This version: September 2, 2003

A Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA)[1]

TEODOSIO PEREZ-AMARAL[†], GIAMPIERO M. GALLO[‡] and HALBERT WHITE[*]

[†] *Departamento de Analisis Economico, Universidad Complutense de Madrid*
28223 Madrid, Spain
(e-mail teodosio@ccee.ucm.es)

[‡] *Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale G.B. Morgagni, 59*
50134 Firenze, Italy
(e-mail gallog@ds.unifi.it)

[*] *Department of Economics, University of California, San Diego, 9500 Gilman Drive*
La Jolla, CA, 92093-0508, USA
(e-mail hwhite@weber.ucsd.edu)

ABSTRACT

A new method, called Relevant Transformation of the Inputs Network Approach (RETINA) is proposed as a tool for model building. It is designed around flexibility (with nonlinear transformations of the predictors of interest), selective search within the range of possible models, out-of-sample forecasting ability and computational simplicity. In tests on simulated data, it shows both a high rate of successful retrieval of the DGP which increases with the sample size and a good performance relative to other alternative procedures. A telephone service demand model is built to show how the procedure applies on real data.

I. Introduction

In the process of model building, a decision must be made as to which among several specifications (possibly belonging to different classes of models) should be chosen to represent a relationship between a dependent variable and other variables of interest. Among these, one may prefer a parametric specification (either linear or nonlinear) where some interpretation of parameter values may be retained, or else suggest the adoption of flexible functional forms where the relationship among the variables is guided by other criteria of explanatory power.

Within each class of models, specification selection is far from trivial[2]: some methods focus on the relationship between a model and its interpretability according to some theory, others are based on hypothesis testing between competing models; some depend upon the trade-off between explanatory power and parsimony in the retained specification, others are based on the performance of a model in explaining a set of data not used for estimation, especially when the flexibility of the specification tested in-sample may signal overparameterization when applied out-of-sample; and so on. The lively debate spurred by the paper by Hoover and Pérez (1999, and the discussions contained in the same issue of the *Econometrics Journal*) and the contributions provided at the 2002 EC^2 conference in Bologna are a proof that the question, far from being resolved, is receiving a great deal of interest, especially in the attempt to produce an automated procedure capable of paralleling the ever-growing computing power available to researchers.

No approach is perfect, especially when misspecification of a model relative to the process which generated the data is always a possibility; hence all approaches to model selection have their particular limitations. Hypothesis testing in support of model choice is well-known to be potentially dangerous (cf. Granger *et al.*, 1995) given the implicit advantage attributed to the model under the null hypothesis in a nested framework or the possible ambiguity of results in a non-nested context. Moreover, one of the undesirable aspects of such an approach is the need to resort to pairwise comparisons.

---

[2] An excellent compendium of issues on statistical model selection is the book by Burnham and Anderson (2002).

In frameworks in which a penalty function for the number of parameters modifies the value of the likelihood function to provide a number which can be used to select a model (as in Akaike's, 1973, AIC or Schwartz's, 1978, BIC) there is always the issue of which form such a function should take, especially given certain undesirable properties of such information criteria in systematically choosing over- or under-parameterized models in some circumstances.

Model selection based on out-of-sample performance is also prone to problems and, in fact, after the pioneering work by Granger and Newbold in the early 1970s (Granger and Newbold, 1973), only in recent years has it become standard practice to adopt testing procedures for predictive ability whereby some measure of performance (such as the Mean Squared Prediction Error but, again, the choice of the criterion is not neutral) is used in a formal hypothesis testing framework (cf. Diebold and Mariano, 1995; West, 1996; White, 2000; Giacomini and White, 2003).

In this paper we present an approach based on earlier work by White (1998), called the Relevant Transformation of the Inputs Network Approach (RETINA). It aims at achieving a flexible and parsimonious representation of the mean of a variable, conditional on a (potentially large) set of variables deemed of interest in situations where one does not have strong priors as to the form of the suitable function linking available information, or the relevance of individual variables. It has the flexibility of neural network models in that it accommodates nonlinearities and interaction effects (through nonlinear transformations of the potentially useful variables in the conditioning set), the concavity of the likelihood in the weights of the usual linear models (which avoids numerical complexity in estimation), and the ability to identify a set of attributes that are likely to be truly valuable for predicting outcomes (which corresponds to a principle of parsimony). In performing model selection, our approach relies on an estimation/cross-validation scheme which is aimed at limiting the possibilities that good performance is due to sheer luck. Some simulation results show that it has good finite sample properties.

We will start (Section II) by discussing flexibility, selective search and out-of-sample forecasting ability which are the important elements of model building and selection embedded into RETINA. The details of the RETINA algorithm are discussed in section III. We illustrate the characteristics and properties of the procedure in Section IV by

means of some simulation experiments with different DGPs where we take into account linearity and nonlinearity, as well as some pecularities in the data (outliers, sparse data, structural break). We provide some insights about the rationale of some features of the algorithm by examining simpler versions of RETINA; further sets of results show how RETINA performs relative to other automated procedures. In Section V we sketch some evidence of the practical issues encountered when applying RETINA to real data. Concluding remarks follow.

