Forecasting Canadian Single-detached Housing Completions Using ARIMA models and Inventional Analysis

Yangming (Derek) Li

[redacted]

Abstract

Canadian housing completions is an important economic indicator. In this article, we are going to use the Box–Jenkins method to identify the ARIMA model. Two models will be used to fit the Canadian housing completions data from 1999 Quarter3 to 2019 Quarter3. After comparing the two models, we use our best model, which includes the inventional analysis to predict the next two years housing completions. We are expecting to see a decreasing trend in the future.

Keywords: ARIMA model, Box–Jenkins method, Dicky-Fuller test,Inventional Analysis

1. Introduction

Housing is a key sector of Canada's economy, which has a dramatic effect on related industries, such as banking, the mortgage sector, raw materials, employment, construction, manufacturing, and real estate. In a strong economy, Canadian are more likely to construct new homes; conversely, in a weak economy, Canadian are less likely to build new homes. Housing completions refer to the number of new residential construction projects that have finished during any particular years. The number of housing completions in Canada is an important indicator of economic strength. There are several different types of the housing unit in Canada: single-detached, multiples, semi-detached, row, apartment and other unit types. This article analyzes the single-detached housing type. We obtained the quarterly single-detached housing completions data from 1999 Quarter 3 to 2019 Quarter3 from the Statistics Canada's database. The data provider is Canada Mortgage and Housing

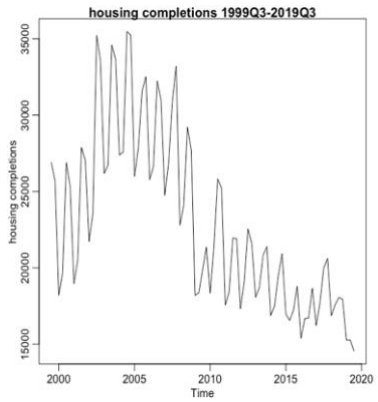
Corporation, which is the crown corporation of the government of Canada. In the beginning , we want to obtain some general information about the data. The mean of data is 22969.53; the median is 21386. The maximum value of the data is 35481, which appeared in Quarter3 of 2004. The minimum value of the data is 14540, which appeared in Quarter 3 of 2019. Since Q3 2019 is the last data point we find in the data. It indicates that it may have a downward trend. The rest of the paper is organized as follows. In Section 2, we will fit the ARIMA model using Box-Jenkins Approach. In Section 3, we will add Intervention Analysis to our ARIMA models. In section 4, we will compare two models described in section 2 and 3. In section 5, we will forecast two years housing completions use our best models. Conclusions are in Section 6.

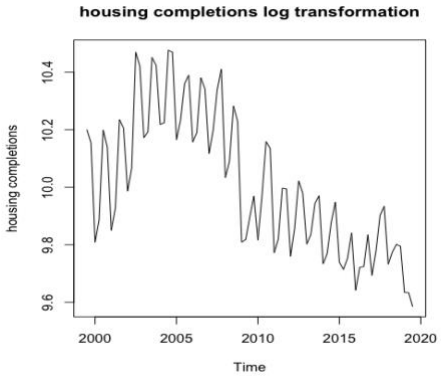
2. Fitting The ARIMA Model Using Box-Jenkins Approach.

We are using the autoregressive integrated moving average (ARIMA) model to analyze the time series described in the paper written by Box and Jenkins (1970). Moreover, the Box–Jenkins method was used to identify ARIMA models. There are three steps to implement the Box–Jenkins method. The first step is data preparation step, which includes plotting the data to see the possible pattern, transforming data to stabilize variance, differencing data to obtain the stationary time series. The second step is called model selection step, which includes examining the ACF and PACF to identify potential models. The last step is estimation and diagnostics, which includes estimating parameters in potential models, selecting the best model by using suitable criterion, checking ACF and normality of residuals, performing Ljung-Box test, and measuring significance level of parameters. The model is adequate if the model satisfies all the criterions.

