Forecasting Residential Dwelling Property Transaction Value in Ireland

Colm Clery

Dataset

This dataset is publicly available from the Central Statistics Office with code HPM02: [HPM02 - Residential Dwelling Property Transactions (cso.ie)]

The dataset contains 9 columns and 144 rows.

```
> colnames( Irish.Household.Sale.Data )

[1] "i..Statistic"     "Month"            "County"

[4] "Dwelling.Status"  "Stamp.Duty.Event" "Type.of.Buyer"

[7] "Type.of.Sale"     "UNIT"             "VALUE"
```

The columns of interest are “Month” and “Value”. Each row corresponds to a single month over the course of 12 years spanning January 2010 - December 2021. “Value” refers to the mean sale value of residential dwelling property transactions closed during each month

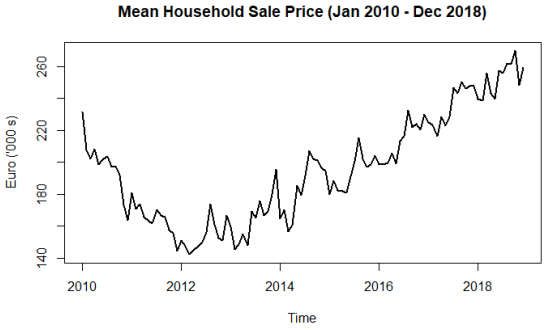
For the purpose of this analysis, the data has been split into training and test sets comprising of data from January 2010- December 2017 and January 2018- December 2021 respectively.

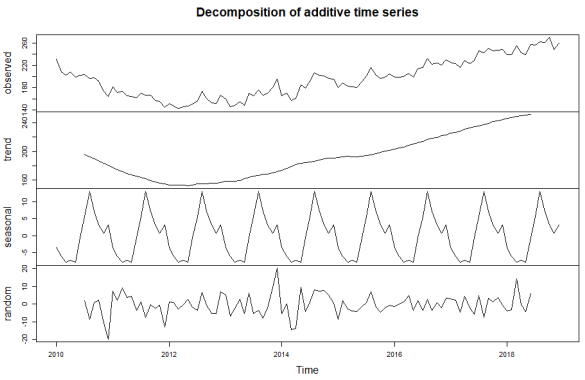
The purpose of this analysis is to quantify and analyse the growing housing crisis in the Republic of Ireland. The questions I seek to answer in this analysis are:

- Is the cost of housing continuing to increase year on year?
- How much would you expect to pay to become a home-owner in the future if trends continue?
- Are there times of the year where residential properties are, on average, significantly cheaper to buy?

Reduction to stationarity

To begin, we visualise the time series and decomposition.





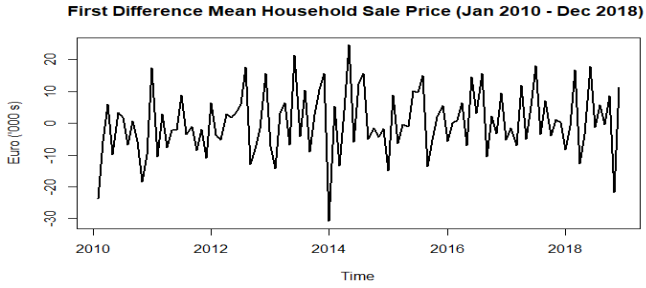
The data appears non-stationary. There also is a large seasonal component which peaks in August and reaches its minimum in May each year. To be sure, we run an augmented Dickey-Fuller test.

H_0: The time series is non-stationarity

H_A: The time series is stationarity.

```
> adf.test(tseries , alternative = "stationary")

         Augmented Dickey-Fuller Test

data:  tseries

Dickey-Fuller = -2.8508, Lag order = 4, p-value = 0.2237

alternative hypothesis: stationary
```

This returns a p value of .2237, which is greater than .05 and insignificant at a 5% level of significance. We accept the null hypothesis that the data is non-stationary. Based on this conclusion, we create an integrated series of order 1.