II. Some Ingredients for Model Building and Selection.

A researcher typically selects a "preferred" parametric model of some aspect of the observed phenomenon, that is, a useful approximation to the data generating process (DGP, which may not even be within the class of candidate models), in that it provides a useful representation of the data characteristics. In general, for given available data (e.g., cross section, time series or panel) and aggregation level (e.g., individual, family, etc.) the choice of a model is influenced by its intended purpose, e.g., estimation of a conditional mean, hypothesis testing or out-of-sample forecasting. For example, at times economists use different models of consumption depending on whether they wish to estimate a given parameter, to choose among competing theories, or to obtain out-of-sample forecasts.

We will frame the usual problem of the description of the behavior of a dependent variable Y as one where, from a (potentially large) set of variables X of possible relevance for Y, we need to model the behavior of the conditional mean $E(Y/X)$: the question is one of selecting which X's are relevant and finding out how they exert their impact on Y (possibly non-linearly). RETINA´s recommended model aims at achieving a useful approximation to the unknown relationship (whatever coherence and rationale the researcher may find for the result), following three main principles: flexibility, selective search, and out-of-sample forecasting performance. The procedure should also accommodate forms of prior knowledge such as the addition or deletion of variables, the introduction of restrictions, or other theoretical or empirical considerations.

Flexibility

Given a lack of information on the form of the relationship linking X to Y (as is common in economics), in order to maintain a certain degree of flexibility one may use a set of transformations of the input variables, say $\zeta(X)$, which embody both *nonlinearities* and *interactions*. Thus, if for each observation i ($i=1, ..., n$) we observe a value of the response variable Y_i and we have available candidate predictor attributes X_{ih}, $h = 1,..., k$, (where k is potentially a very large number), the set $\zeta(X)$ can be made up of transformed variables W_{ij}, $j = 1,...,m$, each of which is obtained as $X_{ih}^{\alpha} X_{il}^{\beta}$, $\alpha, \beta = -1, 0, 1; h, l = 1, \ldots, k$.[3] $\zeta(X)$ will include the original X_{ih}'s, their squares, cross products, inverses and cross-ratios ("level 1 transforms", avoiding possible repetitions of outcomes and divisions by zero).

Since the ultimate goal is to identify a parsimonious set of (transformed) attributes that are likely to be truly relevant for predicting out-of-sample outcomes for Y, we need to be careful that the transformations we choose are not highly correlated with one another, as highly correlated transforms will not provide a great deal of independent predictive information.

Finally, a desirable property of the procedure should be concavity of the likelihood in the parameters (to avoid numerical complexity) and this can be achieved by allowing the effects of the $\zeta(X)$'s on Y to be exerted in a linear fashion.

The model we will be considering has the form:

$$E(Y/X) \approx \zeta(X)'\beta$$

in the regression case or, more generally

$$E(Y/X) \approx F(\zeta(X)'\beta)$$

where F is a suitable link function (e.g., the logistic cumulative distribution function for binary classification problems). We will rule out the appearance of further unknown parameters inside ζ because that may result in non-concavity.

---

[3] The number of transformed variables is therefore $m = k^2 + 5k$.

Selective Search

The task of evaluating all the 2^m possible models when we have m candidate regressors in the set of transformed variables $\zeta(X)$, and then of applying some form of model selection would quickly become impossible for an even moderate number of X variables. Rather, following the ideas in White (1998), we can think of selecting a number (of order proportional to m) of candidate models, inserting new explanatory variables on the basis of their relevance for the problem at hand: one possibility is to rank the candidate regressors according to their correlation in absolute value with the dependent variable. At the same time, in order to control for the degree of dependency of the new information added, we may want to keep the amount of collinearity among the regressors below a threshold parameter λ chosen by the experimenter (as λ approaches 0 new regressors approach orthogonality; as λ approaches 1 new regressors may be highly collinear).

Out-of-Sample Forecasting Ability

Although flexibility is desirable, in order to avoid the overparameterization suggested by a good in-sample fit for the model, we use disjoint sub-samples for estimation and cross-validation, and an out-of-sample prediction performance criterion for model selection as important features of the procedure. The order in which sub-samples are used for estimation and/or cross-validation should not matter, and an extensive search of the model over different orderings of the sub-samples could be performed.

III. The RETINA Procedure.

The ingredients outlined in the previous section find their expression in the algorithm described in Table 1. This also contains further definitions of objects that we will refer to in what follows.

TABLE 1 about here

Some comments are needed to justify and/or clarify some features.