2.1 Data Preparation

We firstly plot the original time series data, which is shown in figure 1. As can be seen from the graph, the number of housing completions raised steadily from 1999-2005, then it decreased dramatically in the period 2007-2009. It declined gradually after 2009. The time series is not stationary. It does not have constant mean and constant variance. Transformations may be useful to equalize the variability over the time series. We perform a useful log transformation to the data. We let $y_t = \log x_t$.We want to suppress larger fluctuations that occur over portions of the series. After taking the log shown in the figure2, we can see the variance become more stable, but it is still not stationary. For this reason, we perform the difference transformation to the logged series $dy_t = \nabla \log x_t$, then take the seasonal difference to the differenced series $ddlx = \nabla_{12}\nabla \log x_t$. As shown in figure3, we find our time series become more stationary after each transformation. The mean and variance look more stable, except for the time around 2009.





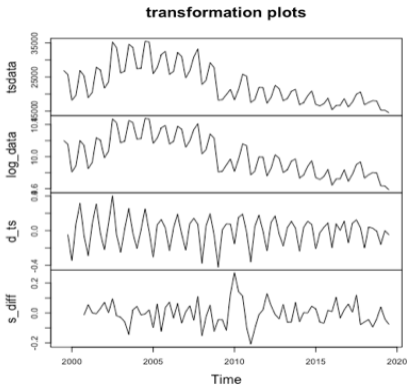


Fig.1 original time series plot **Fig.2 time series plot with log transformation** **Fig.3 plot transformed data($\log x_t$, $\nabla \log x_t$, $\nabla_{12}\nabla \log x_t$)**

For each transformation, we use the Dicky-Fuller test to test the stationary of the time series. The null hypothesis of the Dicky-Fuller test is that the series has a unit root. The alternate hypothesis of Dicky-Fuller test is that the process is stationary. From the table1, We can also see the p-value deceased after each transformation. The p-value of the seasonal differenced data is less than 0.01. It

indicate that we had vary strong evidence against the null hypothesis and the series is stationary. We get a stationary time series after transformation.

Table1

Data	original series	Data after log transformation	Differenced logged data	Seasonal differenced data
Pvalue of ADF test	0.153	0.0989	0.0169	<0.01

2.2 Model Selection

We use the sample ACF and sample PACF to select the models. We plot the sample ACF and sample PACF graph in figure 4 and figure 5. For the seasonal component: It appears that at the seasons, the ACF is cutting off at lag 1s,(s=4). PACF is tailing off at lags 1s 2s 3s 4s. These results imply the seasonal moving average with order 1. We call it SMA(1). For the non-Seasonal component, we inspect the sample ACF and PACF. It is difficult for us to figure out the order p and q.

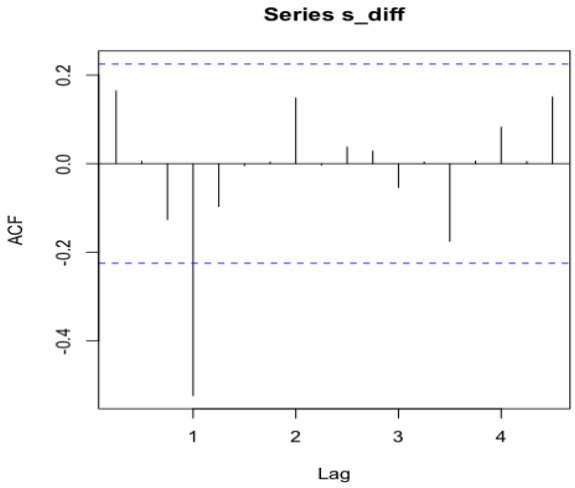


Fig.4 ACF plot

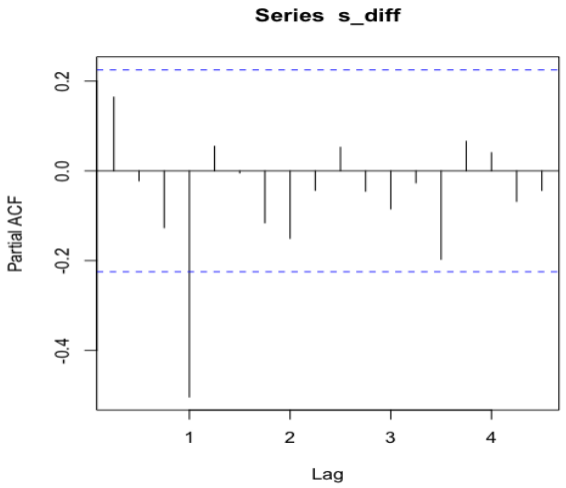


Fig.5 PACF plot

2.3 Estimation and Diagnostics

In the model selections, we can determine it is SMA(1) model, but it is hard to find order p and q. We will try several models to determine the order p and q. We try from the smallest order, which is

