International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 009-017

Estimating Those Transformations That Produce the Best-fitting Additive Model: Smoothers Versus Universal Approximators

D. A. de Waal[1], S. E. S. Campher[2] and J. V. du Toit[3]

[1]SAS Institute Inc., Cary NC, 27513, USA
Andre.DeWaal@sas.com

[2]IM Management Information, North-West University, Potchefstroom Campus,
Private Bag X6001, Potchefstroom, 2520, South Africa
Susan.Campher@nwu.ac.za

[3]Department of Computer Science and Information Systems, North-West University,
Potchefstroom Campus, Private Bag X6001, Potchefstroom, 2520, South Africa
Tiny.DuToit@nwu.ac.za
(Corresponding author)

***Abstract*: When estimating a generalized additive model, a crucial decision that must be made is the choice of underlying technique that will be used to estimate those transformations that produce the best-fitting model. Data smoothers and universal approximators are two opposing techniques that seem to hold the most promise. ACE (alternating conditional expectations) was developed by Breiman and Friedman and utilizes a super-smoother to determine conditional expectation estimates. It was intended to be used as a tool to estimate the optimal transformations for multiple regression problems. Generalized additive neural networks on the other hand depend on the use of universal approximators to compute the nonlinear univariate transformations for the independent variables. These two approaches are compared and illustrated with a suitable example from the literature.**

***Keywords*: Generalized additive neural networks, Smoother, Universal approximator.**

I.Introduction

The procedure of alternating conditional expectations, or ACE in short, was developed by Breiman and Friedman [1] and was intended to be used as a tool to estimate the optimal transformations for multiple regression problems and to use the transformations to build additive models. It is a non-parametric procedure that utilizes data smoothers to estimate the conditional expectations. ACE has been praised as a novel andremarkable achievement [2], and as a powerful tool that brings objectivity to the area of variable transformations in data analysis[3].

Generalized additive neural networks (GANN) were first proposed by Sarle [4] when he investigated the relationship between neural networks and statistical models. A GANN is the artificial neural network implementation of a generalized additive model (GAM) and it uses a separate multilayer perceptron(MLP) to model each univariate transformation. As MLPs are universal approximators capable of modeling any continuous function [5], a GANN should in principle be able to approximate any additive model. When combined with the correct choice of link function, a GANN can be used to estimate any GAM. At least two algorithms exist that may be used to construct GANNs, namely the iterative algorithm of Potts [6] and the automated algorithm of du Toit [7]. GANNs are a more recent development than ACE and therefore not that well known.

The goal of both methods are the same, namely to construct those transformations that produce the best-fitting additive model. Both methods are non-parametric as no assumptions about the structure of the relationship between the independent variables and the dependent variable are made. The most obvious difference between the two approaches is in the use of smoothers versus universal approximators.

A detailed comparison between ACE and GANNs (and the AutoGANN implementation) can be found in [8]. In her thesis, she used three simulated examples as well as three observed data sets that have been used in the literature relating to data analysis. In this paper, one of the three observed data sets, namely the Boston Housing data set [9] will be used as a running example.

The aim of this paper is not to repeat all the results of this thesis, but to provide a high-level overview of the two approaches, to highlight some important differences between the approaches and to discuss some advantages and disadvantages of each approach. This article therefore focuses on the core algorithms and not newly developed variations or suggested improvements.

The rest of the paper is organized as follows. Section II contains a highlevel description of ACE. GANNs are explained in Section III. Section IV contains arunning example and Section V an extended discussion. The article ends withsome conclusions.

II.Alternating conditional expectations

The ACE algorithm [1] is conceptually very simple and elegant. Consider the multivariate case of n observations of the form (y_i,x_i), where $\boldsymbol{x}_i = (x_{1i},x_{2i},\ldots,x_{pi})$ is a vector of p predictor variable observations. When it is assumed that the response variable $Y$is dependent on the predictor variable $\mathbf{X}$= $(X_1,X_2,\ldots, X_p)$ by a relation of the form

$$Y = f(\mathbf{X}) + \varepsilon, \tag{1}$$

the value $f(\mathbf{X})$ is considered to be the conditional expectation of Y given $\mathbf{X}$.