- *Division of the entire sample into three sub-samples.* The choice of three (as opposed to a more customary two when cross-validation is adopted) as the number of sub-samples is mainly heuristic at this stage but is crucially motivated by the need to examine outcomes for different orders in which the samples are fed to the procedure, since we use the first one for estimation alone, the second one for cross-validation and re-estimation, and the third one just for cross-validation. We keep in mind Miller's (1990) result that in sub-sample 1 parameters and standard deviations are biased away from zero, and therefore we estimate parameters again in the second sub-sample. It is therefore advisable to perform the cross-validation once more, with unused information included in the third sub-sample. In view of this, splitting the sample in more than three sub-samples would seem to have diminishing returns. The sub-samples must be disjoint so that the information and the statistics we compute are (at least roughly) independent across samples, and should be as similar to one another as possible to limit the dangers of unaccounted for clusters of heterogeneity. With practical applications, the need for some sort of randomization of the way the observations are ordered may arise (cf. Section V).

- *Safeguards against spurious correlations.* One legitimate concern is related to the possibility that a variable which does not enter the DGP may be erroneously selected in the set of the candidate regressors. This possibility is made more remote by three features of RETINA:

  i) one is the fact that the order suggested by estimation on the first sub-sample and cross-validation on the second sub-sample is scrambled by re-computing the correlations between the candidate set of regressors and Y on the second sub-sample before proceeding to the cross-validation on data from sub-sample 3 (i.e., $W_{(1)}$ will still be present but need not be the first variable to be included in the re-estimation on the second sub-sample).

  ii) The second is the use of λ as a threshold to control for the correlation among regressors: the decision that new variables should be added to the set on the basis of how much of their variability is accounted for by variables already

included entails that what follows $W_{(1)}$ in $\zeta_\lambda(X)$ need not be the same for different values of λ.

iii) The third is that the repetition of the procedure on different orderings of the sub-samples (this would involve a total of six repetitions) may alter the composition of the candidate set of regressors altogether (i.e., $W_{(1)}$ may not be the same variable for different orderings).

- *Information criteria.* The criterion we adopt to select a candidate model has two-steps: starting from the estimation on the first sub-sample we select the model which minimizes the mean square prediction error on the second sub-sample. But in order to select the final model associated with a specific ordering of the sub-samples we then apply an information criterion, which adds to the mean square error a penalty term that depends on the number of parameters of the model: here we use the out-of-sample AIC. The justification for this procedure is pragmatic, rather than theoretical: the results of Sin and White, 1996, do not carry over straightforwardly to this two-step - and then repeated for a different ordering – approach. Nevertheless, without such a penalty we have found that the procedure tends to select somewhat larger models than would be ideal. Imposing the relatively modest AIC penalty usually takes the results in the right direction as seen in the simulation of Section IV. Precisely why this happens should be the object of a theoretical question for further research.

- *Comparison with some other methods in the literature*:

  i) RETINA shares flexibility with Artificial Neural Networks (White, 1989), while maintaining linearity in the parameters within the link function which makes estimation easier; it uses out-of- sample predictive criteria, while ANN often use an in-sample goodness of fit criterion.

  ii) The "general-to-specific" (Gets) approach to model building and selection enjoys a long tradition in econometrics (an overview of the literature is forthcoming in Campos, Ericsson and Hendry, 2003). By framing it into the possibility of its being translated into an automated procedure, Hoover and Pérez (1999) can be credited with reviving the debate on the merits and

properties of the procedure. This methodology starts with a sufficiently complicated model to describe economic phenomena, and by way of statistical hypothesis testing is able to reduce model complexity thereby conveying the same information about the phenomenon of interest in a more parsimonious way. The preliminary results in the discussion by Hendry and Krolzig (1999) and their later work (still in progress: Krolzig and Hendry, 2001; Hendry and Krolzig, 2003a and 2003b) have uncovered much room for improving the procedure. Hypothesis testing as a model selection strategy is also used in stepwise regression (Miller, 1990) to which we will provide some comparisons below. RETINA expands the range of possible regressors through the transforms (a feature which could in principle be adopted by Gets as well) but relies on out-of-sample performance measures rather than residual diagnostics and hypothesis testing on coefficients. The use of sub-samples is also another point in common with Gets though for different reasons: in Gets the splitting is intended to check against the dangers of spurious correlations possibly detected in the overall sample; in RETINA it is a building block of the selective search.

iii) An alternative model building and selection approach is the non-negative garrote (Breiman, 1995) aimed at selecting among the possible (untransformed) explanatory variables in a regression framework by zeroing and/or shrinking coefficient estimates and using cross-validation. Breiman (1995) himself acknowledges that the models selected by the non-negative garrote tend to be overparameterized, but the shrinkage provides accuracy.

iv) Generalized linear models and generalized additive models (Hastie and Tibshirani, 1990) are linear in the parameters. Like our procedure, they can incorporate nonlinear link functions and may accommodate distributional assumptions about the error terms. They provide a general model building environment but are less amenable to automated model selection.

The characteristics of RETINA thus outlined shield the procedure from the pejorative aspects of data mining (as discussed, among others, by Campos and Ericsson, 1999) in that the selective search makes no attempt at maximizing t-ratios and deals with data

interdependence and over-parameterization. Moreover, starting from a large set of variables and enlarging this set through transforms reduce the danger of variable omission. As the procedure is automatic, it does not pay attention to whether estimates are sensible, say, from an economic point of view. This should not prevent the researcher from exercising expert knowledge, especially when several candidate models are available.

