Forecasting House Prices in Ireland

Grossi, M. Y. Xuan

December 17, 2015

Abstract

Widely attributed as the cause of the global market crisis of 2007, the house pricing bubble is an important topic of research. Economists try to explain what causes changes in house prices to understand and create policies that will affect this market ((Unknown, 2010) and (Roche, 2003)).

This project aims to forecast the prices of new houses in Ireland based on available statistical data sets that from a fundamental economic standpoint ((Égert and Mihaljek, 2007)) might influence the housing market ((Auterson, 2014)). Four fundamentally sensitive data sets are chosen of an initial group to serve as predictors. An ARIMA model is then fitted to our data and a forecast of the prices of new houses in Ireland is produced with satisfactory results.

1 Introduction

The housing market is one of the fundamental cornerstones of an economy. The ability to forecast it by using fundamental economic theory is always a hot topic for research. Given the nature of the house pricing market a special type of statistical analysis is necessary. Time series analysis differs from other statistical analysis due to the fact that it needs to account for the fact that data points taken over time may have an internal structure (such as trend, autocorrelation, or seasonal variation).

In this report, an effort is made to forecast the new house prices in Ireland for the periods of 2006 to 2011 by using historical data from 1975 to 2005 and also a set of explanatory variables from the same period. We also show comparatively the forecast data against the real data for visually assessment of the quality of the prediction.

This forecast will be based on time series analysis using the R software platform and various packages. We will use the $ARIMA$ autoregressive model to fit our historical data plus explanatory data and predict future points. These model is very powerful in the sense that it can be applied to non-stationary time series by means of a differencing step. It is important to note that when explanatory variables are added to an autoregressive model this is also known as a dynamic autoregressive model(Hyndman and Khandakar, 2008).

2 The Data

In order to ensure the accuracy and the objectivity of our statistic analysis, we found some authoritative and reliable databases as the sources for data such as the Central Statistics Office[1], which is Ireland's national statistical office, responsible for coordinating the official statistics of other public authorities and for developing the statistical potential of administrative records and The World Bank Open Data[2], which is an open database about country development from all around the globe. Considering what elements might impact the house price market, we choose some datasets that are commonly referred by economic papers on the subject ((Auterson, 2014) and (Égert and Mihaljek, 2007)).

2.1 Population

The census population as measured by the central statistics office of Ireland[3]. Normally, the amount of houses would not be bigger than the population, other words, the population can influence the number of houses needed indirectly.

The original dataset only had the population for each 5 year interval from 1841 to 2011 (Census years). For analysing time series we need to have data with matching frequencies, therefore we used linear interpolation to fill in the gaps. We will be calling this data set henceforth as $cna13$.

2.2 Gross Domestic Product

Gross Domestic Product (GDP) is a measure of a nation's total economic activity. It represents the monetary value of all goods and services produced within a nation over a specified period of time (usually per annum).

This data was obtained from the World Development Indicator dataset[4]. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.

The GDP influence the housing in two basic ways: through private residential investment and consumption spending on housing services. The mentality is very naive, the more money that is available in an economy, the more people will spend in housing what would drive the prices up.

At first we wanted to use estimated household disposable income[5] as this indicator is more precise regarding how much money people have at-hand. However there was no significant historical available for this (data exists only from 2000 onwards). So finally *gdp* (in Euro) becomes our choice.

[1] CSO - http://www.cso.ie

[2] http://data.worldbank.org

[3] Table CNA13 on CSO

[4] The World Bank Data in Ireland

[5] Table CIA01 on CSO

2.3 House Completions

House completions dataset[6] is based on the number of new dwellings connected to ESB[7] electricity network. These represent the number of homes completed and available, and do not reflect any work-in progress. ESB have indicated that there was a higher backlog in work-in-progress in 2005 than usual (estimated as being in the region of 5,200 units). This backlog was cleared through the connection of an additional 2,000 houses in Quarter 1 2006 and 3,000 houses in Quarter 2 2006. In addition, second-hand houses are not included. This dataset will be referred by as $hsa01$.

The amount of new houses is a important part in the house price prediction model(Auterson, 2014). It is very intuitive that house completions would be a good indicator of the supply of new houses.

2.4 Value of House Loans Approved

This measure is the total value of approved house loans[8], which is based on the existing house mortgages in Ireland from 1970 to 2011. This only reflects loans approved for purchases of new new houses. It is worth mentioning that dataset contains an unquantified element of refinancing of existing mortgages, involving the redemption of an existing mortgage and its replacement with a mortgage from a different lender.

The amount of money approved for buying new houses is a very good intuitive measure of demand for new houses. This will be called $hsa08$ throughout this study.

