ニューラルネットワークと線形回帰分析の
ハイブリッド解析法の実際
―建築コスト価格モデルへの適用

浅野　美代子・ユー　K. W.　マーコ

An Introduction to the Hybrid Approach of Neural Networks and the Linear Regression Model: An Illustration in the Hedonic Pricing Model of Building Costs

Miyoko ASANO
Department of Law, Daito Bunka University
Marco K. W. YU
Bartlett School of Graduate Studies, University College London

Abstract

This paper introduces the hybrid approach of neural networks and linear regression model proposed by Asano and Tsubaki (2003). Neural networks are often credited with its superiority in data consistency whereas the linear regression model provides simple interpretation of the data enabling researchers to verify their hypotheses. The hybrid approach aims at combing the strengths of these two well-established statistical methods. A step-by-step procedure for performing the hybrid approach is presented and a hedonic building price model is used to further illustrate the *modus operandi* of the hybrid approach. Our analysis on building price is based on the data in Cheung and Skitmore (2006), and we find that the hybrid approach improves their model by a better structure of the input variables for the building prices.

1. Introduction

The salient benefits of the linear regression model are its clarity of the relationship between the variables at issue and the tractability of the calculation. The concept of fitting the data with a line by minimising the sum of the squared errors is comprehensible to many and the estimated equation provides straightforward interpretation, rightly or wrongly, of the relationship between the variables under study. These make it popular in non-experimental science research such as economics. However, the theories in non-experimental subjects in general and economics in particular rarely imply any specific functional forms and thus mis-specification is one of the problems of using the linear regression model to test the theories or to produce forecasts.

Neural networks, being one of the most flexible non-parametric modelling techniques, are appealing alternatives for prediction of a concerning variable due to its ability in pattern recognition from the statistical perspective (Bishop 1995 and Ripley 1996). Diaconis and Shahshahani (1984) illustrated that neural networks can be regarded as consistent and non-parametric estimates of nonlinear regression functions. Asano et al (2002) further demonstrated the marked advantage of neural networks to approximate structural changes of the population regression function efficiently over other non-parametric regression methods. Owing to its plasticity, neural networks are believed to have the advantages in data consistency and forecast performance.

Asano and Tsubaki (2003) proposed a hybrid approach to the neural networks and linear regression analysis, which draws on the strengths of both methods. The objective is to develop a modelling method with high lucidity and forecast accuracy.

In the remaining of this section, we will explain the hybrid approach with a view to enhancing its accessibility. Section 2 will report the hedonic building price model in Cheung and Skitmore (2006) and we will carry out the hybrid approach analysis of the hedonic building price model in section 3. We will finally conclude the paper in section 4.

1.1 Hybrid approach to neural network and linear regression analysis

In this section we would explicate the hybrid approach proposed by Asano and Tsubaki (2003).

1.1.1 The model and the statistics

Let x be the p-dimensional input vector, and y be the corresponding output variable. Then a hybrid statistical model used for data editing becomes the following

$$y=\beta_0+\sum_{j=1}^{p}\alpha_j x_j+\sum_{i=1}^{q}\beta_i f_i\left(\sum_{j=1}^{p}w_{ij}x_j\right)+\varepsilon,\ \varepsilon\sim N(0,\sigma^2) \qquad (1)$$

$$= A + B + \varepsilon,$$

where, $A=\beta_0+\sum_{j=1}^{p}\alpha_j x_j$, $B=\sum_{i=1}^{q}\beta_i f_i\left(\sum_{j=1}^{p}w_{ij}x_j\right)$ and $f(t)=\dfrac{1}{1+\exp(-t)}$ is the sigmoid function, under the condition that errors ε are independent, normal with constant variance σ, and q is the number of units of the hidden layers.

We proposed a three-step regression (procedure of the three-step regression as below) that corresponds approximately to model (1). To estimate model (1), we proposed the procedure for hybrid approach to the neural network and linear regression analysis (Asano and Tsubaki, 2003) as follows.

1.1.2 Procedure of hybrid analysis

Step 1: By analysis of the neural network using Kurita (1990)'s method of minimizing the Akaike Information Criteria (1974), we decide the numbers of the hidden layers (q). The formula of the Akaike Information Criteria (AIC) in our case is as follows:

$$AIC = n(LnRSS\,/\,n) + 2((p+1) * q + q)$$

where RSS is the residual sum of squares. The output values of q units of the hidden layer are added as the input variables to B of model (1).