IV. Simulations.

Because attempts to obtain analytic results are overwhelmed by the intricacy of the approach, the major proving ground is how well the procedure works. Here we test the ability of RETINA to select a model corresponding to the DGP on simulated data. The goal of the simulations is twofold: to check the performance of RETINA and to compare its capabilities relative to other procedures. Aware as we are of Granger's and Timmermann's (1999) warning that the `true' model is rarely even approximately known in practical situations, we would like it to be clear that in our simulations we have followed the common practice of adopting a metric based on a known DGP, in the spirit that at least the procedure should 'behave well' in such a case (cf. the reply by Hoover and Pérez, 1999).

As far as the first goal is concerned, therefore, we design several data generating processes and vary several parameters, such as the overall sample size n, the amount of correlation ρ among original variables X, and the variance σ of the disturbance term so as to achieve a desired average R^2 for the resulting estimated equations across replications. Apart from some simple linear and nonlinear DGPs, we also probe the sensitivity of the selective search performed by the algorithm by including some peculiarities in the data generation, such as discrete explanatory variables or explanatory variables with sparse data or outright noise in the form of outliers or structural breaks in the DGP. The DGPs are detailed in Table 2.

TABLE 2 about here

To keep matters simple, RETINA was used with the level one transforms of the variables included in the DGP and of an additional irrelevant variable with the same distribution and correlation as x_{i1} and x_{i2}. As a result, the maximum number of candidate regressors (the W_j's) is 25 and the total number of possible candidate models to consider would be $2^{24} = 16777216$ (the constant is always included) versus the order 2x24=48 evaluated by RETINA. We let the parameter λ vary from 0 to 1 by increments of 0.1. Recalling that the model selected by RETINA is the one which has the best out-of-sample forecasting performance relative to the DGP, for reporting purposes we will count a 'success' when RETINA chooses a model which coincides with the DGP and trace the percentage of successes over 1000 replications of the experiments. Table 3 reports such percentages organized by overall sample size and average R^2. A visual rendition of Table 3 is given in Figure 1 where each panel reports the results for each DGP.

TABLE 3 about here

FIGURE 1 about here

The results are strikingly similar in that the pattern which can be discerned is an increase of the percentage of success as the sample size, or the R^2, increase across a wide variety of situations. A refinement of the grid for λ (the parameter that controls for collinearity among regressors) affects the success rate only marginally.[4]

The simulations can also help us understand the benefits of some of the choices made in designing the algorithm. In particular, a useful comparison is relative to simpler versions of RETINA which apply either a split of the overall sample into two rather than three sub-samples (White, 1998); or one in which there is no repetition of the procedure for a different ordering of the sub-samples (e.g., 3-2-1 instead of the initial 1-2-3, cf. Stage III.5 in Table 1), as well as one in which no penalty term for the number of parameters in the model is applied to the mean squared error computed on the third sub-

---

[4] In the n=100, R2=0.50 experiment with DGP1 the results were 72.9% with a 0.1 step versus 73.1% with a 0.01 step at the price of a five-fold increase in the computation time.

sample (cf. 4.b in Table 1). The results (the full details of which are not reported here, but are available upon request) show that

- Concerning the split in two versus split in three sub-samples, the results show that there is a tendency for the former algorithm to overparameterize the selected model, and to stabilize the rates of success fairly far away from 100% when either the overall sample size or the R^2 increase. This is true for all the DGPs tested in which this version of model selection is outperformed by RETINA; the only exception occurs in the linear DGP with $n=100$ and $R^2=0.25$, possibly due to problems with estimation on a third of a small sample rather than on one half.

- When the algorithm is based on the division into three sub-samples, the repetition of the procedure on a different ordering of the sub-samples provides some substantial gains in the successful retrieval of the DGP: for $n=100$ the gain is about 40% for low values of R^2, with a decrease in the gains as n and/or R^2 increase. As an example, one can see the results obtained for the DGP 1 reported in Figure 2 – panel A.

- Finally, one can further simplify the algorithm by not repeating the procedure on a different ordering of the sub-samples and by adopting a different forecast performance criterion, namely using a simple minimum mean square error on the third sub-sample without adding any penalty term for the number of parameters in the candidate models. In such a case the procedure has a better performance in small samples but it tends to stabilize around rates of success of about 75% instead of about 98% for larger samples. Again for a comparison, it is worth looking at Figure 2 – panel B.

FIGURE 2 about here

The second set of experiments is run by comparing RETINA's performance with the results achievable by backward stepwise regression (Miller, 1990) and the non-negative garrote (Breiman, 1995) when both procedures are provided with the same level one transforms of the original variables. Table 4 shows that for the data of the linear DGP RETINA outperforms its rivals. The comparison is computationally demanding as the

execution time for the non-negative garrote is more than two hundred times that of RETINA due to the complicated optimizations and the tenfold cross-validation used by this method[5], and is certainly not exhaustive across DGPs.