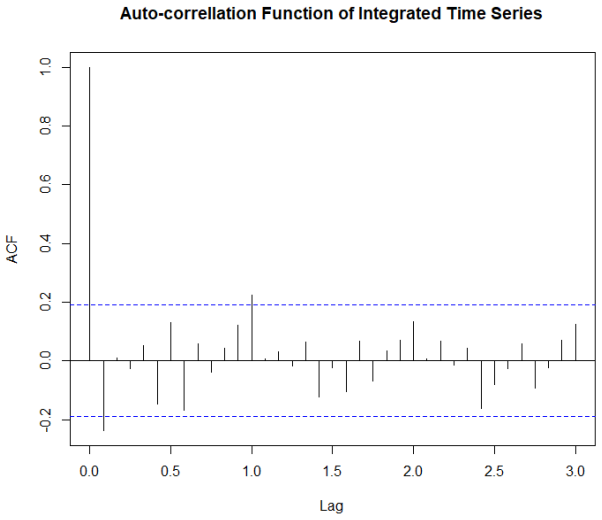


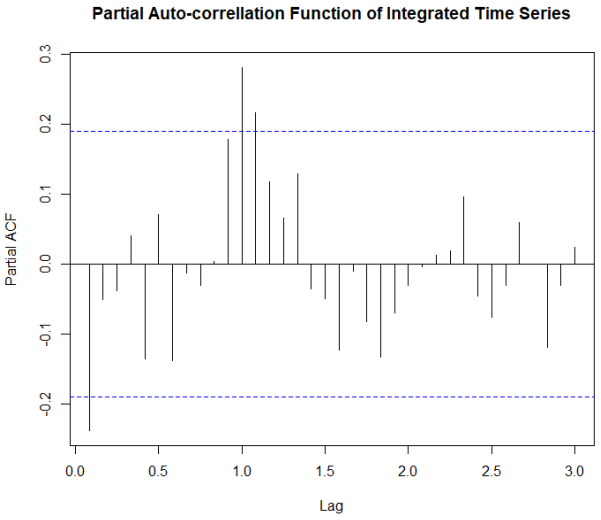

```
 > adf.test(diff(tseries) , alternative = "stationary")$p.value

[1] 0.01
```

Our differenced time series looks stationary and we can reject the Null hypothesis of non-stationarity under the Dickie-Fuller test.

Model Fitting

To fit an ARMA model to our integrated time series, we first begin by generating and investigating the Auto-correllation function (ACF) and Partial Auto-correllation function (PACF).





Consider first the non-seasonal component of the data.

The acf peaks at lag 1.0 (1 year backwards in time) and appears to cut off completely after this. The PACF peaks with significant values at lags 1.0 and the decays in a form which resembles a damped cosine function. This damped cosine function in the PACF values is indicative of a MA(1) and hence conclude that a seasonal MA(1) would be appropriate for this time series.

The non-zero acf terms over lags (1.0, 3.0) which again resemble a damped cosine function. The damped cosine function present in the ACF values indicate that a AR(1) component is appropriate. As a result I also considered seasonal ARMA(1,1),deciding between the two by comparison of AIC values.

```
> t.smodel = arima(tseries, order = c(0,1,0) , seasonal = c(0,0,1))

> AIC(t.smodel)

[1] 794.2374

> t.smodel2 = arima(tseries , order = c(0,1,0), seasonal = c(1,0,1))

> AIC(t.smodel2)

[1] 794.0017
```

The inclusion of seasonal AR(1) reduces the AIC slightly and so I decided on an ARMA(1,1) as my seasonal component.

Now considering the seasonal portion of the data. The acf and pacf functions behave similarly over lags [0,1.0] as they do over lags [1.0, 3.0]. There are peaks in both acf and pacf at lag 1/12 (one month backwards in time) with damped cosine decays in the following lags. Again, I would suggest an MA(1) or ARMA(1,1) as our candidate models.

```
> t.model = arima(tseries , order = c(0,1,1), seasonal = c(1,0,1))

> AIC(t.model)

[1] 777.9979

> t.model.2 = arima(tseries , order = c(1,1,1), seasonal = c(1,0,1))

> AIC(t.model.2)

[1] 779.9772
```

ARIMA(0,1,1)(1,0,1)12 minimises the AIC and so this is my chosen final model.

When deciding on both the seasonal and non-seasonal components, I employed a search of hyperparameters p,q (non-seasonal ARMA) and P, Q (seasonal ARMA). Over the range of 0 to 5 inclusive in both cases to minimise the AIC. Included below is an example of the code employed for search over seasonal component.