Step 2: A stepwise analysis of multiple regression is carried out based on F-test analysis using significant level α in the forward selection and backward elimination methods. Here, the p variables of A of Model (1) are all included. The variable selection process focuses on the hidden layer's output q variables from Step 1 using forward selection. The q' variables are to be selected from the q variables. Therefore, the input variables are now $p + q'$.

Step 3: In order to select the appropriate input variables from the original p input variables, we do an F-test analysis using the usual significant level α in the backward elimination method. All q' variables from Step 2 are included. The model selection process concentrates on the p variables and p' variables are to be selected from the p variables.

After Step 3, we automatically obtain the model of $p' + q'$ input variables. The resulting model is as follows:

$$\begin{aligned} y &= \beta_0 + \sum_{j=1}^{p'} \alpha_j x_j + \sum_{i=1}^{q'} \beta_i f_i\left(\sum_{j=1}^{p} w_{ij} x_j \right) + \varepsilon \\ &= A + B + \varepsilon, \end{aligned}$$

where, $A = \beta_0 + \sum_{j=1}^{p'} \alpha_j x_j$, $B = \sum_{i=1}^{q'} \beta_i f_i\left(\sum_{j=1}^{p} w_{ij} x_j \right)$, and $f(t) = \dfrac{1}{1+\exp(-t)}$ is the sigmoid function, under the condition that errors ε are independent, normal with constant variance σ, and q is the number of units of the hidden layer. Therefore A is the linear regression component; B is the data-editing component.

2. A Hedonic Building Price Model

No matter building a new home or constructing a new office, it is one of the most important investment decisions made by the individual or the organisation. The importance of good building price model cannot be emphasised more. As a buyer of the building, the client would like to obtain accurate forecast of the contract price of constructing his building, so that he can decide the viability of the project and make the corresponding financial arrangement. Moreover, he would also like to understand the drivers of the building price so that, with the advice of his consultants, he can make the best trade-off between the price and other characteristics of the building. Skitmore and Marston (1999) is a collection of the effort on building price modelling mainly made by the academics. Linear regression analysis is one of the popular ways to construct the building price models. On the other hand, Emsley et al (2002) is a solid attempt to apply neural networks to building price modelling and has demonstrated that the ability of modelling the nonlinearity in the data and better fit of the data (measured in mean absolute percentage error) are two benefits of neural networks.

It is generally believed that the larger the building measured by floor area, the higher the construction price of the building. Construction cost consultants, therefore, tend to capture the construction price in a single rate -construction price per floor area- for buildings with different functions (e.g. residential, industrial and office) and qualities (e.g. luxury, high quality and standard). However, this simple modelling strategy will fail to capture any effects of the building shape, height, and the existence of basement to the building prices, so James (1954) proposed a more sophisticated single-rate method about the measure of the storey enclosure rather than the conventional floor area. A number of the characteristics of buildings such as the building shape, floor area, vertical position of the floor area, storey height, total building height and some measures of basement are considered in James' model. Against this background, Cheung and Skitmore (2006) revisited the James' storey enclosure model with 50 empirical data of residential building collected in Hong Kong and claimed that they have found the best model. We will return to their modelling strategy and result later in this section.

Most of the building price modelling research such as Cheung and Skitmore (2006) are motivated by data consistency and forecasting ability. Some of them try to relate their pricing models with the historical production cost of different components of buildings such as Flanagan and Norman (1978) on heights of the buildings and Chau (1999) on the plan shapes of the buildings. However, these pricing models bear no resemblance to the standard economic theory that the price of a product is simultaneously determined by the demand of the consumers and the supply of the producers.

The hedonic[1] hypothesis provides the essential linkage of these pricing models with the standard economic theory. It asserts that the consumer value the products for their utility-bearing characteristics such as floor area for buildings, memory size for computers and power for automobiles. The products are regarded as bundles of their characteristics and the prices of the products are summations of the prices of their characteristics. The hedonic hypothesis implies that the opportunity cost of producing the products also depend on the bundles of characteristics. Rosen (1974) demonstrated that the prices of the characteristics (i.e. the coefficients of the linear hedonic pricing model) of the product are determined by the distributions of producer costs and consumer tastes[2]. He also illustrated that economic theory generally cannot specify a functional form for the hedonic pricing model and consequently the choice of functional form should be an empirical issue.