V. An Illustration with Empirical Data.

A further major proving ground of RETINA's usefulness lies in its application to empirical data and modeling of phenomena for which the functional form linking a dependent variable of interest with a set of available variables is not known a priori. While we will not develop an original application in this section, we will borrow some interesting preliminary results from Perez-Amaral and Marinucci (2002) demonstrating some of the technical difficulties that need to be taken into account when applying RETINA.

The empirical question originates with the desire to model the demand for business telephone toll use by individual firms: as Taylor (1994) discusses, this is a field in which unaccounted heterogeneity can pose serious difficulties and in which the selection of really relevant variables becomes an empirical question. The dataset we have available[6] is a cross-section (dating from 1997) of 13,766 individual businesses across 9 US states, comprising 32 variables related to consumption of telephone services. The phenomenon of interest is the **total local bill**, that is, the dollar amount for short distance calls paid by individual businesses. Among the possible explanatory variables, some are related to the number and type of telephone connections: the number of **Business** lines; the number of **Hunting** lines (a service that bundles all the telephone lines in the same location to be easily accessible with only a single number); the number of **PBX** connections (the connections between a firm's Private Branch exchange and the outside telephone networks) and the number of **CENTREX** lines (a sort of outsourcing of telephone services which avoids the need to purchase equipment at high risk of obsolescence). Others are related to the size and organization of the firm: the **Total** number of employees in the firm; the number of employees working locally (**Here**); **Sqft** (the

---

[5] We have also limited the maximum value of the garrote parameter to 6 or 4 for faster convergence and speed of execution.

surface occupied by the business activities), **Pop**ulation (the size of the area served by the business), and **Sales** (the sales by the firm). For brevity's sake, let it suffice to say that many unreported data in the form of zeros are present, especially in the square footage and sales. After cleaning the data (assuming that errors in reporting are not systematically linked with the dependent variable) the effective sample was reduced to 4,476 observations: to reduce heterogeneity a transformation of the total bill into total bill per local employee was adopted: over the whole sample a linear model had a very low R^2 (equal to 0.041) while the model provided by RETINA gave an R^2 equal to 0.635. Moreover, a random shuffling of the original firms to avoid possible clusters of heterogeneity across sub-samples brought about an improvement to an R^2 of 0.831. The out-of-sample performance computed as the root mean square error on a sub-sample not used for estimation shows that RETINA outperforms the linear model by 2.3 times. The selected model includes a constant and the following transformations (all of which are individually statistically significant): Business/Total; Hunting/Total; Business/Here; Hunting/Here; PBX/Here; CENTREX/Here; Sqfoot/Here; 1/(Here*Sales); 1/(Here*Pop). Far from claiming that the demand for telephone business toll use is correctly specified, what we retain from the exercise is that the results show the usefulness of transformations identified by RETINA beyond what may be suggested by common practice in econometrics (in the case at hand, for example, expressing the original variables as ratios to local employees).

VI. Conclusions.

In this paper we have developed a new method which may prove to be a useful tool for model building and selection when applied to empirical datasets. It can be used also as a data exploratory tool to suggest possible modeling choices and transformations of the explanatory variables to researchers who may want to exercise their own expert judgement in the final choice.

---

[6] The data were provided by PNR, a subsidiary of Indetec International, now TNS Telecoms.

While they do not cover all possibilities, our simulations have delivered a fairly reassuring picture of the performance of RETINA in recovering simple DGPs, especially with regard to the increase toward 1 of the success ratio as n increases.

More simulations are certainly needed, to cover additional relevant cases, and especially to reproduce more and more realistic datasets. The unceasing increase of the computing power makes this task look less daunting than in the past.

Several features of RETINA are still in need of investigation: how other choices of forecast performance measure inserted at the various suitable stages affect the results presented here; the treatment of dependent processes (including the important case of nonstationary variables) in a time-series context and of heterogeneity in a cross-section context; or the use of likelihood-based estimation procedures *in lieu* of regression, when the model to be selected belongs to the class of limited dependent variables models, for example, or when the error distribution assumptions are different from normality.

The most promising comparison left aside here is between PcGets and RETINA. This is in part due to the fact that both procedures are still being developed and in part due to the fact that the software needed to run a fair competition between the two is highly demanding: a full-fledged comparison should be the object of a separate study, to investigate how the different logics of multi-path search and of flexibility express themselves in each methodology, compared side by side.

References.

Akaike, H. (1973). 'Information theory and an extension of the likelihood principle', in B.N. Petrov and F. Csaki (eds.), *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Breiman, L. (1995). 'Better subset regression using the nonnegative garrote', *Technometrics*, Vol. 37, 373-84.

Burnham, K. and Anderson, D. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach*, 2nd ed., Springer-Verlag, New York.