order 1 to see the significant level of parameters. For example, we try to fit the models like ARIMA(1,1,1),$(0,1,1)_4$,ARIMA(1,1,0),$(0,1,1)_4$,and ARIMA(0,1,1),$(0,1,1)_4$.The p-value of the parameters are shown in the Table2. For ARIMA(1,1,1),$(0,1,1)_4$,it has large p-values in the AR parameter and MA parameter. We can conclude that both AR MA parameters are not significant. We consider to decrease the order of the AR parameter to 0, which is ARIMA(0,1,1),$(0,1,1)_4$. It is still not significant in the MA parameter. As a result, we reduce the MA order to 0 to get our first adequate model, which is the model ARIMA(0,1,0)$(0,1,1)_4$. We call it as model1.

Table2

models	**ARIMA(1,1,1),$(0,1,1)_4$**	**ARIMA(1,1,0),$(0,1,1)_4$**	**ARIMA(0,1,1),$(0,1,1)_4$**
P-value(AR)	0.9899	0.4599	0
P-value(MA)	0.9160	0	0.4487

So our model1 is the following:

$$y_t = \log(x_t) \qquad (1 - B^4)(1 - B)y_t = (1 - 0.584B^4)w_t$$

As can be seen from the figure6, the residuals look like white noise. The sample ACF of the residuals is approximately zero for each positive lag. The residuals are approximately normally distributed. We use the Ljung-Box test to test for the autocorrelations. The null hyperthesis is that the autocorrelations are zero for lags h=1, 2, ..., H.The alternative hyperthesis of Ljung-Box test is that the autocorrelations are not zero for lags h=1, 2, ..., H. Since the pvalue of the Ljung-Box is greater than 0.1, we do not have evidence to against the null hyperthesis. The autocorrelations are zero in this case. We check for the significance level of the parameter. We find the pvalue is less than 0.01. It means the parameter is significant. AIC value is equal to -2.446121, which is vary small. As a result, the model 1 is adequate.

Table3

Parameter	Estimate	Standard error	pvalue	AIC
SMA1	-0.584	0.0885	<0.001	-2.446121

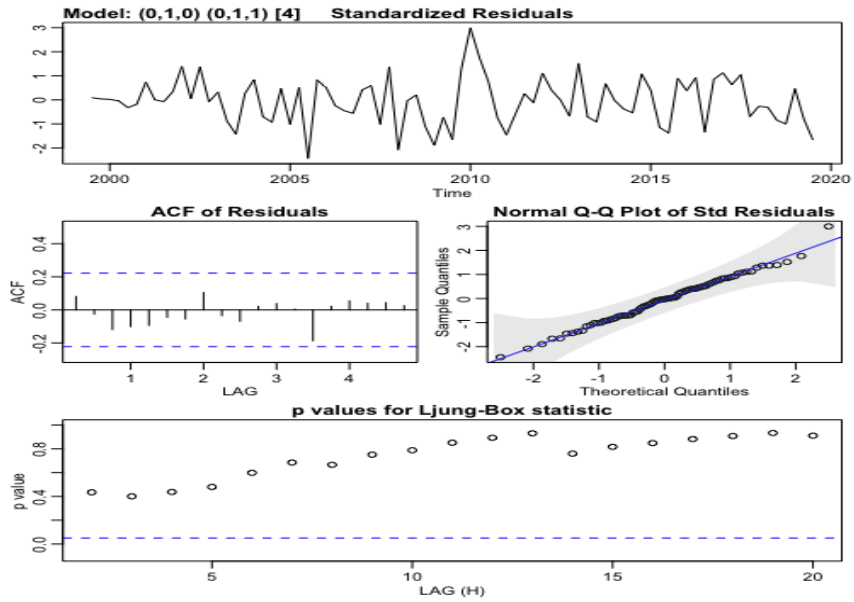


Fig.6 diagnostics plot

3. The ARIMA model and Intervention Analysis

In the last section, we find model1. However, there is a large spike around the year 2019 in the bottom plot of the figure3, which is the seasonal differenced log transformed data. It reminds us of the recession of 2008-2009 in Canada. Reported by Gordon, S. (2017), the global financial crisis that began in 2007 dragged much of the world economy into recession, and Canada was not spared. Since previous models have not investigated the impact of the sudden financial crisis on the housing sector, thus this paper attempts to examine and analyze the impact of this global financial crisis. To assess the impact of this event, the ARIMA intervention analysis will be applied in this study to evaluate the pattern. The present financial crisis is unique in that it was initially triggered in July 2007 and then spread to Canada. According to the Canada Year book (Statistics Canada, 2011)