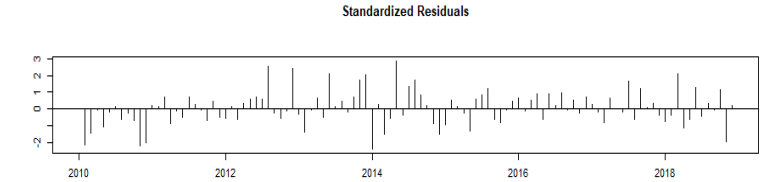
```
ij.min = c(1,1)

min.AIC = AIC(t.smodel)

for (i in c(0,1:5) ){

  for (j in c(0,1:5)){

    t.model.ij = arima(tseries , order = c(0,1,0), seasonal = c(i,0,j))

    AIC.ij = AIC(t.model.ij)


    if(AIC.ij < min.AIC){

      min.AIC = AIC.ij

      ij.min = c(i,j)

    }

  }

}
```

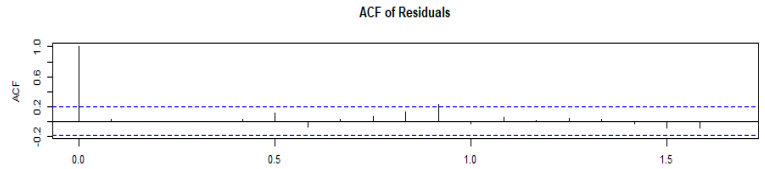
The only improvements that can be made to the AIC under this dataset is by using an ARIMA(0,1,1)(1,0,0)12 model which improves AIC by 0.7024 to 777.2955. The difference is that the seasonal MA part of the model is removed. Including a seasonal AR but not seasonal MA part is not reflective of how I interpret the acf and pacf however, so given the difference in AIC is relatively small and the effect of overfitting should be minimal, I have chosen to include the seasonal MA component. The chosen model is better reflective of the PACF and ACF functions and on this basis, I expect that the forecasts produced by this model will be superior to those of the alternate ARIMA(0,1,1)(1,0,0)12 model which minimises the AIC.

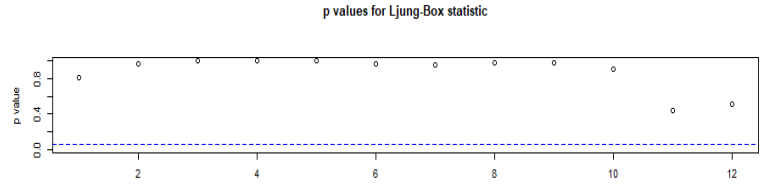
My final chosen model is an ARIMA(0,1,1)(1,0,1)12

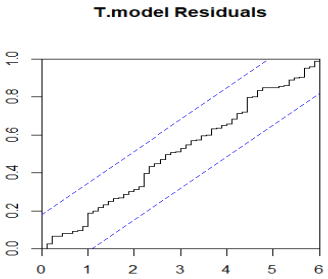
```
Call:

arima(x = tseries, order = c(0, 1, 1), seasonal = c(1, 0, 1))

Coefficients:

          ma1    sar1     sma1

      -0.4716  0.6688  -0.2610

s.e.   0.0925  0.1525   0.2119


sigma^2 estimated as 75.54:  log likelihood = -385,  aic = 778
```

Model criticism









Looking first at the Standardised residuals, we have what appears to be a random scatter. This agrees with the model assumption that errors are independently, normally distributed.

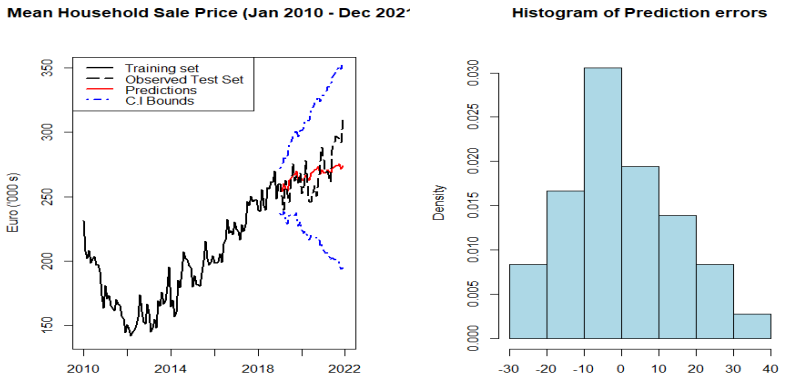
Further, our residuals are mostly uncorrelated, as we can see from our acf of residuals. The only significant autocorrelation is that at lag 11/12 (11 months backwards in time) and most acf values are very close to 0.

Finally, we have p values much larger than the threshold of .05 for all lags in a 1 year period. Thus for all lags in this period, we fail to reject the null hypothesis of the Ljung-Box test and conclude that the residuals are independently distributed. Note that this conclusion includes lag 11/12. Although this lag returns a significant residual ACF, the Ljung-Box test rejects the hypothesis that the lags are dependent and so adjustments to the model to address these ACF values are not necessary.

Finally, considering the cumulative periodogram. Our residual data is relatively smooth about the line y = 2x, which allow us to conclude again that the series of residuals are independently random.