Campos J. and Ericsson, N.R. (1999). 'Constructive data mining: modeling consumers' expenditure in Venezuela', *Econometrics Journal,* Vol. 2, 226-40.

Campos J., Ericsson, N.R., and D.F. Hendry (2003). *Readings in General-to-Specific Modeling*, Edward Elgar, Cheltenham, forthcoming.

Diebold, F.X. and Mariano, R.S. (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics*, Vol. 13, 253-63.

Giacomini, R. and White, H. (2003), 'Tests of conditional predictive ability', UCSD Dept. of Economics, Working Paper 2003-09.

Granger, C.W.J., King, M. and White, H. (1995). 'Comments on testing economic theories and the use of model selection criteria', *Journal of Econometrics*, Vol. 67, 173-87.

Granger, C.W.J. and Newbold P. (1973). 'Evaluation of forecasts', *Applied Economics*, Vol. 5, 35-47.

Granger, C.W.J. and Timmermann, A. (1999). 'Data mining with local model specification uncertainty: a discussion of Hoover and Perez', *Econometrics Journal,* Vol. 2, 220-25.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Monographs on Statistics and Applied Probability 43, Chapman and Hall, London.

Hendry, D.F. and Krolzig, H.-M. (1999). 'Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez', *Econometrics Journal*, Vol. 2, 202-19.

Hendry, D.F. and Krolzig, H.-M. (2003a). 'New developments in automatic general-to-specific modeling', forthcoming in B.P. Stigum (ed.), *Econometrics and the Philosophy of Economics*, Princeton University Press, Princeton.

Hendry, D.F. and Krolzig, H.-M. (2003b). 'Sub-sample model selection procedures in Gets modelling', forthcoming in R. Becker and S. Hurn (eds.), *Advances in Economics and Econometrics: Theory and Applications*, Edward Elgar, Cheltenham.

Hoover, K.D. and Perez, S.J. (1999). 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search', *Econometrics Journal,* Vol. 2, 167-91.

Inoue, A. and Kilian, L. (2003). 'On the selection of forecasting models', WP 214, European Central Bank.

Krolzig, H.-M. and D.F. Hendry (2001). 'Computer automation of general-to-specific model selection procedures', *Journal of Economic Dynamics and Control*, 25, 831-66.

Miller, A. J. (1990). *Subset Selection in Regression*, Monographs on Statistics and Applied Probability 40, Chapman and Hall, London.

Pérez-Amaral, T. and Marinucci, M. (2002), 'Econometric modeling of business telephone toll demand for individual firms using a new model selection approach, RETINA', presented at the 13th Regional Conference of the International Telecommunications Society, Madrid.

Schwartz, G. (1978). 'Estimating the dimension of a model', *Annals of Statistics*, Vol. 6, 461-64.

Sin, C.-Y. and White, H. (1996), 'Information criteria for selecting possibly misspecified parametric models', *Journal of Econometrics*, Vol. 71, 207-25.

Taylor, L. D. (1994). *Telecommunications Demand Modelling: Theory and Applications*, Dordrecht, Kluwer.

White, H., (1989). 'Learning in artificial neural networks: a statistical perspective', *Neural Computation*, Vol. 1, 425-64, (reprinted in White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory* Oxford, Blackwell).

White, H. (1998). 'Artificial neural network and alternative methods for assessing naval readiness'. Technical Report, NRDA, San Diego.

White, H. (2000). 'A reality check for data snooping', *Econometrica*, 68, 1097-1126.

West, K. D. (1996). 'Asymptotic inference about predictive ability', *Econometrica*, 64, 1067–84.

TABLE 1

The RETINA algorithm[†]

Stage 0 - Preliminaries

    1. Data building and sorting

        a. Generate the set of transformed variables $\zeta(X) = \{W_1, \ldots, W_m\}$.

        b. Divide the sample into three sub-samples.

Stage I - Isolating a "candidate" model

    2. Using Data on the First Sub-sample

        a. Order the variables in ζ(X) according to their (absolute) sample correlation with the dependent variable in the first sub-sample alone. Let $W_{(1)}$ be the variable with the largest absolute correlation with Y, $W_{(2)}$ be the second most correlated, and so on.

        b. Consider various sets of regressors all of which include a constant and $W_{(1)}$: each set of regressors $\zeta_\lambda(X)$ is indexed by a "collinearity threshold" $0 \le \lambda \le 1$ and is built by including $W_{(j)}$ (j=2,…,m) in $\zeta_\lambda(X)$ if the R^2 of the regression of $W_{(j)}$ on the variables already included is $\le \lambda$.

        c. The number of sets of regressors is controlled by the number of values of λ between 0 and 1 chosen, say, ν.

    3. Using Data both on the First and Second Sub-sample

        a. Estimate each model by regressing Y on each set of regressors $\zeta_\lambda(X)$ using the data on the first sub-sample only and compute an out-of-sample prediction criterion (the cross-validated mean squared prediction error) using the data on the second sub-sample only. This involves the estimation of ν models.

        b. Select a "candidate" model as the one corresponding to the best out-of-sample performance $\zeta_{\lambda^*}(X)$.