3 Study: A Time Series Analysis

Time series analysis is a very large research topic given its large usability in a multitude of environments and several methods for analysing such data exists ((Cowpertwait, 2009) and (Coghlan, 2015)), which gives a researcher an ample choice of approaches to work with. That is why it is of paramount importance that an appropriate model is chosen.

3.1 Model choice

The choice of a model depends entirely on the characteristics of the data we are trying to forecast and whether the data appears to be random or not ((Kumar and Anand, 2014)). If a time series is random then it is expected that the correlations between its successive values is close to zero. If it is discovered that there exists a correlation between the consecutive observations of a time series than an autoregressive model is likely to be a good choice as it has the ability to cope with such dependencies.

[6] Table HSA01 on CSO

[7] A licensed operator of electricity distribution in Ireland.

[8] Table HSA08 on CSO

For this project we chose the autoregressive integrated moving average or $ARIMA(p,d,q)$ model for its ability to take into consideration not only the autoregressive nature of the data, but also can cope well with non-stationary, seasonal time series. Given the available data for our project is yearly, we will not be evaluating any seasonal component and it will not influence the results as the model is built for both seasonal and non-seasonal time series. The full equation of the $ARIMA(p,d,q)$ without any explanatory variables.

$$ARIMA(p,d,q) = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_q e_{t-q}$$

According to (Hyndman, 2014) we can introduce the explanatory variables by letting the error e be autocorrelated. This can be expressed by the backshift notation ($B^d y = y_{t-d}$) example $ARIMA(1,1,1)$ equation below, where $x_{1,t}$ to $x_{k,t}$ are our explanatory variables ($cna13$, $hsa01$, $hsa08$ and gdp).

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + n_t,$$

$$(1 - \phi_1 B)(1 - B)n_t = (1 + \theta_1 B)e_t,$$

where e_t is white noise.

3.2 Model Parameters

The p, d and q parameters must be chosen for the $ARIMA(p,d,q)$ model. Parameter p is the coefficient of the auto regressive part, d is the integration order (or the number of differences that make our time series stationary) and q is the coefficient for the moving average. In this project we use the (Hyndman and Khandakar, 2008) algorithm that traverses the model space and tries to minimize the corrected Akaikes Information Criterion (AIC_c) value in a two-step approach.

First step is to initialize the model with $ARIMA(p,d,q)$ and choose the one with minimum AIC_c, where d is determined by unit root tests and $(p,q) \in [(2,2),(0,0),(1,0),(0,1)]$. The second step will vary the chosen p and q by ±1 until no smaller AIC_c value can be found. For this project the optimal model was found to be $ARIMA(3,0,0)$[9] and the coefficients can be seen in Table 1.

	AR^1	AR^2	AR^3	Intercept	cna13	hsa01	hsa08	gdp
Coeff.	0.8548	-0.0684	-0.4110	-0.1414	0.8269	0.1994	0.0732	0.5243
s.e.	0.1754	0.2365	0.1655	0.0302	0.1960	0.0402	0.0306	0.0712
est.σ^2=4.613e-05, log likelihood=109.85, $AIC = -201.7, AIC_c = -193.12, BIC = -188.79$								

Table 1: Table of $ARIMA(3,0,0)$ coefficients with non-zero mean

3.3 Model fitting and forecasting

After determining the optimal model parameters we must fit the model from our data. For this project we have split our data set into two; one for training

[9] The model is equivalent to an $AR(3)$ model.

and another for testing our forecast. Considering that the available data period for the Irish prices of new houses varies from 1975 to 2013 we have chosen 1975 to 2005 as the training set and from 2006 to 2011[10] as the testing data set.

The fitted model is now used to forecast the future values of the Irish new house prices from 2006 to 2011. Table 2 shows the forecast raw values and Figure 1 the graphical comparison of the real values versus the forecast values.

Year	Forecast	Real	Lo 80	Lo 95	Hi 80	Hi 95
2006	304,660.80	305,637.00	301,967.70	300,542.20	307,353.80	308,779.40
2007	303,497.10	322,634.00	299,954.30	298,078.90	307,039.80	308,915.20
2008	273,847.30	305,269.00	269,881.10	267,781.50	277,813.50	279,913.10
2009	237,711.60	242,033.00	233,736.90	231,632.80	241,686.40	243,790.50
2010	228,514.20	228,268.00	224,450.40	222,299.20	232,578.00	234,729.20
2011	233,964.90	230,303.00	229,642.20	227,353.80	238,287.70	240,576.10