2.1 Cheung and Skitmore (2006) Model

Cheung and Skitmore (2006) revisited the model proposed by James (1954) and built a hedonic model of building price per m^2 with various measures of the physical sizes of the high-rise buildings. The physical sizes are about three components of a high rise building namely the basement, podium and tower, but not all of the buildings in the data possess basements or podium floors. The model is motivated by the belief that the building prices are affected by building shape measured by the length of perimeter on plan, average floor area, average storey height, total building height, and the existence of the basement.

The best model they found is called 'Regressed Modified Model for Amended Storey Enclosure Method' (RMASEM). They have adopted leave-one-out cross validation and step-wise regression technique to select the input variables out of 19 potential input variables. In brief, the strategy they use is mainly to select the best-fitting model while imposing some penalties for unstable coefficients in the cross validation method (Leamer 1983).

Below is Cheung and Skitmore (2006)'s RMASEM:

$$Y = \beta_0 + \beta_1 \cdot spt + \beta_2 \cdot fb + \beta_3 \cdot pb + \beta_4 \cdot fpt + \beta_5 \cdot sb \qquad (2)$$

where β_0, β_1, β_2, β_3, β_4 and β_5 are the coefficients to be estimated and the definitions of the input variables are tabulated below.

[1] For a comprehensive treatment of the hedonic function, please see Triplett (2004).

[2] This heterogeneous consumers situation is very true in this application because the 'taste' or preference of the consumers of the building is highly influenced by the land value, development controls and the legislations.

Table 1: Variables of the hedonic building price model.

Variables	Definitions
Y	It is the building price per floor area measured in m^2. The building price refers to the accepted tender price but it excludes the prices for foundations, building services, external works, preliminaries and contingencies. Since it is not the property price, it also excludes the land value. It is adjusted for inflation by the relevant tender price index produced by Levett and Bailey Chartered Quantity Surveyors Ltd.
spt	It is the average storey height of the podium and tower and is defined by the formula *spt* = (*a***sp*+*b***st*)/(*a*+*b*); *a* is the number of storey of the podium; *sp* is the average storey height of the podium; *b* is the number of the storey of the tower; *st* is the average storey height of the tower.
fb	It is the average area per storey for the basement.
pb	It is the average perimeter on plan for the basement. For example, if the basement is a 20m x 10m rectangular shape. Then the perimeter on plan is the sum of the length of all four sides of the rectangular which is 60m.
fpt	It is the average area per storey for the podium and tower and is defined by the formula *fpt* = (*a***fp*+*b***ft*)/(*a*+*b*); a is the number of storey of the podium; *fp* is the average floor area for the podium; *b* is the number of the storey of the tower; *ft* is the average floor area for the tower.
sb	It is the average storey height of the basement.

This model means that the building price per m^2 are correlated with the average storey height of the podium and tower, average area per storey for the basement, average perimeter on plan for the basement, average area per storey for the podium and tower and average storey height of the basement.

2.2 The Linear Regression Result and Discussion

The estimated coefficients and the corresponding t-values of the model (2) using Cheung and Skitmore (2006)' s dataset are reported in the following table.

Table 2. Result of the linear regression[3]

Variables	constant	*spt*	*fb*	*pb*	*fpt*	*sb*
Coefficient	-6126.35	3797.07	0.70	-3.52	-0.17	-165.95
t-value	-3.17	5.67	2.54	-1.98	-2.25	-1.71

All variables are significant at 10% level. The positive coefficient for *spt* is as expected since the dependent variable is the total building price per floor area, other things being equal, the higher the average storey height of

[3] This result is in line with that reported on Cheung and Skitmore (2006: pp. 398). The minor difference is believed to be due to the rounding of the data and the adoption of different software packages to carry out the liner regression analysis.

the superstructure (podium and tower), the higher the price per m^2 because constucting higher walls per storey should cost more. The positive coefficient for *fb* also appears to be reasonable since one would expect a larger basement would push up the total building price on average. The reason is that the construction of basement involve expensive excavation work, which does not apply to the podium and tower. The negative coefficient for *fpt* also seems in order because of the economy of scale. Since the majority of the floor area is in the podium and tower, *fpt* may be a good proxy of the size of the building project. One would generally expect that the larger the size of the project, the lower the per m^2 price.