Stage II – Search Strategy

    4. Using Data both on the Second and Third Sub-sample

        a. Search for a more parsimonious model: estimate all models including a constant and all the regressors in $\zeta_{\lambda^*}(X)$ one at a time in the order they were originally included, but also in the order produced by the procedure sub 2.a, this time on the basis of the correlations in the second sub-sample.

        b. Perform an evaluation of the models out-of-sample (using the data on the third sub-sample) calculating a performance measure (the cross-validated mean squared prediction error, possibly augmented by a penalty term for the number of parameters in the model).

Stage III – Model Selection

    5. Repeat Stage I and II Changing the Order of the Sub-samples; Produce a Candidate for Each Sub-sample Ordering

    6. Select the Model which has the Best Performance over the Whole Sample

---

[†] GAUSS Code for running RETINA is available upon request.

TABLE 2
The Data Generating Processes in the simulation experiments[†]

DGP 1: linear

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \sigma u_i \qquad i = 1,\ldots,n$$

where $\alpha_0 = \alpha_1 = \alpha_2 = 1$, x_{i1} and x_{i1} are jointly normal with correlation ρ between regressors equal to either 0.5 or 0.9. The error term u_i is i.i.d. N(0,1), σ is calibrated to achieve an average R^2 of the resulting estimated equations across replications equal to 0.25, 0.50 and 0.75, respectively.

DGP 2: ratio

$$y_i = \alpha_0 + \alpha_1 \frac{x_{i1}}{x_{i2}} + \sigma u_i \qquad i = 1,\ldots,n$$

everything else is as in DGP1 except that $\rho = 0.5$ only.

DGP 3: product

$$y_i = \alpha_0 + \alpha_1 x_{i1} x_{i2} + \sigma u_i \qquad i = 1,\ldots,n$$

and everything else is as in DGP1.

DGP 4: linear with binary regressor

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_3 x_{i3} + \sigma u_i \qquad i = 1,\ldots,n$$

with $\alpha_3 = 1$, and x_{i3} is a discrete explanatory variable which takes the value 1 with probability 0.5 and 0 otherwise, and everything else is as in DGP1.

DGP 5: linear with sparse regressor

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_4 x_{i4} + \sigma u_i \qquad i = 1,\ldots,n$$

with $\alpha_4 = 1$, and x_{i4} is i.i.d. N(0,1) and a correlation $\rho = 0.5$ with x_{i1} with probability 0.2 and zero otherwise, and everything else is as in DGP1.

DGP 6: linear with outliers

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \sigma v_i \qquad i = 1,\ldots,n$$

where everything is as in DGP1 except that we expect 5% outliers in the error term v_i which is equal to u_i when its absolute value is less than 1.96, and it is equal to u_i multiplied by 5 otherwise.

DGP 7: linear with structural break

$$y_i = \alpha_0 + \alpha_1^* x_{i1} + \alpha_2^* x_{i2} + \sigma u_i \qquad i = 1,\ldots,n$$

with $\alpha_1^* = \alpha_2^* = 1$ for $i = 1,\ldots,n/2$ and $\alpha_1^* = 0.5, \alpha_2^* = 1$ otherwise.

TABLE 3
Percentages of successful retrieval of the DGPs by RETINA
for different DGPs[†], sample sizes, and R^2's

DGP	Sample size	$R^2=0.25$	$R^2=0.5$	$R^2=0.75$
1: linear				
	100	*22.8*	*72.9*	*97.7*
	200	*42.8*	*93.9*	*98.3*
	1000	*98.6*	*99.1*	*99.1*
2: ratio				
	100	*39.9*	*72.6*	*93.7*
	200	*49.8*	*82.2*	*97.4*
	1000	*73.6*	*94.7*	*99.1*
3: product				
	100	*75.9*	*96.2*	*98.6*
	200	*94.2*	*99.1*	*99.1*
	1000	*99.5*	*99.4*	*99.2*
4: linear with binary regressor				
	100	*9.8*	*43.4*	*88.1*
	200	*24.6*	*72.6*	*95.5*
	1000	*89.5*	*95.7*	*95.9*
5: linear with sparse regressor[‡]				
	100	*NA*	*NA*	*NA*
	200	*6.8*	*32.8*	*66.8*
	1000	*62.6*	*92.6*	*97.1*
6: linear with 5% outliers				
	100	*19.5*	*64.9*	*95.4*
	200	*47.5*	*93.2*	*98.7*
	1000	*98.6*	*98.9*	*98.9*
7: linear with structural break				
	100	*9.6*	*38.9*	*85.9*
	200	*20.3*	*67.5*	*97.7*
	1000	*92.5*	*98.5*	*98.7*

[†] The correlation among the original variables X is $\rho=0.5$.

[‡] For this DGP, and n=100 we ran into the problem of zero regressors in sub-samples.

TABLE 4.