Table 2: Table of scaled prices of new Irish houses

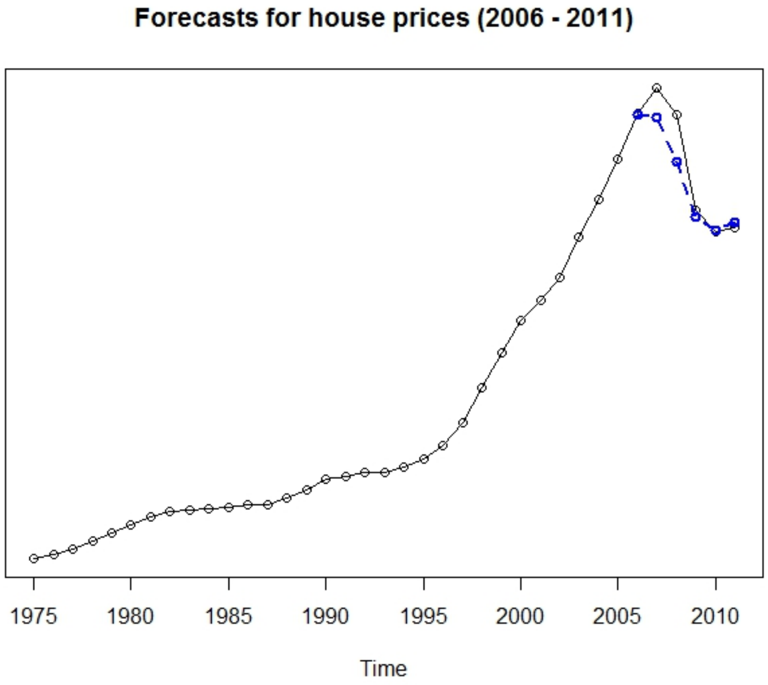

Figure 1: Real prices for new Irish houses versus forecast prices.

[10] As the predictor data sets only go up to 2011 we must limit this data set as well.

The visual comparison of the forecast values versus the real values is very satisfying but we must also analyse the forecasts residuals. It is desired that the residual errors of a forecast be normally distributed with zero mean and constant variance and have no correlation between successive forecast errors. This is to ensure that the residuals are just noise and do not hide any unexplained variables which would indicate a poor forecast model.

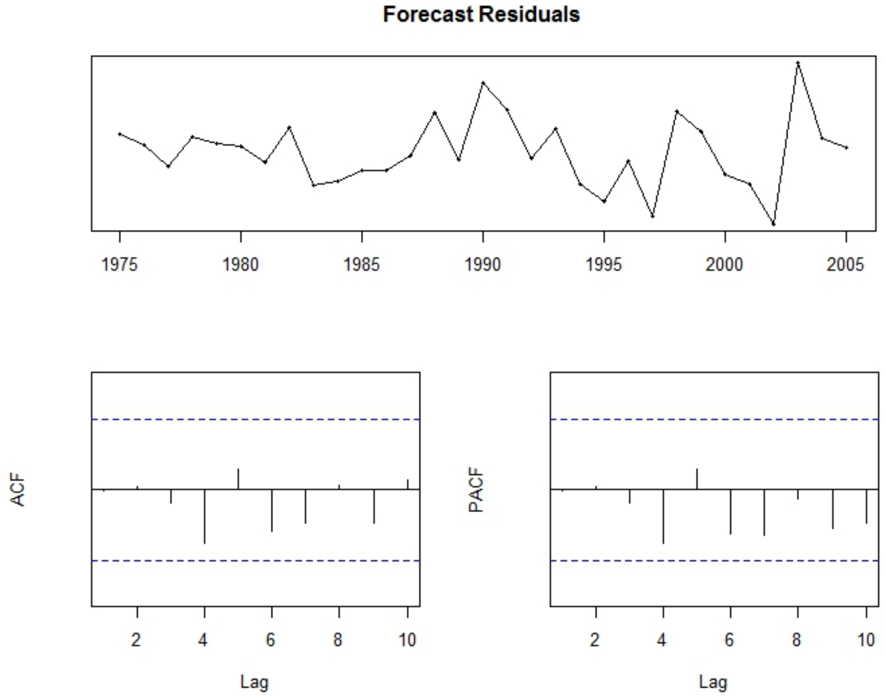

Figure 2: Forecast residuals errors, autocorrelation and partial autocorrelation.