While estimating $f(\mathbf{X})$ as a multivariate function could prove to be challenging, additive models simplifies the problem - it involves the estimation of p one-dimensional functions of the components of $\mathbf{X}$:

$$f(\mathbf{X}) = \sum_{j=1}^{p} f_j(X_j). \tag{2}$$

The ACE procedure now uses a super-smoother to estimate the conditional expectations. [10] describes the super-smoother as a variable span smoother on linear fits using cross-validation to determine the optimal span. The super-smoother uses local averaging by fitting a least squares straight linethrough neighboringpoints of an observation, x_i. The value of the linear fit at x_i is taken to be the measure of average for the $y$values in the neighborhoodof x_i.

The reason for using a least squares straight line rather than a simple average is twofold. Firstly, the x-values are almost never equally spaced in practice. Using a simple average will not reproduce straight lines.Secondly, a simple average calculation will be less accurate near the boundaries of the x-domain as it is not possible to keep the span symmetric.

A key feature of the super-smoother is the automatic calculation of the span at each x-value. Usually, the analyst chooses the best span to use for the specific problem and applies it over the entire problem domain. Using a constant span over the entire domain is not optimal, especially in cases of heteroscedasticity or where the second derivative of $f$changes over the domain [10]. The optimal span to use at each x-value (as well as the corresponding smooth value) may be obtained by selecting that span that minimizes an estimate for the expected squared error.

The ACE procedure creates a model of additive form and maps each one-dimensional function to a variable transformation. In addition, it also transforms the response variable, Y.

III.Generalized additive neural networks

Generalized additive models involves the estimation of p one-dimensional functions of the components of

$$f(\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} f_j(X_j) \tag{3}$$

where $f(\mathbf{X})$ is considered to be the conditional expectation of Y given $\mathbf{X}$.

The generalized linear model (GLM) was introduced by [11] and involves the addition of a link function, g, to the linear model. The link function relates $f(\mathbf{X})$ to the expected value of the response variable,$\mu=E\{Y\}$:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j(X_j). \tag{4}$$

The link function is usually assumed to be known - in the case of the linear logistic model, for example,

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) \tag{5}$$

and the response value is bound by zero and one. Other types of link functions are the identity, inverse hyperbolic tangent and the log link functions.

A generalized additive model (GAM) combines the generalized linear and additive models by transforming the generalized linear model to a generalized sum of (potentially) non-linear functions:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j(X_j). \tag{6}$$

Hastie and Tibshirani [12] introduced the generalized additive model where each function f_i is estimated with a scatter plot smoother, s_i, so that the model becomes a (nonparametric) sum of smooths. They proposed an iterative smoother procedure called the *local scoring algorithm*. It involves the repeated fit of an additive model using the backfittingalgorithm.

A more recent development is the use of multilayer perceptrons to estimate the univariate functions [6], [4]. A separate MLP with a single hidden layer of h units is used for each variable:

$$\begin{aligned} f_j(x_j) &= w_{1j}\tanh(w_{01j} + w_{11j}x_j) + \cdots \\ &+ w_{hj}\tanh(w_{0hj} + w_{1hj}x_j) \end{aligned} \tag{7}$$

and h could vary across inputs.

To ensure that the linear model is a special case, a skip layer may also be included:

$$f_j(x_j) = w_{0j}x_j + w_{1j}\tanh(w_{01j} + w_{11j}x_j) + \cdots + w_{hj}\tanh(w_{0hj} + w_{1hj}x_j). \quad (8)$$

Sigmoid functions like the hyperbolic tangent and logistic functions are the most popular activation functions for MLPs, as they are bounded, monotonically increasing and differentiable [13].

A GANN model has the form:

$$g(E\{Y\}) = w_0 + \sum_{j=1}^{p} f_j(X_j) \quad (9)$$

Where

$$f_j(X_j) = w_{0j}X_j + \sum_{k=1}^{h} w_{kj}\tanh(w_{0kj} + w_{1kj}X_j). \quad (10)$$

This equation clearly represents the form of the generalized additive model. When each w_{kj} is equal to zero, it simplifies to the form of the generalized linear model.

In the original formulation of GANNs by Potts [6], the link function was chosen from a list of available link functions depending on the properties of the dependent variable in the chosen data set. This will restrict the flexibility of the link function to a few default functions, which is not optimal [8]. But, as the output activation function (the inverse of the link function) is also a univariate function, it can be approximated with a universal approximator (MLP) in a similar way as the univariate transformations for the independent variables. This improvement to the basic GANN architecture was implemented in the AutoGANN system and this functionality is exploited in the final example in Section E.

Backfitting is unnecessary for GANNs and any method suitable for fitting of MLPs (such as Newton-type methods) can be used to simultaneously estimate the parameters of the model.

