Forecasting the Case/Shiller Index for the Los Angeles Metro Area

by Mak Kaboudan

July 16, 2009
15th International Conference on Computing in Economics & Finance
The S&P Case/Shiller Index

- S&P CSI is of median home prices representing each of 20 major metropolitan statistical areas within the US and a composite of all 20. The index is published monthly.
- The Los Angeles Metropolitan Statistical Area (MSA) is one of the 20 for which the index is computed. The index for LA (LXXR) is the focus of this presentation.
- The index is published monthly.
Historical trends of the Los Angeles MSA S&P/Case–Shiller index (LXXR) plot against the 10–MSA composite index (CSXR) and the 20–MSA composite index (SPCS20R).
The objective of this research is to predict LXXR for 24 months (April of 2008 through March of 2010).

Data used is monthly and covers the period from January of 1992 through March of 2008.
The forecast is developed in stages employing agents (hence, a multi-agent modeling strategy):
1. One agent identifies those variables suspected of creating temporal variations in LXXR.
2. Several agents then forecast those variables.
3. More agents finally model and forecast the monthly percent change in LXXR (%D_LXXR).
The Focus Variable: %D_LXXR
The Agents

- A GENETIC PROGRAMMING SEARCH AGENT
- A NEURAL NETWORK TRAINING PARADIGM
- LINEAR REGRESSION MODELS
GENETIC PROGRAMMING

GP is a computationally intensive search technique designed to optimize a specified function. It is used here to obtain “nonlinear-regression” type models by minimizing their estimation mean square error (MSE).

TSGP (for Time Series Genetic Programming) software is used to obtain randomly assembled models that predict %D_LXXR. The program assembles a large number of equations (1000). It then uses them as a population that can breed new equations. Breeding takes place using self-reproduction, cross-over, and mutation.
GP ARCHITECTURE

Assemble initial population

Assemble a new population

Compute fitness values (MSE)

MSE < 0.001

MSE > 0.001

Terminate run

Generations evolved < maximum specified

Generations evolved = maximum specified
Example of an assembled equation

The parse tree represents the equation:

\[ Y = 4 \times X + \sin Z \]
CROSSOVER Breeding of 2 Individuals

**Individual 1:** \(+ \ln X4 \ln X7\) or \(Y = \ln X4 + \ln X7\)

**Individual 2:** \(- (\sqrt{\left( + X1 \left( * 8 X2 \right) \right)} X3\) or \(Y = \sqrt{X1 + 8 \ast X2} - X3\)
MUTATION
Breeding of one Individual

Genes assembled to mutate
Neural networks are an information-processing paradigm based on the way the densely interconnected parallel structure of the human brain processes information. A network takes explanatory variables as input to learn how to produce their closest fitted values of a dependent variable as output. Networks are typically trained with static backpropagation and require differentiable, continuous nonlinear activation functions such as hyperbolic tangent or sigmoid.
The Multi-Agent System

Stage 1: Selection of X variables

Stage 1: Explanatory variables identification

\( X_{v1} \)  \[ \cdots \]  \( X_{vn} \)

\[ \downarrow \]

GP

\[ \downarrow \]

\( X_{v1} \)  \[ \cdots \]  \( X_{vn} \)
## Explanatory Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COL</strong></td>
<td>Cash out loans or percent of amount borrowed above the amount used to finance a purchase</td>
<td><a href="http://www.freddiemac.com/finance/cmhpi/#old">www.freddiemac.com/finance/cmhpi/#old</a></td>
</tr>
<tr>
<td><strong>FS</strong></td>
<td>Number of houses for sale in thousands</td>
<td><a href="http://www.census.gov/hhes/www/histt10.html">www.census.gov/hhes/www/histt10.html</a></td>
</tr>
<tr>
<td><strong>SOLD</strong></td>
<td>Number of units sold in thousands</td>
<td><a href="http://www.census.gov/hhes/www/histt10.html">www.census.gov/hhes/www/histt10.html</a></td>
</tr>
<tr>
<td><strong>ES</strong></td>
<td>Excess supply</td>
<td>[ ES_t = FS_{t-1} - SOLD_t ]</td>
</tr>
<tr>
<td><strong>CMI</strong></td>
<td>Construction material index</td>
<td><a href="http://www.census.gov/const/www/newresconstindex_excel.html">www.census.gov/const/www/newresconstindex_excel.html</a></td>
</tr>
<tr>
<td><strong>CHI</strong></td>
<td>Change in housing inventory</td>
<td>[ CHI_t = FS_t - FS_{t-1} ]</td>
</tr>
<tr>
<td><strong>MR</strong></td>
<td>30-year real mortgage rate</td>
<td><a href="http://www.freddiemac.com/dlink/html/PMMS/display/PMMSOutputYr.jsp">www.freddiemac.com/dlink/html/PMMS/display/PMMSOutputYr.jsp</a></td>
</tr>
<tr>
<td><strong>LOAN</strong></td>
<td>Indexed loan</td>
<td>[ LOAN_t = LXXR * LPR ]</td>
</tr>
<tr>
<td><strong>LAPI</strong></td>
<td>Los Angeles real personal income</td>
<td><a href="http://www.bea.gov/regional/sqpi/default.cfm?sqtable=SQ1">www.bea.gov/regional/sqpi/default.cfm?sqtable=SQ1</a></td>
</tr>
<tr>
<td><strong>ESDV</strong></td>
<td>Excess supply dummy variable.</td>
<td>[ ESDV_t = 1 \text{ if } CHI_t &lt; \text{average } CHI_t, \text{ and } = 0 \text{ otherwise.} ]</td>
</tr>
</tbody>
</table>
Predication the Explanatory Variables