construction report section, overall employment of construction fell 5.7% in 2009. In 2010, the economy saw signs of improvement, and employment in construction advanced by 4.9%. As a result, we can assume Canada housing market and housing completions had a level shift in the year 2009. We add a dummy variable that is defined as $I_t = \begin{cases} 1, & t \leq 2009 \\ 0, & t \geq 2009 \end{cases}$. We use the Box–Jenkins method again and find the best model with the dummy variable. We run the diagnostics of our second model. As can be seen from the figure7, the residuals look like white noise. The sample ACF of the residuals is approximately zero for each positive lag. The residuals are approximately normally distributed. We use the Ljung-Box test to test for the autocorrelation. The null hyperthesis of the Ljung-Box test is that the autocorrelations are zero for lags h=1, 2, ..., H.The alternative hyperthesis is that the autocorrelations are not zero for lags h=1, 2, ..., H. Since the pvalue of the Ljung-Box is greater than 0.1, we do not have evidence to against the null hyperthesis. The autocorrelations are zero in this case. We check for estimated AR and MA parameters. We find the parameters are significant. Most importantly, the pvalue of the dummy variable is less than 0.01.This parameter is significant. It means there is a level shift in 2009. AIC value is equal to -2.552442 in this case. The diagnostics indicate that our second model is also adequate.

Therefore,We get our second ARIMA(1,0,0)(0,1,1)$_4$ model. we call it as model2.

$y_t = \log(x_t)$ $(1 - B^4)$ (1-0.8647B) y_t=(1-0.5589B^4)$W_t - 0.1726I_t$

Table4

Parameter	Estimate	Standard error	pvalue
AR1	0.8647	0.0667	0
SMA1	-0.5589	0.0941	0
XREG	-0.1726	0.056	0.0029

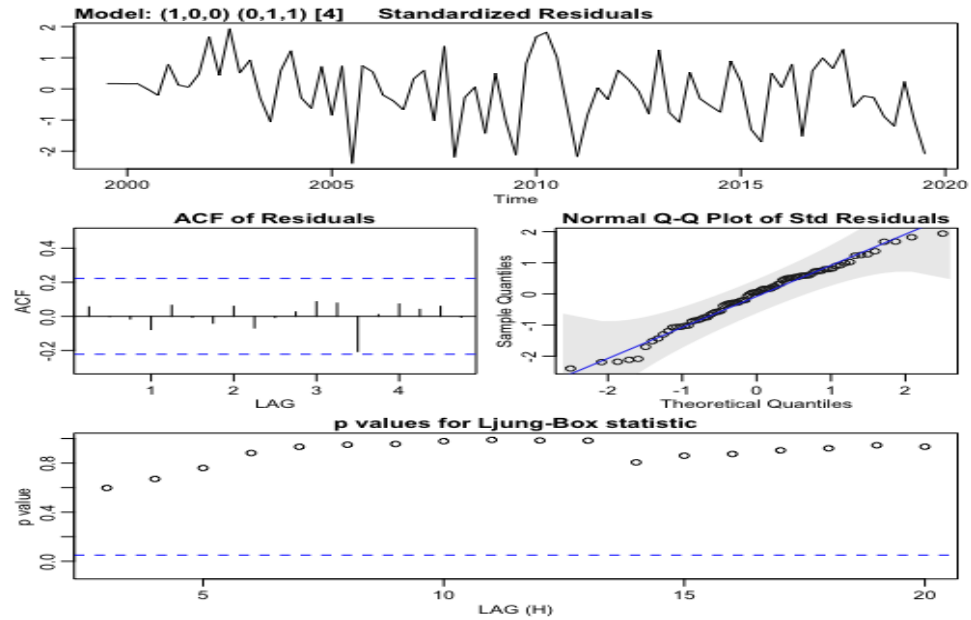


Fig.7 diagnostics plot

4. Model Comparison

For both model1 and model2, we want to compare them with their mean square error and their Akaike Information Criterion (AIC) values.