A recipe for constructing GANNs based on the inspection of partial residual plots was given in Potts [6]. The algorithm was automated by du Toit in his thesis on generalized additive neural networks [7].

The problem of constructing the best GANN now reduces to a GANN architecture selection problem (determining the optimal number of nodes, *h*, for each variable – *h* could vary across inputs). In the AutoGANN modeling node [14], [15] in SAS® Enterprise Miner™, informed search, heuristics and special operations from genetic algorithms are combined to find the best GANN model as quickly as possible, based on objective model selection criteria or cross-validation error. Examples of objective model selection criteria are the Schwarz Bayesian Criterion (SBC) and Akaike Information Criterion (AIC). The results of an analysis of the significant relationships between the independent variables and the target variable are used to construct a GANN model that becomes the root of a search tree of potential GANN models. The search continues until a time limit is reached or until the search space is exhausted. The results include a ranking of all the evaluated models as well as partial residual plots and fit statistics of the best model. The algorithm is completely automatic and no user interaction is required while searching for the best model.

This approach is very different to that of ACE. In the GANN approach, the starting point for each transformation is a smooth sigmoidal curve which is transformed and made more or less complex (by adding or removing more *tanh* curves) until a satisfactory fit is obtained. ACE is a nonparametric procedure based on the iterative calculation of bivariate conditional expectations. The conditional expectations are estimated using a smoothing method applied to the set of observations. In the super-smoother of Breiman and Friedman, various smoothers are repeatedly applied to the observations as well as the residuals until the final smooth is obtained. The results could therefore still be highly nonlinear and jagged. This is unlikely to occur in the GANN, as the "S" shape of the *tanh* curve dictates a smooth function.

IV.Example

To illustrate the differences between the two approaches, one of the original examples studied by Breiman and Friedman [1], namely the Boston housing market data [9], is redone with the two approaches. The data set was used to determine how various factors might affect the housing values in the Boston Standard Statistical Metropolitan Area in 1970. The data set has 506 observations of the target variable, median value of owner-occupied homes (*MEDV*), and 13 explanatory variables. In the original experiment only 4 inputs were used. The inputs were: *RM*, average number of rooms in owner units; *LSTAT*, proportion of population that is lower status; *PTRAT*, pupil-teacher ratio by town school district; and *TAX*, full property tax rate. Harrison and Rubinfeld's analysis suggested the following transformations: *RM* is replaced by RM^2, *LSTAT* is replaced by *log*(*LSTAT*) and *MEDV* is replaced by *log*(*MEDV*).

In Breiman and Friedman's original article on ACE [1], an R^2 of 0.89 was reported for this problem. Several attempts have been made to reproduce this result without success. The implementation of ACE by Campher [8]obtained an R^2 of 0.81. A near identical R^2 was also obtained with the R implementation [16]. Breiman and Friedman's R^2 therefore appears to be overly optimistic and for the rest of this paper an R^2 of 0.81 will be assumed.

Figures 1 to 5 contain the variable transformations for *log*(*MEDV*), RM^2,*log*(*LSTAT*), and *PTRAT* computed with the R system. The transformations appear to be nearly identical to that reported by Breiman and Friedman [1].