Percentages of successes of RETINA versus other procedures for DGP1 and $R^2=0.5$

Procedure	Sample size	$R^2=0.25$	$R^2=0.50$	$R^2=0.75$
RETINA				
	100	*22.8*	*72.9*	*97.7*
	200	*42.8*	*93.9*	*98.3*
	1000	*98.6*	*99.1*	*99.1*
Stepwise				
	100	*8.5*	*9.0*	*9.0*
	200	*8.2*	*8.2*	*8.2*
	1000	*8.7*	*8.7*	*8.7*
Non-negative Garrote				
	100	*3.8*	*5.0*	*21.3*
	200	*10.1*	*11.3*	*32.4*
	1000	*34.2*	*54.3*	*73.4*

Figure 1: Percentages of successful retrieval of the DGPs by RETINA for different DGPs, sample sizes, and R^2's
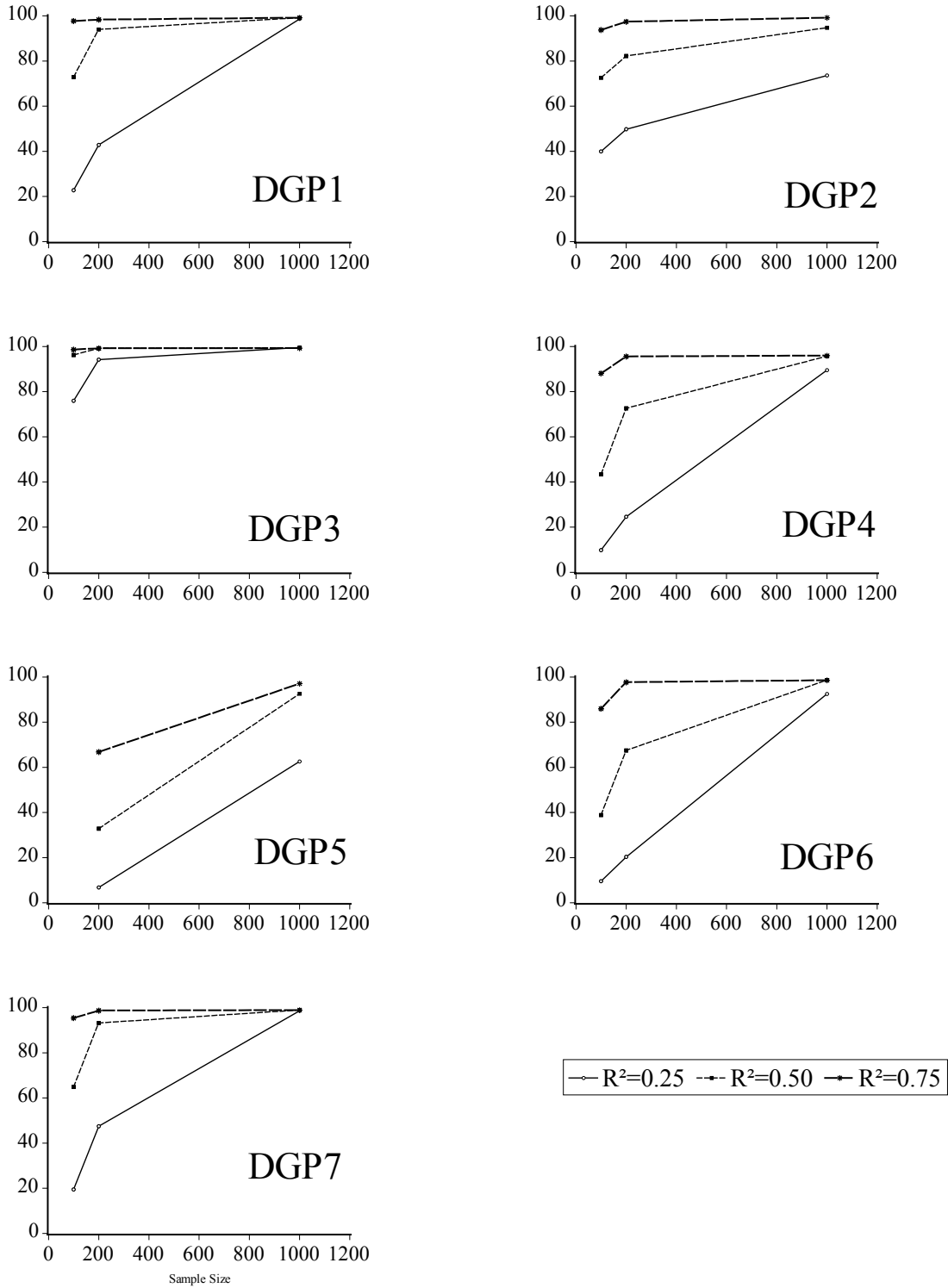
Figure 2: Performance comparison between versions of RETINA: Data from DGP1
Panel (A) without the repetition of the procedure on different sub-sample reordering
Panel (B) same as (A) and without the penalty term in the forecast performance criterion
index