For comparing their mean square error, we first deleted last ten observations, which is from 2017 quarter2 to 2019 quarter 3.We call it observed values Y_t :9.779114 9.901986 9.934017 9.732581 9.773777 9.801344 9.794565 9.633711 9.633973 9.584659. We use model1 and build-in function in R Sarima.for to forecast these ten time points deleted before. For model1, we get our forecast values which are $\widehat{Y_t}$:9.738064 9.788819 9.869281 9.687604 9.732161 9.782916 9.863378 9.681701 9.726258 9.777013. We calculate our mean square error $\frac{1}{n}\sum (y_i\text{-}\hat{y}_i)^2$ = 0.007533216. Moreover, we forecast these 10 points by using our model2. For our model2, we get our forecast values which are 9.735647 9.779128 9.860877 9.680609 9.722643 9.766032 9.847700 9.667363 9.709336 9.752673. We calculate our mean square error $\frac{1}{n}\sum (y_i\text{-}\hat{y}_i)^2$ = 0.00676. We can find model2 has the smaller mean square error than model1. It means model2's fitted data is better than model1.

Moreover, model2's AIC is -2.55, which is smaller than model1's AIC -2.44. We can conclude that model2 is better than model1 in the mean square error and AIC perspectives.

To understand the model better, we also plot the original data, model1, and model2 in the same graph in figure 8. We draw the original data with black lines. We use blue lines to represent the model1 fitted values without the level-shift . We use red lines to represent the model2 fitted values with the level-shift model. As can be seen from the picture, both model1 and model2 fitted the data vary well. In general, it looks model1 has larger variance than model2. The difference can be seen more clearly around 2009. In the time point 2009, the red line, which represents the model1 has a significant drop to the value near 9.6. However, the actual data is 9.8. At the same point, model2 with the level shift just dropped to 9.7, which is closer to the observed value. It means model1 needs to use more variance to explain the massive decline in 2009, which causes it to loss more mean square error. Model2 use the level shift to illustrate that drop, which makes it gain more mean square error.

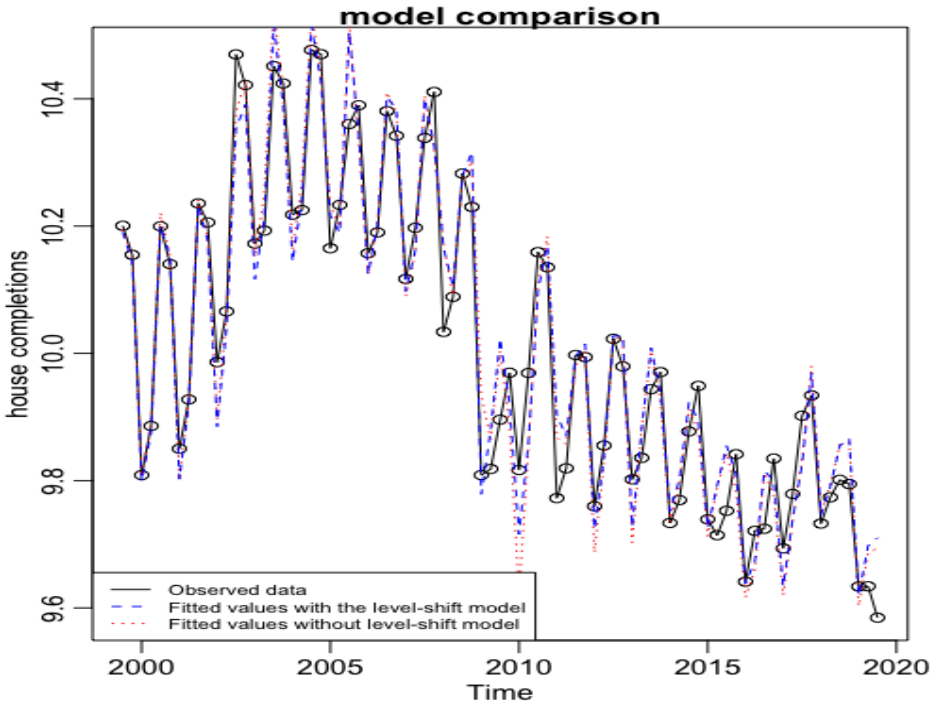


Fig.8 model comparison plot

5. Forecasting The Next 10 Data Points.

We finally get our best model which is model2. Without deleting any points, we use the build-in function in R Sarima.for again to forecast ten time points ahead for model2. The values we predict are 15327.12 for 2019 Q4, 13089.11 for 2020 Q1, 13725.32 for 2020 Q2, 14097.87 for 2020 Q3, 14874.76 for 2020 Q4, 12713.14 for 2021 Q1, 13340.63 for 2021 Q2, 14711.41 for 2021 Q3, 14475.07 for 2021Q4, and 12377.61 for 2022 Q1. We also draw the 95% confidence region for our predict values in figure9. As can be seen from the graph, the blue line, which is the forecasting line for the next two years goes down. It means there is still a downward trend in the following two years. We can say that housing completions will be continually decreasing in the next two years.