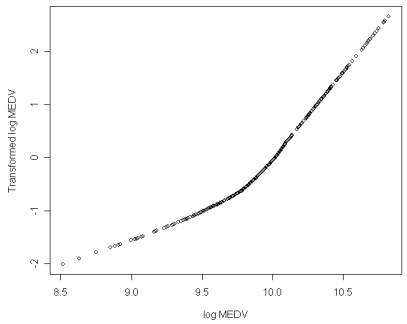

Figure 1. Log *MEDV* transformation

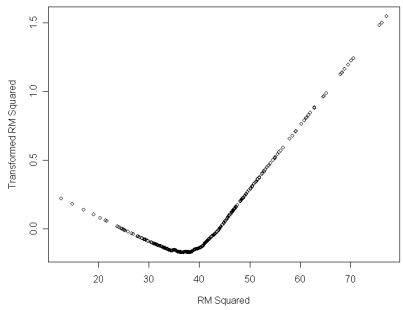

Figure 2. RM squared transformation

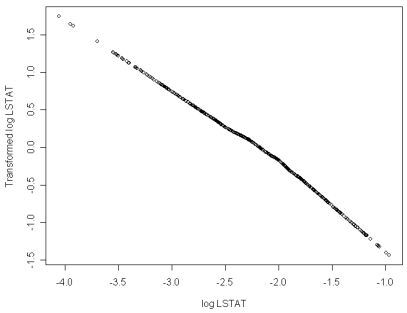

Figure 3. Log LSTAT transformation

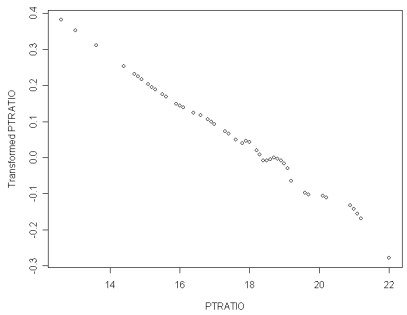

Figure 4. PTRAT transformation

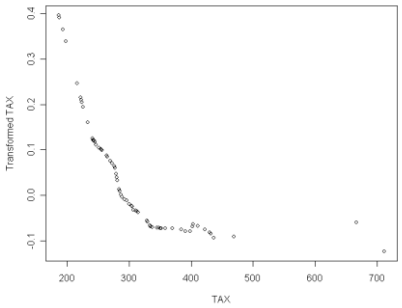

Figure 5. TAX transformation

All four variables are included in the model and an R^2 of 0.81 is obtained. The generalized additive neural network was computed with the AutoGANN modeling node [14], [15] in SAS®Enterprise Miner™. As the data set is relatively small, SBC and not cross-validation was used to obtain the best GANN. Figures 6 to 9 contain the target transformation as well as the partial residual plots (and therefore the transformations) for *MEDV*, *RM*, *LSTAT* and *PTRAT*. The reasons for using the original variables and not the transformed variables (log(*MEDV*), log(*LSTAT*) and *RM*squared) are explained in the next section.

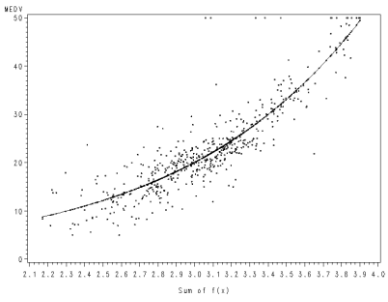

Figure 6. MEDV transformation

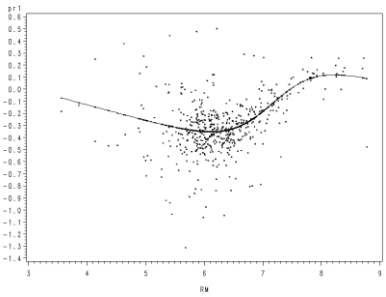

Figure 7. Partial residual plot for RM

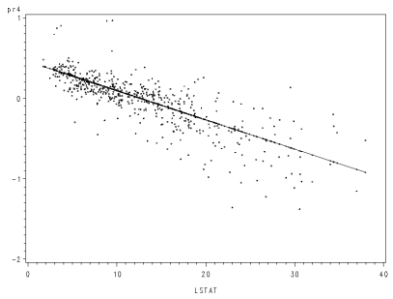

Figure 8. Partial residual plot for LSTAT

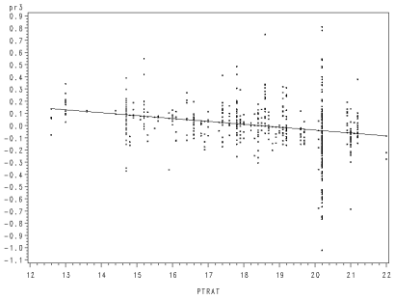

Figure 9. Partial residual plot for PTRAT

Only three of the four variables are included in the best GANN model with an R^2 of 0.795 and an SBC-value of 1485.

At first glance it seems that the GAM estimated with ACE outperformed the GANN. A detailed analysis however reveals some interesting and subtle distinctions between the two approaches. This is discussed in detail in die following section.

V.Discussion

The differences between the two approaches will be discussed under the following headings, namely *Predictive Accuracy*, *Utility*, *Stability*, *Understandability and Interpretation* and *Novelty*.

A.Predictive accuracy

The difference of 0.015 in R^2 between the GAM estimated with ACE and the AutoGANN modeling node is due to the following reasons. First, because the AutoGANN system uses objective model selection criteria to select the best model, there is a cost associated with the inclusion of a variable in the model. When the Gaussian error model applies, SBC is defined as:

$$SBC = n\ln(\hat{\sigma}^2) + K\ln(n) \tag{11}$$

where

$$\hat{\sigma}^2 = \frac{\sum \varepsilon^2}{n} \tag{12}$$

are the estimated residuals for the particular model, K the number of estimated parameters and n the training sample size. As can be seen from the formula, the penalty term, $K\ln(n)$, increases with an increase in the number of parameters.