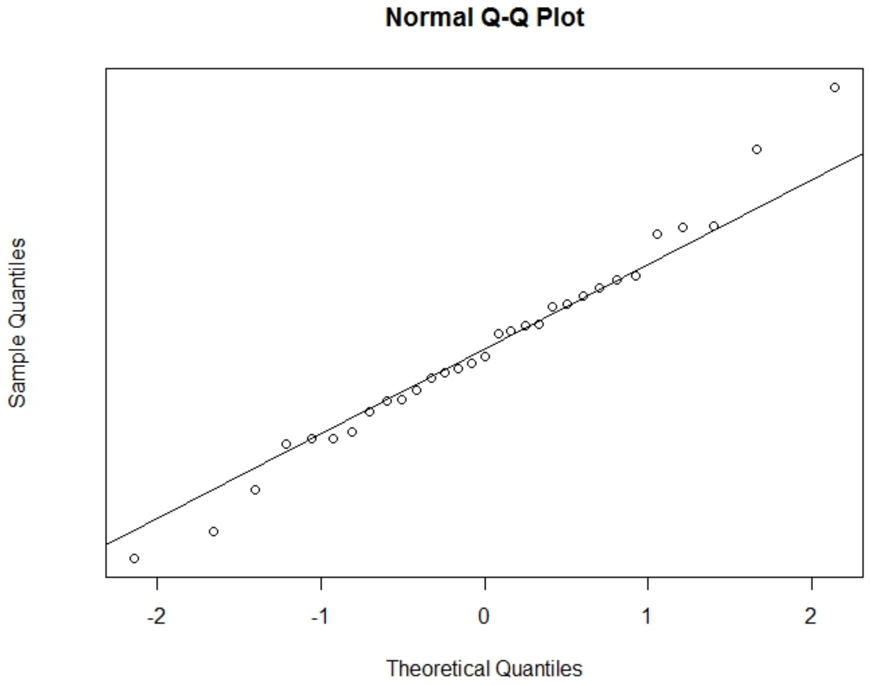

Figure 3: Normal Q-Q plot of forecast residuals.

Analysing Figure 2 we can see that there are no correlated lagged errors and Figure 3 shows a compatible normal distribution and finally we run a Box-Ljung test that fails to reject our null hypothesis that there are no autocorrelations in the forecast errors at the first 10 lags (Table 3).

χ^2	df	p-value
7.6182	10	0.6661

Table 3: Table of Box-Ljung test results for H_o: No autocorrelations

4 Conclusions

This project showed how to apply an autoregressive $ARIMA$ model to a time series using four fundamentally reasonable external predictors to forecast the prices of new houses in Ireland. It is shown that a model of $ARIMA(3,0,0)$ was a good fit for our data and the forecast residuals showed no autocorrelation and that they seem to follow a normal distribution with mean zero and constant variance.

The forecast value of the house prices was then compared to their real value and the result was very satisfactory. Considering that the period for which the forecast was ran coincided with the property market bubble burst of 2006 to 2008 and the model was able to pick up these abrupt changes makes the results more gratifying.

It would be beneficial for future work to repeat this study with monthly data, instead of yearly (provided this data is available). This would allow exploring the seasonality of the data, which from an economically fundamental point of view would make sense given that house prices may be impacted by holiday season and perhaps end of year, when people usually have more expenditures and are less likely to be looking to buy.

Another avenue of research would be to compare several time series analysis methods on the same data set. For our example we could improve upon this work by trying VAR and $GARCH$ models for example.

References

Unknown, "Irish housing market - fundamentally strong or a speculative bubble," http://misc.mortgagebrokers.ie/images/blogimages/2010/August2010/IrishHousingMarket-fundamentallystrongoraspeculativebubble.pdf.pdf, 2010, (Visited on 12/13/2015).

M. Roche, "Will there be a Crash in Irish House Prices?" *Quarterly Economic Commentary: Special Articles*, vol. 2003, no. 4-Winter, pp. 1–16, 2003. [Online]. Available: https://ideas.repec.org/a/esr/qecsas/2003winterroche.html

B. Égert and D. Mihaljek, "Determinants of house prices in central and eastern europe," *Comp Econ Stud*, vol. 49, no. 3, pp. 367–388, sep 2007. [Online]. Available: http://dx.doi.org/10.1057/palgrave.ces.8100221

T. Auterson, *Working paper No 6 Forecasting house prices.* London: Office for Budget Responsibility, 2014.

R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for r," *Journal of Statistical Software*, vol. 27, no. 3, 2008. [Online]. Available: http://dx.doi.org/10.18637/jss.v027.i03

P. Cowpertwait, *Introductory time series with R.* New York: Springer-Verlag, 2009.

A. Coghlan, "A little book of r for time series," https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf, 11 2015, (Visited on 12/15/2015).

M. Kumar and M. Anand, "An application of time series arima forecasting model for predicting sugarcane production in india," *Studies in Business and Economics*, vol. 9, no. 1, pp. 81–94, 2014.

R. Hyndman, *Forecasting : principles and practice.* Heathmont, Vic: OTexts, 2014.