Modeling Cartel Gasoline Margins with Known Structural Breaks

Matthew Edwards and Benjamin Evans
Department of Economics
Queen's University

April 26, 2018

Abstract

This paper models weekly Canadian retail gasoline margins (percentage) for the town of Sherbrooke, Quebec, from June 9th 1998 to present. We break the series into four pieces, using three known structural breaks: The exposure of a gasoline cartel later investigated by the Competition Bureau (June 6, 2004), the 2008-2009 financial crisis, and the commencement of oil fracking in America in late 2014. We find that the model of best fit is different for each series, and conclude that these structural breaks have a significant impact on the underlying process. Performing the same analysis for the market of Montreal, we find two things. The cartel margins in Sherbrooke are statistically identical in process to the margins in Montreal. Second, the margins differ after cartel exposure in Sherbrooke, but not in Montreal. We conclude that there is little evidence of the cartel's impact on margins, but the launch of the investigation did impact the behavior of Sherbrooke retailers. The other two shocks affect both markets equally.

1 Introduction

The market power of retail gas stations varies considerably among and within Canadian provinces. With high startup costs (large entry barriers), the Canadian market consists of relatively large, but few, firms. In this oligopoly setting, the only way for firms to profitably engage in anticompetitive behavior is through collusion. In 2004, the Competition Bureau investigated the possibility of such a cartel in the province of Quebec[1]. Evidence revealed that prices were artificially raised through collusion of several retail stations across four different cities, over a period of at least four years. More recent works confirm the effectiveness of this cartel, finding statistically significant price asymmetries following the announcement of this investigation (Clark and Houde 2012), as well as a significant decline in the average change in price level (Erutku and Hildebrand 2010).

In the years since, Canadian gas stations underwent at least two other known exogenous events, in 2008-2009, and 2014. The former refers to the Great Recession, where multiple shocks caused contractions in almost every major macroeconomic aggregate (Christiano, Eichenbaum and Trabandt 2015). The consensus is that this crisis lasted from late 2007 to the summer of 2009, and several recent papers model the effects of this and other crises as structural breaks (Wang and Zivot 2000; Baumeister and Kilian 2016a). Due to the dramatic spike in margins during this time, we omit these observations from our analysis.

[1] All documentation on the charges and sentences can be found at the Competition Bureau's website: "http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03079.html"

The change in 2014 comes from the beginning of fracking by U.S. oil companies. When crude oil supply increases, the input prices of intermediate refinements necessary for production of gasoline fall. This lowering of cost augments retail gas margins, as we will show later. There are several works on the direct impact of this event on U.S. markets (Baumeister and Killian 2016b) and we hypothesize that there were spillover effects into Canada. Quebec is one of Canada's largest crude oil importers; in 2012 alone, the province imported almost 14 billion dollars of crude, which was, at the time, Quebec's largest import (Chassin 2013).

This paper uses weekly retail gasoline margins from the towns of Sherbrooke and Montreal, QC, to examine the structural changes caused by these three known events. In particular, we use a series of quantitative and qualitative techniques to fit ARIMA models to each of the four periods. We then suggest a unified theory (model) that can explain these changes. We find that (from left to right per Figure 1), low-order ARIMA models fit the second and fourth periods reasonably well, but the first and third segments are closer to non-stationarity.

2 Data

In 2004, The Canadian Competition Bureau conducted an investigation after being notified of potential collusive behaviour in four Quebec gasoline markets; Magog, Victoriaville, Sherbrooke, Thetford Mines. They determined that, in all four markets, the collusive period started in 2002. Clark and Houde (2009) determined that the cartels in these markets collapsed after the Competition Bureau announced their investigation in June 2006 by showing a drop in profit margins of the alleged cartel members following the announcement. However, it is important to note that the Competition Bureau initially began their investigation in June 2004. This followed the release of a news article from La Nouvelle, a local newspaper, stating that a gas station owner in Victoriaville was being pressured by other owners to fix his prices. We believe this article acted as a "whistle blower" which forced the cartel to act more cautiously. Thus, in our analysis, we will determine whether there is a break in profit margins following the release of this newspaper article.