TAX is not included in the model as it contributes very little to the overall reduction in error and therefore also the improvement in R^2 of approximately 0.01. The improvement is not enough to force the variable's inclusion into the GANN model. In other words, a GANN model with *TAX* excluded has a smaller (better) SBC-value than a GANN model with *TAX* included. This explains why *TAX* is not included in the GANN model.

Second, the transformations suggested by Harrison and Rubinfeld [9] are nearlycorrect/optimal (if they were 100% correct/optimal, Figures 1 to 5 would havecontained only straight lines). To correct these transformations with more applicablenon-linear transformations will require several extra degrees of freedom in the GANN. The slight improvements do not compensate for the extra degrees of freedom needed to model these additional transformations (again because SBC was used in the AutoGANN system as model selection criterion). It is therefore better to start from the original data set and attempt to model the non-linear transformations correctly, than to attempt to correct Harrison and Rubinfeld's suggested transformations. This explains why the GANN model was estimated using the original data set and not the data set with the transformed variables.

Third, some of the transformations suggested by the ACE algorithm are highly non-linear (e.g. *RM* squared and *TAX*) and jagged (e.g. *PTRAT*). In general, the transformations suggested by the GANN are smoother than that suggested by ACE and seems to be less affected by noise in the data. These more general trends have the effect of further reducing the obtained R^2 on the training data set. The likelihood of overtraining is therefore also reduced. An advantage of the GANN model may be that it is better suited to prediction and generalization, rather than being restricted to the interpretation and understanding of the relationships between the dependent and the independent variables.

The three reasons just given are responsible for most of the difference in R^2 between the two approaches.

A problem with ACE is the inference of response values when a new case falls outsidethe training problem space. ACE produces prediction values which are the same as themodel values defined on the closest edge of the observed problem space. This is notthe case with the GANN model, since the univariate functions remain defined outsidethe boundaries of the observed predictors. The predictions made just outside the observed problem space may be inaccurate but could still give an indicationof what the response would be if the observed relationships were to continuetheir current trends.

B.Utility

The results presented in this paper were obtained using the standard defaults of the ACE procedure in R and the AutoGANN modeling node in SAS®Enterprise Miner™. In its default form, both methods were easy to use and are

applicable to most regression problems. In her thesis, Campher [8] found that specifying the user parameter values for the ACE procedure is not trivial.

This includes the bass enhancement indicator used in the super-smoother. The bass enhancement indicator is used to reduce high variance in the super-smoother and values from 0 to 10 may be used. A value of 0 results in no bass enhancement and the normal variable span smoother is applied. Using a value of 10 is equivalent to applying a constant woofer span.

In their rejoiner, Breiman and Friedman [17] highlight that the super-smoother variable span selection produces less accurate estimates and that forcing a larger span (greater bass enhancement) produces better estimates. They also give an example in which the threshold value for the iterative procedure termination rule has to be reduced in order to produce acceptable variable transformations.

The limit the user should place on the execution time of the AutoGANN algorithm is not always apparent. Since the size of the model search space is directly related to the number of predictor variables, it is advisable to increase the running time accordingly when estimating more complex models. As the automated algorithm is loosely based on the recipe of Potts [6], where convergence is usually obtained within a small number of iterations (but with objective model selection criteria replacing human judgment), the automated algorithm should behave in a similar manner. A good GANN model is usually obtained quickly, but the identification of the optimal GANN model may take considerably longer.

Because the AutoGANN algorithm is based on search, the time it takes to find a suitable GANN architecture and model limits its usability to large problems as multiple neural networks has to be constructed and trained. A way to circumvent this problem is to ignore objective model selection criteria and cross-validation error, to force all transformations to be nonlinear and to overlook variable selection. Only one GANN architecture has now to be constructed and trained (*h* - the number of hidden nodes for each input is set to 1 or 2 depending on the complexity required for the transformations) with a dramatic decrease in the time needed to compute the transformations. An example of the application of this strategy is given in Section E.

C.Stability

The ACE method has a few factors contributing to instability. Breiman and Friedman [1] points out that the ACE model is highly dependent on the type of smoother used in the algorithm. Even the super-smoother does not always produce desirable results, especially near the boundaries of the problem input space. The ACE model is also unstable in the order in which the problem variables are entered, especially in the case of weak association between the predictor and response variables [17].

