Land price modeling with genetic algorithms and artificial neural network procedures

Dominique Fischer [1] and **Peddy Lai** [2]

[1] Department of Estate Management , Universiti A, Malaysia
[2] Peddy, Pi-Ying Lai, National Pingtung Institute of Commerce, Taiwan

Corresponding author: domfischer@gmail.com

Abstract

Genetic programming (GP) is a heuristic procedure that can favourably be compared to parametric treatments or distribution free Artificial neural network treatments. This article provides a brief description of the genetic programming algorithm used here to predict land prices in the city of Kaoshiung (Taiwan). The GP results are then compared to results obtained, from the same data set, from two different ANN models and with a standard multiple regression models.

Genetic programming's results are comparable in precision to the ANN results. However both ANN models perform much faster and one of them (labeled ANN2) provides much more explicit information.

This article is simply an exploration in the applicability of GP to problems that do not respond well to the requirements of parametric statistical models.

Keywords: artificial neural networks, genetic programming, land prices, Taiwan.

Introduction

Land price predictions or valuation have traditionally relied on multiple regression treatments (so called hedonic analysis). However, the standard econometric approach to price determination requires strong assumptions on the statistical characteristics of the input variables. Most of the time, the required statistical requirements are brushed away and thus the results may not be entirely convincing from a strict statistical inferential view point.

New generations of non-statistical models are now available at reasonable prices and learning time investment costs. These models can produce accurate price predictions or valuations without having to violate any of the sacrosanct statistical hypothesis. These instruments are purely heuristic models that In fact, require no hypothesis at all.

Artificial Neural Network (henceforth noted ANN) has been applied to similar property related problems for quite a few years now [i]. The procedure works well but it is still far of being a standard tool in property analysts' toolbox.

On the other hand, genetic programming (henceforth GP) that provides a more powerful and versatile family of algorithms has not yet been applied in our field. This paper thus experiment with the instrument and compares the relative performance and usability of GP and other artificial intelligence based programs such as artificial neural network.

The three types of models (GP, ANN and standard regression analysis) are applied to a set of data collected in the Kaoshiung Municipal territory for year 2006. The land prices per square meters are market prices and they are 'explained' by the same variables that are used by the Kaoshiung valuation authorities.

Genetic programming

Pioneered by John Holland in the 60s, Genetic Algorithms has been widely studied, experimented and applied in many fields in engineering worlds. GP is now also tentatively applied to economic and financial problems ii. As far as we could find out, it has not yet been applied to real estate pricing issues

Genetic programming relies on adaptive heuristic search algorithms that can be analogized to a Darwinian natural selection through mutations, combination and elimination. The analogy is, of course, only an evocative metaphor. As cautioned elsewhere (Fischer, 2007) if the metaphor should not be overplayed it certainly provides a language that helps the description of the GP principles and algorithms. GP does not 'optimise' solutions, but it gradually selects the 'fittest' among competing sets of programs. The heuristic simile could be described as follows.

From an initial set of data ('the primordial soup') simplistic models are built to combine all the existing inputs in order to obtain a certain result (the target). In fact the so-called models are simply 'bit strings' of fixed length with values limited to 0 and 1. The results are turned into decimal numbers and compared in their proximity to the targeted results. The 'fitness' of each string is ranked and the ranking is used to proceed to the next set of calculations. Among a large number of alternative combinations and transformations of the input variables, the best performers (the one that get closer to the target) are selected and two of the 'survivors' are chosen and 'mated' by combination of their components. The recombined model (the descendant) is then including with the surviving models and entered again in the tournaments. The process is repeated over and over again (over 90 000 'tournaments' in this land pricing example).