Stage 2: Estimation of X variables to obtain their forecasts

Flow chart starting with input variables \( z_{vl} \) for \( v = 1, \ldots, V \) explanatory variables and \( l = 1, \ldots, L \) lags) that explain variations in \( X_1 \). Identical inputs are used by the two agents (GP and ANN) to produce solutions \( GP \, S_1 \) and \( ANN \, S_1 \). Competitive and/or cooperative strategy then determines the final output as a forecast of \( X_1 \). The selected forecasts are then used as input to predict \%D_LXXR in the final stage.
Explanatory Variables’ Estimation

\[
\text{COL}_t = f(\text{MR}_{t-6}, \ldots, \text{MR}_{t-18}); \quad (1)
\]

\[
\%D_{\text{MR}}_t = f(\%D_{\text{MR}}_{t-6}, \ldots, \%D_{\text{MR}}_{t-18}); \quad (2)
\]

\[
\text{FS}_t = f(\text{FS}_{t-1,12}, \text{COL}_{t-1,12}, \text{LPR}_{t-1,12}); \quad (3)
\]

\[
\%D_{\text{SOLD}}_t = f(\%D_{\text{MR}}_{t-6,12}, \%D_{\text{LAPI}}_{t-1,12}); \quad (4)
\]

\[
\text{CMI}_t = f(\text{CMI}_{t-3,12}); \quad (5)
\]

\[
\%D_{\text{LOAN}}_t = f(\%D_{\text{MR}}_{t-3,18}); \quad (6)
\]

\[
\%D_{\text{LAPI}}_t = f(\%D_{\text{LAPI}}_{t-3,6}, \%D_{\text{CPI}}_{t-12,16}, \%D_{\text{HWR}}_{t-12,16}). \quad (7)
\]
### Estimation statistics of fitted explanatory variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>MSE</th>
<th>$R^2$</th>
<th>Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>%D_COL</td>
<td>7.39</td>
<td>0.94</td>
<td>ANN</td>
</tr>
<tr>
<td>%D_MR</td>
<td>6.10</td>
<td>0.26</td>
<td>GP</td>
</tr>
<tr>
<td>FS</td>
<td>2.98</td>
<td>0.61</td>
<td>ANN</td>
</tr>
<tr>
<td>%D_SOLD</td>
<td>18.81</td>
<td>0.50</td>
<td>ANN</td>
</tr>
<tr>
<td>CMI</td>
<td>9.84</td>
<td>0.96</td>
<td>GP</td>
</tr>
<tr>
<td>%D_LOAN</td>
<td>1.31</td>
<td>0.38</td>
<td>ANN</td>
</tr>
<tr>
<td>%D_LAPI</td>
<td>0.06</td>
<td>0.71</td>
<td>ANN</td>
</tr>
</tbody>
</table>
Flow chart starting with $X_j$ explanatory input variables into each modeling technique. Solutions from the different models are identified by ‘S’ (GP S, ANN S, ..., etc.). Respective MSE computations determine the best model each technique produces. ANN is then used to estimate models that fit residuals output from GP and from RM.
GP’S BEST EQUATION