However, the negative coefficients for *pb* and *sb* appear to be perverse. Both variables relate to basements which should have a positive impact on the price per m^2 because of the expensive excavation work involved. Apart from the effect of excavation, *pb* and *sb* are more likely to have a positive effect on *Y*. The effect of average storey height (*sb*) is explained above, and *pb* refers to the perimeter on plan which should be positively correlated with the amount of external wall. Therefore, holding the area (*fb*) constant, the larger the *pb*, the higher the ratio of the amount of external wall to the floor area. Since external wall is believed to be a relatively expensive element of buildings, a higher *pb* should be correlated with a higher *Y*.

Rosen (1974), however, remind us that the structural interpretation of the hedonic model is not available since the hedonic model is a complex reduced form of the underlying demand and supply model. Pakes (2003 and 2005) justified the improbable negative coefficients for utility-bearing attributes by imperfect competition which is plausible in differentiated product market that products are heterogeneous.

Another possibility of the perverse signs is the mis-specification of the model that some variables are omitted, the functional form adopted is incorrect or both.

3. Analysis Using the Hybrid Approach

We analyse the hedonic building price model by the procedure for the hybrid approach laid down in section 1. We show the procedure by using this case. The previous applications of the hybrid approaches were to time series data (Asano and Tsubaki 2003 and Asano et al 2006) and this is the first attempt to extend the application to cross sectional data.

The model (2) has 5 variables excluding the constant, so p=5. We analyse the hedonic building price model by using the fit single-hidden-layer neural network NNET function of S-plus. Before the analysis we transform the output variable *Y* to [0, 1], and the 5 input variables are standardised. We show the result of the linear regression in table 3.

Table 3. Result of the linear regression on transformed data

Variables	constant	*spt*	*fb*	*pb*	*fpt*	*sb*
Coefficient	-0.523	0.160	0.604	-0.479	-0.067	-0.063
t-value	20.95	5.67	2.54	-1.98	-2.25	-1.71

Residual standard error: 0.1765 on 44 degrees of freedom

Adjusted R Squared: 0.5375

3.1 Step 1 of the procedure.

First of all we show the result of Step 1. We analyse data using model (3). In order to fix the q, we calculate AIC using various q.

$$y=\beta_0+\sum_{i=1}^{q}\beta_i f_i\left(\sum_{j=1}^{5} w_{ij}x_j\right)+\varepsilon, \tag{3}$$

We show the *AIC* of q=1,2,3 and 4 in table 4.

Table 4: *AIC* of neural networks (q=1,2,3 and 4)

Number of unit of hidden layer: q	*AIC*
1	-180.4
2	-174.7
3	-190.2
4	-172.1

We fix three units of hidden layers from the values of *AIC*. We show all weightings of the feed-forward three layers neural networks on table 5.

Table 5: Weights of Neural Networks (Step1)

inputvariable	W_1	W_2	W_3
constant *spt* *pb* *fpt* *sb*	4.2 16.7 -0.5 1.4 -7.4 8.3	8.6 27.7 25.1 -23.6 -5.6 1.7	20.5 -27.6 -15.0 -6.2 13.5 7.6
b_0	b_1	b_2	b_3
0.696	-0.507	0.655	-0.366

We also show the Sigmoid Decomposition Graph on figure1.