The 'models' are build by the combination of the most basic manipulations of the data. The required transformations are limited to the following operations:

1. Arithmetic operations such as addition, subtraction, multiplication, division, absolute values, square roots. These operations are used to combine numerical constants.
2. Boolean operators such as AND, OR, GREATER THAN, TRUE, FALSE.
3. Conditional logical operators such as: IF, IF THEN,

values, multiplications and square roots. And a typical formulation could look like:

```
abs(v[0]) - 0.5) + v[0]) + ((abs(v[0]) - 0.5) + v[0])) - sqrt(0)) + v[0]) / 0.5);
```

This code line illustrates how one variable (Vo) is modified by very simple operations.

The best performing program is then selected and can be used to apply as a price predicting instrument. However the output of this particular software is presented in a programming language (Assembler or C++) and the program must be then integrated and run separately.
A mathematical solution can be offered for the simples problems but, in our case, the solution is too long to be provided as an output.

The objective of this experimental research is simply to compare the results obtained from Genetic programming, ANN and multiple regression analysis of the same set of land price data.

The data set

The initial set of data (around 500 data points) has been cleaned up and massaged in order to eliminate missing data and dubious cases. The resulting set was reduced to 386 data points. Summary statistics are presented in table 1.

Table 1: The variables used in the comparative analysis

	1	2	3	4	5	6	7	8
	distance from station	zoning coded	Road width	near road	usage	FAR	area	$/m3
Mean	5108.16	11.71	14.62	0.36	3.74	3.89	202.28	5.03
Standard Error	172.33	0.18	0.43	0.02	0.06	0.09	37.35	0.18
Standard Deviation	3732.13	3.90	9.30	0.48	1.32	1.95	808.87	3.99
Kurtosis	-0.87	-1.34	5.97	-1.64	-0.35	0.13	206.79	26.40
Skewness	0.87	-0.62	1.92	0.60	-0.97	1.18	12.74	3.72
Range	10297.00	15.00	56.00	1.00	4.00	6.90	14375.57	44.90
Minimum	1339.00	2.00	4.00	0.00	1.00	1.50	0.43	1.00
Maximum	11636.00	17.00	60.00	1.00	5.00	8.40	14376.00	45.90

Basic exploratory cross tabulations are presented in appendix 1. These cross-tabulations are used to screen the pertinent variables and to provide some indications of the signs of the co-variations.

1. Land prices per square meters have been collected for 11 districts of the Kaoshiung municipal. The counties are listed in column 1 and their distances to Kaoshiung central station was calculated (column 1 of table 1).

2. In column 2 The authorised zoning is coded according to the following categories (c2=Commercial 2 R4=Residential 4 SR=Specific Retail A=Agriculture

3. The width of the trunk road connection is given in meters in column 3. Presumably wider road indicate better accessibility and should thus have a positive influence on prices. This proximity indication is complemented by the measure of the distance to the main artery (near = 1, not near = 0) in column 4. In fact, to anticipate our results, these two variables do not appear to act as 'accessibility' proxy but more as 'nuisance' proxies.

4. Column 5: The effective usage of the lots are then coded as Use (0=residential 1=shop & house use 2=parking lot 3=warehouse 4=retail shop 5=factory 6=office). However, since no residential (0) appear in column 4, we had to presume that residential and shop activities were combined in the data collection. This should not be too surprising: this 'housing-shop' configuration is very common in Taiwanese cities.

5. The authorised Floor to Land ratios are given in column 6 and Lots surfaces are in column 7. This variable may have been thought to help capturing some scale effect on land pricing. In fact, it did not…

6. The only other information used in the calculation are price per square meter (column 8). This variable is the target of the GP and ANN calculations and the dependent variable in the multiple regressions.

The data was cleaned up and then subdivided in two subsets. A first set of 200 data points used as the 'training set' and a set of the remaining 186 data points used as the 'validation set'. This organisation of the data sources is required by the GP software used here. The training of the model is done on the first set and then selectively tested on the validation set.

Result of the multiple regression analysis

The Multiple regression model could be written as:

Land price per m2 = f (distance from station, road width, road proximity, zoning, FAR, Usage, lot size) + error term

The model is not overly sophisticated, however this limited set of untransformed inputs seems to be doing a satisfactory job at explaining land prices.