The variables used in this analysis are average profit margins in the Montreal and Sherbrooke markets for regular gasoline. To measure the cartel, we use only the Sherbrooke market because it is the largest available series and the effect need only be observed in one city. The data used to determine profit margins are collected by Kent Group. This organization conducts weekly surveys on daily after-tax retail and rack (wholesale) prices for regular gasoline from all firms in various Canadian markets. The observed values in the data are averages of surveyed prices for all firms in each market. The profit margins in our analysis are determined by taking the retail price, subtracting the rack price (the cost for retailers) and dividing by retail price. The rack prices are collected from the Montreal wholesale market, the closest metropolitan area to Sherbrooke. We assume that Sherbrooke gasoline retailers purchase from Montreal suppliers. Data are weekly from June 9th, 1998 to March 13th, 2018, for a total of 1032 observations. We purposely exclude July 29th, 2008 to April 7th, 2009 because there are outliers which are not seasonal components of our series.

Table 1 shows the summary statistics by city for each of the four periods. We see clear differences across time. The series are much more volatile in the early periods, and tends to stabilize as we move closer to present day. Also, margins are on average, higher before than after the investigation by a significant amount (over 25 percent). Note also that the means switch from being above to below the median in some of the periods. The series are never higher than in the beginning

Table 1: Summary Statistics for Percentage Retail Margins (By Structural Period)

Period	City	Min	Max	Mean	Median	Variance	N
1(9/6/1998-8/6/2004)	Sherbrooke	0.4169	0.7434	0.5727	0.5587	0.00389	314
2(8/6/2004-29/7/2008)	Sherbrooke	0.3089	0.5385	0.4330	0.4336	0.00231	216
3(14/4/2009-4/11/2014)	Sherbrooke	0.3405	0.5305	0.4052	0.4040	0.00069	290
4(11/11/2014-Present)	Sherbrooke	0.3681	0.5314	0.4570	0.4607	0.00090	175
1(9/6/1998-8/6/2004)	Montreal	0.4316	0.7486	0.5682	0.5500	0.00430	314
2(8/6/2004-29/7/2008)	Montreal	0.3037	0.5283	0.4276	0.4305	0.00202	216
3(14/4/2009-4/11/2014)	Montreal	0.3433	0.4745	0.4131	0.4140	0.00049	290
4(11/11/2014-Present)	Montreal	0.4139	0.5433	0.4744	0.4737	0.00045	175

We see that the variance of the series declines over time. Also, margins are, on average, much greater before the news article than after, but seem to be back on the rise as of the last period.

(Figures 1,2 and 3 show this), as gas margins may fall over time in the face of increased competition.

It is important to note some of the limitations in our data. The Competition Bureau identified and charged several, but not all gas stations in Sherbrooke for their involvement in the cartel. Thus, we know which stations were involved, but we cannot isolate for them. This would be a confidentiality breach on individual gas stations and their historical price setting. Thus our data contains the entire market of Sherbrooke (with collusive and non-collusive retailers). It would be ideal to isolate for cartel members only and observe their profit margins before, during and after the collusive period. The inclusion of non-collusive retailers may lead to a lessened effect on observed profit margins. However, we still expect a significant drop in the market. We also assume all gas stations in Sherbrooke and Montreal face the same cost, but it is possible that several retailers have long-term agreements with suppliers for cheaper than average rack prices.

3 Empirical Analysis

With known breakpoints, we split each series into four periods, as labeled in Table 1. We fit the two series separately, and find that the best candidate models change with each period. Both qualitative and quantitative techniques are used to select the best candidates.

3.1 Period 1: 9th June 1998 - 8th June 2004

3.1.1 Sherbrooke

This first period contains 314 observations, and is the longest. The series has a clear downward trend, which we remove by an Ordinary Least Squares regression of margin on time. Both coefficients significant at all standard levels. Per Figure 1 and Table 1, we can see that this series is the most volatile, but it is mean reverting (it reverts to the trend, which we remove), suggesting stationarity. We find the autocorrelation and partial autocorrelation of the detrended series. There is a high degree of persistence in the ACF, which signals a relatively high value for $\hat{\phi}_1$. The PACF indicates a small order autoregressive component ($p = 2$ or $p = 3$). We use a series of tests for stationarity to determine whether this coefficient is exactly equal to one. Note that differencing this series induces negative serial correlation, which is an initial sign that the detrended process is not unit root. Also, the ACF decays faster than series we know to be fractionally or fully integrated, such as stock data.

