the Temperature data. Very simple models with large MSE are fitted at in the middle values, whereas more complex models are fitted around the extreme values. This indicates it is more complex to estimate extremes of temperature. The models fitted were as follows:

1) \[
\log(y) = -2.46 + 0.38 \sqrt{x_5}
\]
2) \[
y = -1.36 + 0.02x_1^{3/2} - 0.04x_{10}
\]
3) \[
y = -1.05 + 0.005x_2 + 0.1x_3^{3} + 0.016x_{10}
\]
4) \[
y = -0.4 + 0.1x_5^{3/2}
\]
5) \[
y = 0.81 - 0.06x_1 - 0.018x_5^{3} - 0.032x_9
\]
6) \[
\dot{\phi}(y,x,5) = 2.22 - 1.5 \log(x_3) + 0.62 \log(x_5) - 0.04x_9^{3/2} - 0.15x_{10} + 0.046x_{11}
\]

| Table II: Ensemble Partition of the Response Space for Temperature. |
|--------------------|--------------------|-----------------|-----------------|
| **Min Value** | **Max Value** | **Size** | **MSE** |
| -1.71 | -1.71 | 313 | 0.45 |
| -1.71 | -1.05 | 825 | 0.035 |
| -1.05 | -0.79 | 524 | 0.02 |
| -0.79 | 0.04 | 2030 | 0.053 |
| 0.04 | 1.22 | 2588 | 0.07 |
| 1.22 | 2.52 | 846 | 0.05 |

The simple ensemble used to generate the models given in Table II can be extended to include alternative inclusion criteria. A generalised ensemble algorithm is described in Algorithm V. Although as yet untested, we would hope this algorithm would produce more stable results with reduced error.

**Algorithm 3 Ensemble Regression**

**Train Ensemble**

1) For each subset
   a) Produce 4 models using the Stepwise Algorithm IV and one of AP, KS, Runs or GQ stopping condition
**Predict Ensemble**

1) For each new case
   a) for each subset Combine predictions and variance estimates from each model to form
   b) Combine the predictions from each subset, inverse weighted by the distance of the prediction from the range of the subset.

**VI. CONCLUSIONS**

A commonly employed method of handling issues like heteroscedasticity is to transform the data to stabilize the variance. Support vector machines implicitly search the transform space with the objective of linearising the problem. However, there has been less focus on the analysis of residual distributions in the SVM and NN community than in statistical research. The first objective of this paper is to highlight to the NN community how statistical tests can help guide the model building process. In computational intelligence research, parameters are often set by grid search or other computationally intensive algorithms, and overfitting is mitigated against by extensive cross validation, which is also demanding of computing resources. This could be somewhat mitigated against by a shift of emphasis towards attempting to stabilise the problem (as measured by the type of statistics we are suggesting) rather than simply minimizing SSE.

The second objective is to propose a stepwise algorithm that combines two commonly used transformation procedures to form a transformed model. This algorithm produces models that are better on the competition data sets than linear models fitted in the untransformed space. However, we also demonstrate that the residuals of transformed models also exhibit significant deviation from independence and normality. A revised stepwise approach that adopts stricter entry criteria for variables based on the four statistics described in Section II illustrate that the algorithm is not finding a linearising transformation. This may indicate that a linear model in the transformed Box-Tidwell/Yeo-Johnson space is in fact inappropriate for this data. This negative result could form the justification for using a non-linear modelling algorithm. Alternatively, it could indicate that a piecewise approach would improve results. We propose two ways of splitting the data to fit piecewise models using the statistics described: The first performs splits on the attribute space
based on the heteroscedasticity of introducing variables. The second split the data set by the response variable values.

ACKNOWLEDGMENT
This research is supported by EPSRC grants GR/T18479/01, GR/T18455/01 and GR/T/18462/01.

REFERENCES