The regression statistics are not spectacular but quite good for this type of treatment. A coefficient of determination of 48% indicates that almost half of the price per square meters can be explained by our six variables. The overall results are significant (F = 25.6) and the input variable coefficient have the expected signs (more or less) and are all strongly significant except for the variable ‘usage’

Table 2: Multiple regression on land per m2. Kaoshiung districts

Regression Statistics	
Multiple R	0.696346129
R Square	0.484897932
Adjusted R Square	0.466019846
Standard Error	2.242606727
Observations	200

For our purpose, the interesting results are mostly in the size and dispersion of the residuals (the gap between the observed prices and the prices predicted by the model).

Obviously the mean of the residual is almost zero (-1.52642E-15) but the standard error of 2.20 Taiwan dollar per square meter is not negligible. This variability must be compared to the average price of 4.76 NTW). To anticipate our next results, this standard deviation is much larger than the ones obtained either with ANN or GP testing.

The ANN results

Two different ANN software packages were tested. The results were very close (with respective standard deviation of residuals being:

Table 3: Comparing targets and results with two different ANN packages

	ANN 1	ANN 2
Mean difference between target and prediction	-0.35	0.16
Standard deviation of the differences	1.89	1.51

So the second model (ANN 2) did perform better but, to boot, its output is much more ‘presentable’ that the output of the competing software package.

The graphic presentation of the results is sufficiently evocative to illustrate our results in table 4 and figure 1.

Table 4: ANN1 results

	Training set	Test set
# of rows:	165	34
CCR:	n/a	n/a
Average AE:	0.633056749	1.10395072
Average MSE:	2.016284842	3.6459857
Tolerance type:	Relative	Relative
Tolerance:	10%	30%
# of Good forecasts:	119 (72%)	19 (56%)
# of Bad forecasts:	46 (28%)	15 (44%)

RSquared: 0.7551
Correlation: 0.8800

Figure 1: Actual vs. Forecast

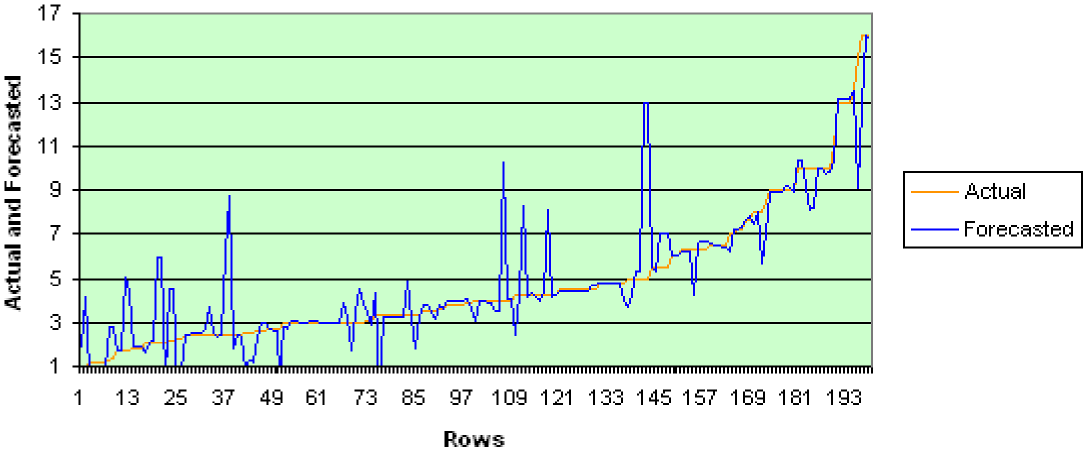
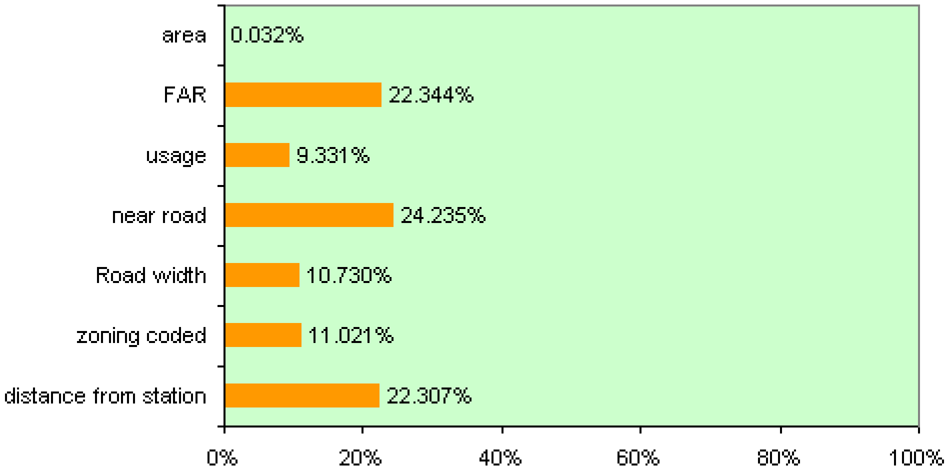