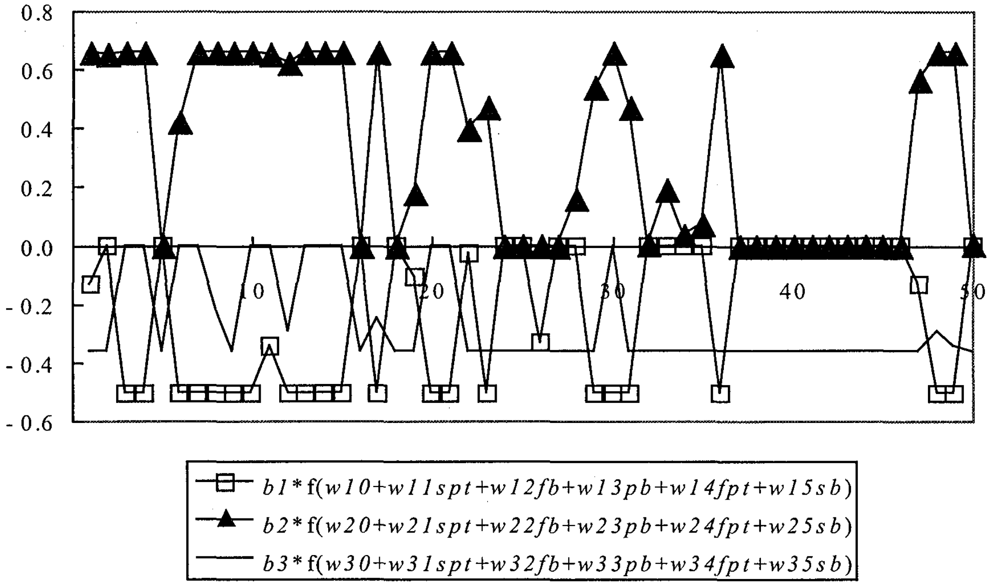

Figure1: The Sigmoid Decomposition Graph of input variables of the building price model. (q=3)

If the data is a time series and the x-axis denotes time, the Sigmoid Decomposition graph sometimes shows structural changes of the data. Here, we order the values of the model's components by the values of the predicted Y. Figure 2 shows the Sigmoid Decomposition Graph of building price. The first and second outputs of neural networks are added up and the third output of neural networks is shown separately.

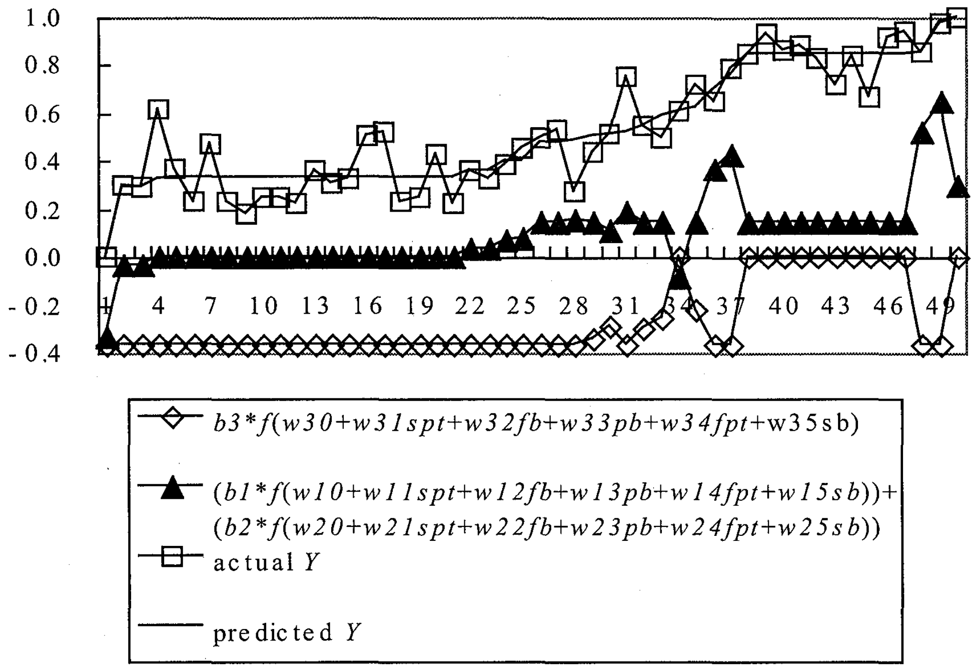

Figure 2. The Sigmoid Decomposition Graph of building price

3.2 Step 2 of the procedure

A stepwise analysis of the multiple regression is carried out based on the F-test analysis using significant level 0.01 in the forward selection and backward elimination methods. Here, all 5 variables of part A of Model (1) are included. The variable selection process focuses on the hidden layer's 3 output variables from Step 1 using forward selection. By using ANOVA of S-plus the first and second variables are selected from the 3 variables based on the stepwise analysis. Therefore, q'=2 and the input variables are 5+2 = 7 in Step 2. We show the result of the regression in table 6.