The first stationarity test is the Augmented Dickey-Fuller Test, which puts the null hypothesis of unit root ($\hat{\gamma} = 1$) against alternatives of either stationarity ($\hat{\gamma} - 1 < 0$), or explosiveness ($\hat{\gamma} - 1 > 0$). For lags chosen between 10 and 20, we find that the test retains the null under both alternatives, with p-values of at least 0.4 (significant at all standard levels). However, when we run the Phillips-Perron test for non-stationarity, with the same null and alternative hypotheses as ADF, we find that the test rejects the unit root, with a p-value of less than 0.01. With conflicting results, we run a third test, called the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The null hypothesis can be either trend or level stationarity (we choose level stationary, since the series is already detrended), against the alternative of a unit root. This test fails to reject the null when using short lag truncation, but the p-value is 0.036. For long lag truncation, we do not reject the null. This is evidence of stationarity. Overall, we do not have sufficient evidence to suggest the process is unit-root or explosive (or trend stationary). We proceed to ARMA fitting.

The leading candidate according to the PACF and ACF of the detrended sequence is an AR(2), so we start by fitting that. We also fit an AR(1), ARMA(1,1), ARMA(2,1) and ARMA(2,2). Of these four, the ARMA(2,1) gives the smallest Akaike and Hannan-Quinn Information Criteria, but the ARMA(1,1) gives the best Bayesian Information Criterion. We take the two sets of residuals, and run portmanteau tests as well as Lagrange Multiplier (Breusch-Pagan) tests for independence. In both models, we fail to reject the null hypothesis of independence using the Ljung-Box and Box-Pierce test statistics. We also fail to reject for a wide range of lags in the BP test. It appears both models fit well. Note that the ACFs and PACFs of the model residuals look nearly identical, with some of the lag structures from the ARMA(1,1) being smaller in magnitude.

Overall, we conclude (qualitatively and quantitatively) that the first period is best estimated with the ARMA(1,1):

$$margin_t = \underset{(0.0051)}{(0.647)} - \underset{(2.83e-05)}{(0.00047)}t + \underset{(0.0279)}{(0.9015)}margin_{t-1} + \epsilon_t - \underset{(0.0665)}{(0.1883)}\epsilon_{t-1}$$

Note the high AR(1) coefficient, which explains the persistence we see in the initial ACF.

3.1.2 Montreal

As with Sherbrooke, we remove a linear time trend from the data before fitting a model. We find that both the negative slope coefficient and positive intercept are significant. The Autocorrelation and Partial Autocorrelation are very similar to Sherbrooke. The ACF shows long memory, and the PACF shows a small AR component (perhaps 1 or 2). We find that Sherbrooke and Montreal look very similar, suggesting that they may both have the same underlying process. High persistence in the ACF of Montreal suggests there may be a unit root, but we dismiss this initially because differencing induces negative serial correlation.

Next, we formally test for stationarity, as we did for Sherbrooke. We find mixed results. Under the augmented Dickey-Fuller test, we find (under alternatives of both stationary and explosive) failure to reject the unit-root null, with a p-value of 0.5. The choice of lag is identical to the ADF test for Sherbrooke. However, the Phillips-Perron test indicates rejection of the null, with a p-value of less than 0.01. This, taken with the ACF and PACF, suggests that the process may be a low-order ARMA model with high coefficients (near one) that skew the test results. With this in mind, we fit an AR(2) to the data. This gives coefficients of 0.61 and 0.25 (both significant), which together are very close to 1. One of the stationarity conditions for the AR(2) is that the coefficients sum to a value less than one, indicating that our process is close to non-stationarity.