Figure 2: Relative input importance



One of the problems with these results is that you can identify the strength of the variables impacts but not their signs. For example here we can observe that both the Floor Area Ratios (FAR) have more or less the same weight than the distance from the station however we should know that, presumably, that one has a negative effect (distance) whereas the other (usage density) has a positive effect.

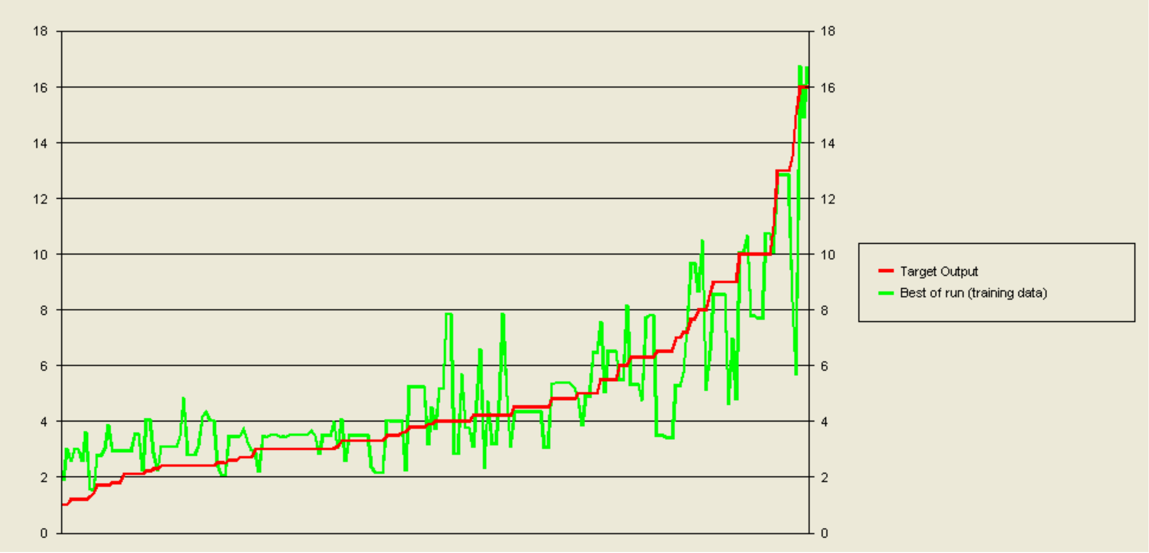
It is one of the reasons why regressions, or at least thorough cross tabulations, are always required when you perform this type of analysis. To be facetious, we could conclude that these intelligent models are intelligent enough to calculate the impact to the last 10 decimals, but too dumb to tell us the direction of this impact.

Genetic programming results

We must now admit that the results of the Genetic programming software tested here are much less 'presentable'. However, their repeated performance (at least for this limited example) is more stable and accurate and we conjecture that a better (but much more expensive) package could provide more attractive and usable outputs.