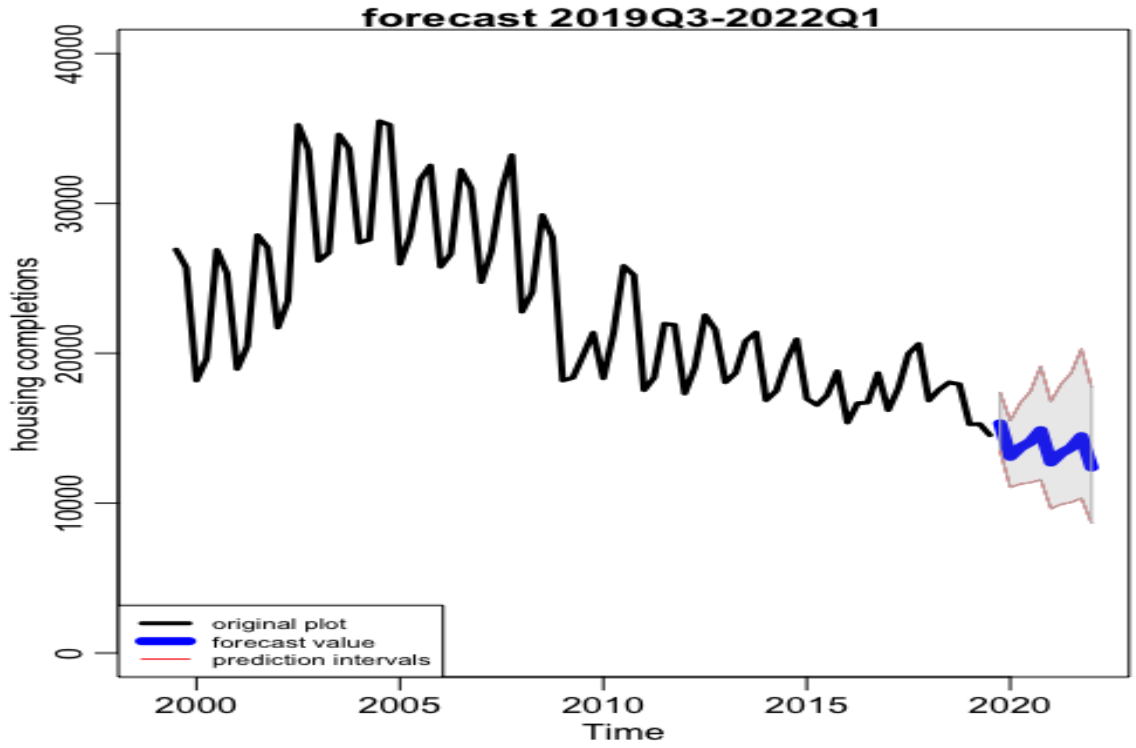


Fig.9 forecasting plot

6. Conclusion

In conclusion, Model2 is our best model.Our best model indicates the canadian housing completions will still perform a downward trend in the next two years. However, since the Canadian housing market is very sensitively related to the Canada policy and economic, it may have some upward trend in the future. We need to perform the inventional analysis again to explain the uptrend. In the future study, the non-linear model could be used to analyze this data.

7. References

[1] Box, G. E. P., & Jenkins, G. M. (1970). Statistical Models for Forecasting and Control.

[3] Gordon, S. (2017, October 27). Recession of 2008–09 in Canada. *the Canadian Encyclopedia*. Retrieved from https://thecanadianencyclopedia.ca/en/article/recession-of-200809-in-canada

[4] Statistics Canada, . (2011). Canada Year book. Ottawa: Statistics Canada.

[5] Statistics Canada. Table 34-10-0135-01 Canada Mortgage and Housing Corporation, housing starts, under construction and completions, all areas, quarterly