The AutoGANN model search strategy does not display this behavior of instability, providedthe time allowed for the procedure is set at an appropriate value. The order in whichthe predictor variables are listed has no influence on the algorithm. The transformationsoccurring in the final model could also be made more or less complex by selectivelyadding or deleting hidden nodes from the GANNs hidden layer. This gives the modeler the opportunity to modify the final model incorporating domain specific or other information (e.g. to force a linear transformation). Similar functionality is alsoprovided in the ACE algorithm in R.

The result of using a data-dependant model selection method based on a single selection criterion is a single approximating model that does not account for model uncertainty. The AutoGANN modeling node incorporates Bayesian modelaveraging to account for model uncertainty, although this functionality was not exploited in this paper.

D.Understandability and Interpretability

Apart from predictive accuracy and stability, understandability of a particular model is also important. More often than not predictive model users need to understand the relationships between problem variables and have to be able to interpret the prediction results. In the case of ACE, the method supplies the user with (optimal) variable transformations to interpret the variable relationships. They give insight into the relationships between the inputs and the target. These transformations, together with the additive nature of the model, make the model easier to interpret. The same applies to the partial residual plots of the AutoGANN system. Since a GANN is based on a generalized additive model, the black-box association with neural networks is overcome.

E.Novelty

Linear statistical models for regression have been around for a long time and the theory and application thereof has been studied intensively. With the growth of interest in knowledge discovery and data mining, more flexible non-linear modeling methods such as ACE and Artificial Neural Networks (ANNs) have gained interest. Problems associated with extremely flexible models such as these include spurious relationships being identified and the tendency to over fit the data [18]. Strategies that may be used to ease these problems include the use of objective model selection criteria (with penalized goodness-of-fit functions), restriction of model complexity and the shrinking of an over fitted model. ACE and AutoGANN make use of some or all of these strategies. ACE has a bass control that can be set to obtain better generalization, it restricts the model to be additive and the stepwise variable selection procedure shrinks the model. The AutoGANN system uses objective model selectioncriteria such as AIC and SBC that penalize complexity. The model form isalso restricted to a generalized additive model.

In general, ACE has been widely praised for its contributions to data analysis, mainlyin terms of finding optimal variable transformations.[19] ascribe the scarce practical application of ACE to its susceptibility to noise andthe fact that it requires a large observations-to-variable ratio in order to obtain reliable results. To resolve this issue they have developed a modified ACE method, called GA-

ACE, incorporating genetic algorithms and data compression.

GANN models estimated by the AutoGANN system compare well with ACE in terms of prediction. It however seems to be more conservative in its variable transformations as it incorporates objective model selection criteria and cross-validation [7]. To illustrate this point consider the GANN model estimated on the Boston housing market data without regard for the best SBC value.

A GANN model with an R^2 similar to that of the model based on the transformations by ACE can be estimated by forcing the GANN to model all transformations as non-linear transformations (*h* was set to 1 and a skip layer was included for all inputs as well as for the output activation function - the inverse of the link function) and ignoring the objective model selection criteria. Although this is not ideal, it shows the influence of incorporating objective model selection criteria as implemented in the AutoGANN system has on the final GANN model. Figures 10 to 14 contain the results.