Table 6. Result of the linear regression in Step 2

variable	constant	*spt*	*fb*	*pb*	*ftp*	*sb*	*out1*	*out2*
coefficient	0.4318	0.115	0.227	-0.123	-0.040	-0.042	0.837	0.868
t-value	10.00	4.08	1.25	-0.67	-1.61	-1.31	4.37	6.57

Residual standard error: 0.1267 on 42 degrees of freedom

Adjusted R Squared: 0.7673

3.3 Step 3 of the procedure.

In order to select the appropriate input variables from the original 5 input variables of part A of model (1), we do an F-test analysis using the usual significant level 0.05 in the backward elimination method. All 2 variables of part B of the model from Step 2 are included. The model selection process concentrates on the 5 variables of part A of model (1). We found that the input variables *fpt* and *sb* were dropped from the part A. Therefore, the input variables are now 3+2 = 5. We show the coefficients of Step 3 on table 7.

Table 7: Result of the linear regression on the hybrid model.

Variables	constant	*spt*	*fb*	*sb*	*out1*	*out2*
Coefficient	0.399	0.108	0.0923	-0.0618	0.773	0.923
t-value	10.35	3.87	3.62	-2.06	4.21	7.35

Residual standard error: 0.1277 on 44 degrees of freedom

Adjusted R Squared: 0.7580

3.4 Result of the hybrid approach and consideration.

We show the regression results of Steps 1, 2 and 3 in Tables 3, 6 and 7. The adjusted R-Squared has increased from 0.5375 before Step 1 to 0.7580 in Step 3 which indicates a marked improvement in data consistency. Also *pb*, which is believed to have perverse sign, is dropped from the linear component of the model.

We consider the meaning of *out1* and *out2*. In Figure 1 we noticed that *out1* and *out2* should be added up, so we show *out1+out2* in figure 3. Here in order to consider the result of regression in table 7, we show the Asano-Bhattacharyya graph (2006) ordered by the *spt*.

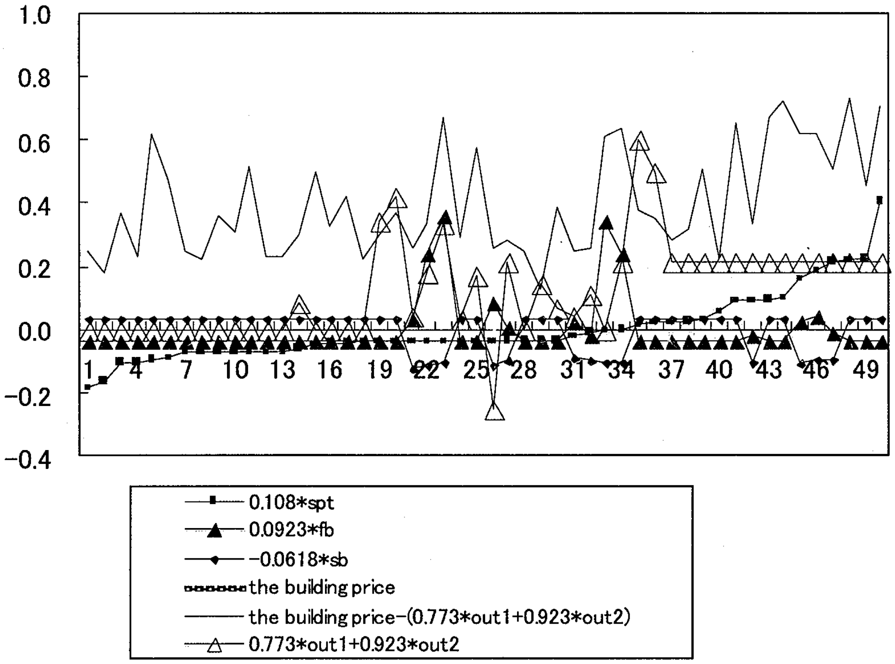

Figure3: Asano-Bhattacharyya graph (2006) ordered by the *spt*

In figure 3, we notice that by the order of *spt*, the △- line (0.773 * *out1*+0.923 * *out2*) can crudely be divided into three segments:

i) the △- line (data-editing component of the hybrid model) equates to 0 when *spt* is less then 2.75m;

ii) the △- line varies when *spt* is between 2.75m and 2.84m;

iii) the △- line equates to 0.2124 when *spt* is larger than 2.84m.