The graphic description of the 'evolutionary' drift toward the best fit is presented below:

Figure 3: The Genetic programming rambling through toward the target



And the summarised results can now be compared with the previous 'ANN winner'

Table 5: Comparing ANN2 and GP deviation from the target

	ANN 2	GP
Mean difference between target and prediction	0.16	-0.15
Standard deviation of the differences	1.51	0.54

For all the efforts involved in learning and running (over and over again…) the Genetic programming model we could not call this a spectacular victory. However, GP wins in accuracy.

Both models should now be tested on a real predictive task: this will be the topic of further research and, probably, the access to more flexible Genetic programming packages.

Conclusion

This brief methodological papers compare old stuff with new stuff and even newer stuff. In fact, none of the new stuff is really new. Artificial intelligence algorithms have been with us for over 30 years. They still have had little impact on our discipline, where we typically favour the old stuff.

To a large extent we are perfectly justified to rely on well worn statistical traditional tools as long as we are aware of their limitations. They work reliably and are sufficiently widespread to allow comparative analysis.

Artificial Neural Network and Genetic programming still do not have this advantage of comparability and certainly have no advantage of transparency. For this reason at least, we cannot see how they will increase their role in property analysis outside academia.

This article confirms that Genetic programming (despite it's exaggerated claim to a Darwinian biological analogy) can be a powerful and adaptable tool to solve much more complex problems as it has done in the various fields of operation research, engineering, bio-medical sciences and quantitative finance.

Appendices

	Number of lots	average price per m2	Distance to central station in metres	average lot area in m2	FAR
Cianjhen	21	5.58	5688	60.069	3.510
Cianjim	23	8.46	1642	222.594	4.983
Cijin	26	1.98	6172	67.846	2.892
Gushan	46	3.57	2432	90.357	4.033
Lingyo	88	7.38	2294	81.044	4.633
Nanzih	26	3.70	9536	467.868	2.892
Samnim	48	6.77	2242	462.388	4.285
Siaogang	90	2.40	11636	126.621	2.500
Sinsing	16	7.57	1339	100.506	4.794
Yangchang	32	5.37	2545	57.583	4.641
Zuoying	53	4.79	4880	494.948	4.426
Grand Total	469	5.03	5108.16	202.28	3.89

	Use(0=residential 1=shop & house use 2=parking lot 3=warehouse 4=retail shop 5=factory 6=office)					
Count of usage	usage					
name	1	2	3	4	5	Grand Total
Cianjhen	4			4	13	21
Cianjim	11		1	7	4	23
Cijin	1			1	24	26
Gushan	5	4	1	13	23	46
Lingyo	8	6		65	9	88
Nanzih	2	6		5	13	26
Samnim	3	8	1	17	19	48
Siaogang	12			65	13	90
Sinsing	1	2		4	9	16
Yangchang	1			21	10	32
Zuoying	3	34		1	15	53
Grand Total	51	60	3	203	152	469

Correlation matrix

Correlation matrix (Pearson (n)):

Variables	distance from station	zoning coded	Road width	near road	usage	FAR	$/m2
distance from station	**1**	0.391	-0.189	-0.099	0.018	-0.421	-0.412
zoning coded	0.391	**1**	-0.160	0.149	-0.002	-0.577	-0.515
Road width	-0.189	-0.160	**1**	-0.010	-0.246	0.430	0.256
near road	-0.099	0.149	-0.010	**1**	0.115	0.002	-0.147
usage	0.018	-0.002	-0.246	0.115	**1**	-0.166	-0.095
FAR	-0.421	-0.577	0.430	0.002	-0.166	**1**	0.479
$/m2	-0.412	-0.515	0.256	-0.147	-0.095	0.479	**1**

Correlation between land area an prices (is missing from the previous table)

	area	$/m2
area	1	
$/m2	-0.03544	1

[i] References bvbv.

[ii] Lit…..