Given this set of Diagnostics, I conclude that the model shows goodness of fit to the training data.

Forecasting



The prediction errors have a mean of 1.16 and a standard deviation of 14.3

The prediction errors have a mean that is close to zero and the distribution of errors is relatively symmetrical. Further, All observed values are within the confidence interval of predictions. This would imply that the model has good prediction power.

The standard deviation of prediction errors is high in comparison to the model estimate for standard deviation, which equals 8.7. This high standard deviation may be a sign of overfitting.

Recall that I chose to include the seasonal MA in this model despite this increasing AIC. Rerunning the prediction procedure without the inclusion of a seasonal MA returns a mean and s.d of prediction errors of 4.4 and 15.6 respectively. First of all this is satisfying as it validates my decision to include a seasonal MA.

Further, it increases the difference in variance of training and test error and so, we can conclude that this is not the source of overfitting. Perhaps it is in the method of fitting itself and that a regularisation method would be appropriate.

Discussion

In conclusion, I believe that the model shows good prediction power and accurately reflects the structural change in the underlying dataset over time. I think that the model suffers slightly from overfitting and that regularisation methods would improve the prediction power of the model further.

Interpreting the model is rather disheartening to those looking to become a homeowner anytime soon. There model captures what is an ongoing and structural increase in Mean Household sale price year on year and the model predicts that this continues to increase. This model predicts values for December 2024 being approximately €278,000, which would be a 96% increase from the value of €142,062 recorded in March 2012.

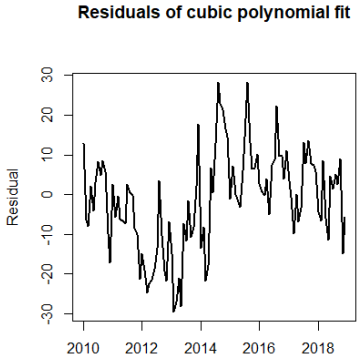
A positive takeaway may be that the seasonal component sites a structural increase in mean value of household sales during the summer holiday months (June July August). Therefore prospective home buyers may be advised to buy during the May where the seasonal component is at a minimum. The minimum being in May is a surprising and interesting result as intuitively I would have expected the minimum to be in the winter months.

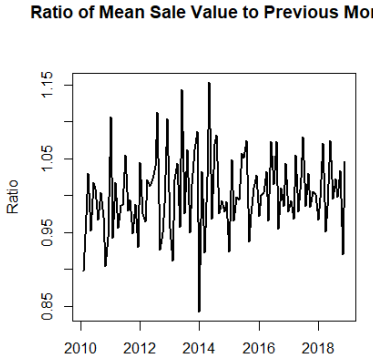
No model is perfect and that includes this one. Improvements that can be made to this model could be as follows:

1. Applying an appropriate transformation to the raw data.

I have integrated the data set to order 1 in fitting this ARIMA model. In doing so, I have made the implicit assumption that there is an underlying linear trend in the data. However, the trend of the time series is not linear and so this assumption is incorrect. The trend is a curve in the approximate time period [2012 , 2015] and outside this region is relatively linear. Also in this period our model returns a number of large standardised residuals exceeding 2 in absolute value. This is unexpected and I believe a result the assumption of linearity being inappropriate during this time.

Transforming the data before differencing or transforming the data so that is it stationary (and differencing is no longer necessary) may improve this model. I considered a polynomial fit as well as transforming the data to be the ratio of Mean Sale value to Mean Sale Value in the previous month.





The residuals of the cubic polynomial are return an insignificant p-value by the Dickey-Fuller test (p = 0.07899). We fail to reject the null hypothesis of non-stationarity and would not consider this an option for model fitting.

The series of ratios however returns a highly significant p-value < .01 by the Dickey Fuller test. We reject the null hypothesis and conclude this data is stationary. An ARMA model fitted to this data may perform better than the model I have fitted to the raw data.

2. Choice of Statistic

One of the questions posed at the beginning of this analysis was "Is the cost of housing continuing to increase year on year? (and how much would you expect to pay to become a home-owner in the future if trends continue?)"

With this question in mind, it may have been more appropriate to model the median value of household sales rather than the mean. The Mean value is susceptible to be influenced by large values, which in this case would be the sale of mansions worth hundreds of millions of euros. Further, it may be influenced by homes that are extremely cheap such as derrelect homes or "Fixer-uppers" which would cost a sizable amount to renovate and add to the true cost of the home. These sales are not necessarily relevant in trying to answer this question. So perhaps the median, which would not be influenced in this way, would be more a more appropriate statistic.