The tentative interpretation of the result is that the true model is highly non-linear, so Cheung and Skitmore (2006) mis-specified the model by using a linear regression model. It is plausible since Ekeland et al (2004) demonstrated that the nonlinearities are the generic features of the equilibrium in hedonic models. Also some valid variables may be omitted from their model such as some measures of the internal finishes used in each project.

4. Conclusions

We have laid down a step-by-step procedure for the hybrid approach which aims at combining the clarity for interpretation in the linear regression model and data consistency in neural networks. The procedure has been applied to the hedonic building price model proposed by Cheung and Skitmore (2006). The hybrid approach has improved the model remarkably in terms of data consistency and alluded that the underlying true model may be non-linear. We hope this paper makes the hybrid approach more accessible to other researchers and students.

References

Akaike, H. (1974) "A New Look at the Statistical Model Identification" *IEEE Transactions on Automatic Control*, 19, 716-723.

Asano, M., Yu, M., Bhattacharyya, P. and Tsubaki, H. (2006) "A Hybrid Approach to Neural Networks and Linear Regression: A British Tender Price Index Modelling", the 2006 Japanese Joint Statistical Meeting, 5-8 September, Tohoku University, Sendai City, Miyagi.

Asano, M., and Tsubaki, H. (2003) "Hybrid Approach with Neural Networks and the Linear Regression Analysis and its Application to Prediction of the Amount of Water Supply in Tokyo 23 Districts", *Japanese Journal of Applied Statistics*, Vol. 31, No. 3, pp. 227-238 (in Japanese).

Asano, M., Tsubaki, H., and Yoshizawa, T. (2002) "Effectiveness of Neural Networks to Regression with Structural Changes", *Applied Stochastic Models in Business and Industry*, Vol. 18, pp. 189-195.

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Chau, K. W. (1999) "On the Issue of Plan Shape Complexity: Plan Shape Indices Revisited", *Construction Management and Economics*, Vol. 17, pp. 473-482.

Cheung, F. K. T. and Skitmore, M. (2006) "A Modified Storey Enclosure Model", *Construction Management and Economics*, Vol. 24, pp. 391-405.

Diaconis, P., and Shahshahani, D. (1984) "On Nonlinear Functions of Linear Combinations", *SIAM Journal on Scientific and Statistical Computing*, Vol. 5, No. 1, pp. 175-191.

Ekeland, I., and Heckman, J. and Nesheim, L. (2004) "Identification and Estimation of Hedonic Models", *Journal of Political Economy*, Vol. 112, No. 1, Part 2, pp. S60-S109.

Emsley, M., Lowe, D., Duff, A., Harding, A. and Hickson, A. (2002) "Data Modelling and the Application of a Neural Network Approach to the Prediction of Total Construction Costs", *Construction Management and*

Economics, Vol. 20, pp. 465-472.

Flanagan, R, and Norman, G. (1978) "The Relationship between Construction Price and Height", *Chartered Surveyor B and QS Quarterly*, Summer, pp. 69-71.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer.

James, W. (1954) "A New Approach to Single Price-Rate Approximate Estimating", *RICS Journal*, Vol. 33, No. 11, May, pp. 810-824.

Kurita, T (1990) "A Method to Determine the Number of Hidden Units of Three Layered Neural Networks by Information Criteria", *Transaction of Institute of Electronics, Information and Communication Engineers*, Vol. J73-D-II, No.11, pp.1872-1878 (in Japanese).

Leamer, E. E. (1983) "Model Choice and Specification Analysis", in Griliches, Z. and Intriligator, M. D. (eds.) *Handbook of Econometrics*, Vol. 1, pp. 285-330.

Pakes, A. (2003) "A Reconsideration of Hedonic Price Indexes with an Application to PC's", *American Economic Review*, Vol. 93, No. 5, pp. 1578-1596.

Pakes, A. (2005) "Hedonics and the Consumer Price Index", manuscript.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Rosen, S. (1974) "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy*, Vol. 82, No. 1, pp. 34-55.

Skitmore, M., and Marston, V. (1999) *Cost Modelling*, London: E&FN Spon.

Triplett, J. (2004) "Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Product", *OECD Science Technology and Industry Working Papers*, 2004/9, OECD Publishing.

(2006年9月25日受理)