\[ Y = \cos\{\sin[X_{1t-3} \cdot \{\cos(X_{1t-6}) + X_{1t-3}\}] - X_{1t-6}\} \ast \sin[\sin(X_{2t-5}) + \{X_{3t-9} \cdot \sin(X_{2t-5} + X_{1t-3})\}] + \{X_{3t-9} \ast \{\sin(X_{2t-5}) + DV_t + X_{3t-9}\} + \{\sin(X_{2t-5} + X_{1t-3} \cdot \{\cos(X_{1t-6}) + X_{1t-3}\})\} - \sin\{X_{1t-3} \ast \{\cos[\sin(X_{2t-5}) + DV_t] + X_{1t-3}\}\} + X_{1t-6}\}\] - DV_t - 2 \ast X_{4t} \quad (8)

\[ Y = \%D_LXXR, \; X_1 = CHI, \; X_2 = \%D_SOLD, \; X_3 = COL, \; X_4 = \%D_MR, \; \text{and} \; ESDV = DV.\]
\[ Y = 7.49 - 39.41 \times 4_{t-6} - 0.74 \times 3_{t-5} + 0.55 \times 2_{t-11} \\
(0.75) (14.3) (0.26) (0.18) \\
+ 3.40 \ln(5_{t-9}) - 3.87 \times 5_{t-12} - 1.47 \times 1_{t-1} + 0.42 \times 1_{t-6}. \\
(0.93) (0.77) (0.17) (0.22) \]

\( Y = \%D\_LXXR, \ X1 = CHI, \ X2 = \%D\_SOLD, \ X3 = COL, \ X4 = \%D\_MR, \) and \( ESDV = DV. \)
- Residuals produced by GP are fit using ANN. The same exact explanatory variables are used in obtaining the new models.
- Residuals produced by ANN are fit using GP.
- The idea is that whatever one technique missed (the residuals) can be explained using another (ANN & GP).
- The new outcomes are GP2 & ANN2.
Estimation Statistics

Estimation statistics of the five agents and their averages

<table>
<thead>
<tr>
<th></th>
<th>GP 1</th>
<th>GP 2</th>
<th>OLS</th>
<th>ANN 1</th>
<th>ANN 2</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.008</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>R²</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.994</td>
<td>0.995</td>
<td>0.997</td>
</tr>
</tbody>
</table>
• The five fit and forecast outcomes are then used as input to produce three new forecasts. This is similar to taking an average of all except that GP, ANN, & OLS agents produce their own weights to reconcile the produced forecasts.

• The new outcomes are GP_R, ANN_R, & OLS_R.

• Once these are determined to produce superior fitting results, they are averaged to determine if an average improves the forecast further.
Reconciliation Equations

- Using GP:
  \[ %D_{LXXR} = 0.4194 \text{ GPNN} + 0.5806 \text{ NN} \]

- Using OLS:
  \[ %D_{LXXR} = 0.011 + 0.463 \text{ GPNN} + 0.543 \text{ NN} \]
  \[ (0.006) \quad (0.053) \quad (0.052) \]

Estimation statistics of the agents and their average:

<table>
<thead>
<tr>
<th></th>
<th>GP_R</th>
<th>ANN_R</th>
<th>OLS_R</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.005</td>
<td>0.008</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.997</td>
<td>0.994</td>
<td>0.997</td>
<td>0.997</td>
</tr>
</tbody>
</table>
Ex post Prediction from Best Forecasting Paradigm
(MSE = 33.83)
Best Forecast

- %D_LXXR
- AveR
- Ex post
- Ex ante

Date: 10/18/93 to 3/23/10
Best Forecast
Conclusion

• The forecasts obtained by the employed agents were marginally different and can be viewed as too similar to select one as best.

• The average forecasts obtained using cooperation between the different agents employed produced the least mean square error. Actual index values published after the forecasts were completed (for April through November 2008) suggest that the average of the five forecasts were more consistent with reality.

• More importantly, given the strong similarity between the forecasts obtained, it is easy to conclude that the LA metropolitan area housing market will remain depressed until the end of the forecast period considered in this research. Prices will stabilize in 2009 but resume their decline early in 2010.
CSI Foundation

- The index, which is designed to measure changes in the market value of residential real estate in each Metropolitan Statistical Area (MSA), tracks the values of single-family housing within the United States. It measures changes in housing market prices with homes sold held at a constant level of quality by utilizing data of matched sale pairs for pre-existing homes. For each MSA, a three-month moving average is calculated. The monthly moving average is of sales pairs found for that month and the preceding two months. Each pair is then allocated to one of three price levels (low, middle and high) determined according to the position of the first price of the pair among prices in the period of the first sale.