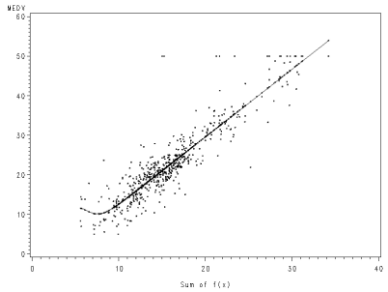

Figure 10. MEDV transformation

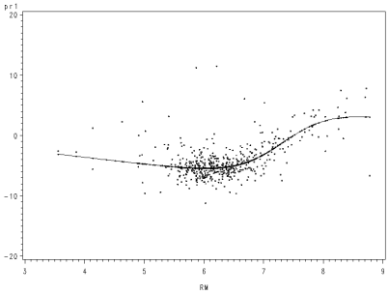

Figure 11. Partial residual plot for RM

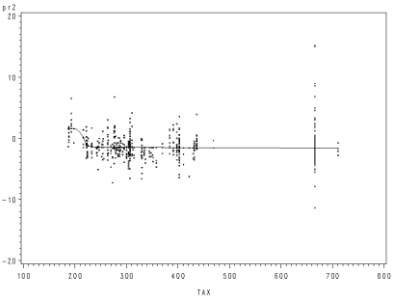

Figure 12. Partial residual plot for TAX

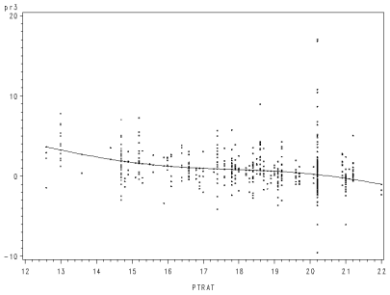

Figure 13. Partial residual plot for PTRAT

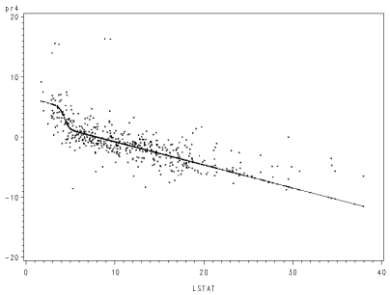

Figure 14. Partial residual plot for LSTAT

This model has a R^2 of 0.812 (and an SBC-value of 1528). The transformations are still smooth, although they have become highly non-linear. The results are similar to that obtained with the ACE algorithm, but the authors prefer the GANN model given in Section III, as it is more parsimonious. This result may also indicate that users of the ACE algorithm should be aware of the possibility of overtraining.

VI. Conclusions

The estimation of accurate generalized additive models will become more and more important as the move away from linear models gains momentum (e.g. in the Credit Scoring industry). It is therefore of the utmost importance that algorithms aredeveloped that could assist the modeler in the estimation of stable, accurate andeasy to interpret models. ACE was one of the first attempts to estimate optimaltransformations for multiple regression problems and although it received praiseas a novel and remarkable achievement, it practical application is limited. Thisleaves room for alternative strategies and algorithms to be developed. One recentlydeveloped alternative is based on the use of universal approximators and exploitedin the AutoGANN system. It incorporates objective model selection criteria andcross-validation into the process and provides an easy to use graphical interfaceto the modeler which should simplify the modeling process.

It is very difficult to motivate a preference for any of the two stated approaches. Each approach has its advantages as well as its disadvantages. Both approaches are conceptually simple to understand, but not trivial to implement. The main difficulty with ACE is in the implementation of the super-smoother (see the technical report of Friedman [10] and the

thesis of Campher [8]).It also generates highly non-linear and sometimes jagged smooths and may be prone to over fitting. But, it is the faster algorithm.

The success of the GANN approach depends on the specification of the correct GANN architecture, which is not a trivial task that could be time consuming. But, this architecture selection problem has been completely automated in the AutoGANN system and therefore does not present a hurdle to the modeler any more. An advantage of the GANN approach is that it directly generates functions that may be used to score new data. It furthermore adds an extra level of objectivity to the modeling process by exploiting objective model selection criteria and cross-validation to assist in model selection. This is absent from the ACE approach. The GANN approach also provides an elegant and straightforward introduction to ANNs that could facilitate the adoption of ANNs by modelers not familiar with machine learning techniques.

It is noteworthy that GANNs perform at least as well as ACE as demonstrated in this paper and in the thesis of Campher [8]. As ACE was praised as a remarkable achievement, GANNs should be seen in the same light. The practical application of GANNs will however depend on its acceptance (or not) as a viable modeling technique in the predictive modeling community.

Acknowledgements

The authors wish to thank SAS® Institute for providing them with Base SAS® and SAS® Enterprise Miner™ software used in computing all the results presented in this paper. This work forms part of the research done at the North-West University within the TELKOM CoE research program, funded by TELKOM, GRINTEK TELECOM and THRIP.

References

[1]L. Breiman, and J. H. Friedman,"Estimating optimal transformations for multiple regression and correlation",*Journal of the American Statistical Association,* Vol.80, No. 391, 1985, pp 580–598.

[2] E. B. Fowlkes, and J. R. Kettenring, Comment on "Estimating OptimalTransformations for Multiple Regression and Correlation", by L. Breiman and J. H. Freidman,*Journal of the American Statistical Association,* Vol. 80, No. 391, 1985, pp 607–613.

[3] D. Pregibon, and Y. Vardi, Comment on "Estimating Optimal Transformationsfor Multiple Regression and Correlation", by L. Breiman and J. H. Friedman,*Journal of the American Statistical Association,*Vol. 80, No. 391, 1985, pp 598–601.

[4] W. S. Sarle, "Neural networks and statistical models",*Proceedings of the Nineteenth Annual SAS®Users Group International Conference*, 1994.

[5] Ripley, B. D., *Pattern Recognition and Neural Networks*, CambridgeUniversity Press, Cambridge, United Kingdom, 1996.

[6] W. J. E. Potts, "Generalized additive neural networks",*In 'KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, 1999, pp 194–200.

[7] Du Toit, J. V., *Automated Construction of Generalized Additive Neural Networks for Predictive Data Mining*. PhD thesis, School forComputer, Statistical and Mathematical Sciences, North-West University, South Africa, 2006.

[8] Campher, S. E. S., *Comparing Generalised Additive Neural Networks with Decision Trees and Alternating Conditional Expectations*, Master's thesis,School for Computer, Statistical and Mathematical Sciences, North-West University, South Africa, 2008.

[9] D. Harrison, and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air",*Journal of Environmental Economics and Management*,Vol.5, No. 1, 1978, pp 81–102.

[10] Friedman, J. H., *A variable span smoother*, LCS technical report no. 5SLAC PUB-3477, Stanford, CA: Department of Statistics, Stanford University, 1984.

[11] J. A. Nelder, and R. W. M. Wedderburn, "Generalized linear models", *Journal of the Royal Statistical Society*,Vol.135, No. 3, 1972, pp 370–384.

[12] Hastie, T. J. and R. J. Tibshirani, *Generalized Additive Models*,Monographs on Statistics and Applied Probability, Vol. 43, Chapman and Hall, London, 1990.

[13] G. Zhang, B. E. Patuwo, and M. Y. Hu, Forecasting with artificialneural networks: The state of the art,*International Journal of Forecasting*, Vol.14, No. 1, 1998, pp 35–62.

[14] D. A. De Waal, Generalized additive neural network modeling,*Invited Talk,M2007, SAS' 10th Annual Data Mining Conference*, Las Vegas, 2007.

[15] D. A. De Waal, and J. V. Du Toit,"Generalized additive models from a neural network perspective",Workshops Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA, IEEE Computer Society, 2007.

[16] Wood, S. N., *Generalized Additive Models: An introduction with R*, Textsin Statistical Science, Chapman & Hall/CRC, London, 2006.

[17] L. Breiman, and J. H. Friedman, Rejoiner for "Estimating OptimalTransformations for Multiple Regression and Correlation",*Journal of theAmerican Statistical Association,*Vol. 80, No. 391, 1985, pp 614–619.

[18] D. Hand, "Data mining: statistics and more?", *The American Statistician*,Vol. 52, No. 2, 1998, pp 112–118.

[19] I. Esteban-Díez, M. Forina, J. González-Sáiz, and C. Pizarro, "GA-ACE: alternating conditional expectations regression with selection of significant predictors by genetic algorithms", *Analytica chimica acta*, No. 444, 2006, pp 96-106.

Author Biographies

André de Waal was born in South Africa on the 13th of December 1963 and received his Ph.D. in theoretical computer science from the University of Bristol in the United Kingdom during 1994. He spent the next year in Germany and Belgium continuing his research in logic programming and automated theorem proving. During 1996 he returned to South Africa to take up his position as lecturer at the School of Computer Science and Information Systems at the then Potchefstroom University for Christian Higher Education (which later became the North-West University), where he was later promoted to associated professor. During 1999 he became one of the founder members of the Centre for Business Mathematics and Informatics at the same university. He became responsible for the data mining program in the centre and shifted his research focus to include neural networks and predictive modeling. He is the co-developer of the AutoGANN and AutoMLP modeling nodes in SAS Enterprise Miner and has published several research papers on the automation and use of generalized additive neural networks. He joined SAS Institute in Cary,

NC during December 2010 to take up the position of analytical consultant in the global academic program.

Susanna E. S. Campher, née Coetzee, was born in Potchefstroom, South Africa in 1977. She obtained the degree BSc in computer science and mathematics from the Potchefstroom University for Christian Higher Education, South Africa in 2000 and the degree BSc(Hons) in computer science from the same university in 2001. In 2008 she obtained an MSc degree in computer, statistical and mathematical sciences from North-West University in South Africa. She is currently working as a business intelligence specialist at the North-West University and her career interests include data warehousing and data mining.

Tiny du Toit was born in South Africa on 26 August 1975. He received his B.Sc. (computer science and mathematics) in 1997, B.Sc.Hons. (computer science) in 1998 and M.Sc. (computer science) in 2000 from the Potchefstroom University for Christian Higher Education in South Africa. In 2006 he obtained his Ph.D. (computer science) from the North-West University in South Africa. Currently he is a senior lecturer at the North-West University and is doing research in the field of artificial intelligence.