Plotting the residuals of the AR(2), we see reasonably good behavior. All lags of the ACF are within the confidence interval (except for lag zero), as are those of the PACF. This is indicative of white noise, meaning the model is a good fit. Before we can be sure that the residuals are white noise, we run several tests for serial correlation. The Box-Pierce and Ljung-Box portmanteau tests both fail to indicate rejection of the null hypothesis (no serial correlation), as does the Breusch-Pagan (Lagrange Multiplier) test. All three tests were run with a maximum lag of 15 but results still hold for all positive lag choices below that.

Knowing that the AR(2) fits the series well, we fit several other models related to it. The ARMA(2,1) and ARMA(1,1) also fit the data well, but the AR(1) does not produce serially uncorrelated residuals. To determine which of the three candidates is the best choice, we use several information criteria. We find that the Akaike and Bayesian information criteria indicate the ARMA(1,1) is better, but the Hannan-Quinn criterion favors the ARMA(2,1). Since the majority of evidence favors the more parsimonious ARMA(1,1), we conclude that this model is the best fit. The model, including trend, is formally represented by the following equation:

$$margin_t = \underset{(0.0052)}{(0.650)} - \underset{(2.84e{-}05)}{(0.00052)}t + \underset{(0.0234)}{(0.931)}margin_{t-1} + \epsilon_t - \underset{(0.0654)}{(0.334)}\epsilon_{t-1}$$

Interestingly, this is the same model that we find for Sherbrooke, indicating that the presence of a cartel had no statistically significant effect on the pre-investigation process for Sherbrooke. The coefficients differ slightly between the two markets, but the difference is negligible.

3.2 Period 2: 8th June 2004 - 29th July 2008

3.2.1 Sherbrooke

This subset of the original series contains 216 observations, stopping just before the great recession in late 2009. Per Figure 1, we see a clear downward trend in the data, which we assume to be linear. We detrend these margins as before. Coefficients on the intercept and slope are statistically non-zero at all major levels of significance. We then take the residuals from this regression for ARMA fitting.

To begin, we run tests to determine whether the residuals are stationary. We run the Augmented Dickey-Fuller test first and reject the unit root null for the stationary alternative. This holds at all major levels of significance. The Phillips-Perron test for unit roots also rejects the null. The reported p-value is 0.01, but this is taken from a table, and the actual value is much smaller. Note that in the ADF test, the results hold for chosen lags in the range of 10-20[2]. The PP test results hold, regardless of whether we use small or large lag truncation. Taken together, both of these tests provide strong evidence that the residuals are stationary, and can be fitted with ARMA

[2]Using too many lags reduces the power of the test. As a starting point, we use the rule given by Schwert (1989), $k = 12 * (T/100)^{1/4}$.

models.

To begin, we plot the autocorrelation and partial autocorrelation functions of the detrended series. On the autocorrelation, we see a smooth decay which suggests an autoregressive structure. The partial autocorrelation shows that the order of AR component is small, since the graph falls within the confidence bounds after only three lags. Using these clues as a starting point, we fit an AR(3) process to the model. For completeness, we also fit AR(2), AR(4), and ARMA(3,1) models.

For all models, we see significant coefficients at the one percent level, with the exception of the fourth autoregressive coefficient in the AR(4) ($\hat{\phi}_4$). Note that all models produce residuals that are reasonably well-behaved. In particular, each series of residuals falls within the confidence intervals on the correlogram (with the exception of lag zero), and each series has statistically insignificant partial autocorrelations[3]. In order to reduce the choice of candidates, we use Akaike and Bayesian Information Criteria. This produces useful, but mixed results. We find the AIC shows that the AR(3) is optimal, but the BIC indicates that the AR(2) is preferred. To break the tie, we use in-sample forecasting.

We split this series into a training set (observations 1 to 166), and a testing set (the last 50 observations). We use point forecasting to generate predictions for the final 50 observations[4]. Effectiveness of forecasts are evaluated using mean squared forecasting error. For robustness, we run 500 repetitions of the forecasts, and we find that the AR(3) model gives the lower error every time. Results do not change, even when we increase the size of the testing set from 50 to 100. This evidence strongly suggests that, for the second period of our series, the AR(3) is the best fit among our candidate models.

The model is only complete if the residuals are white noise. We already know that the residuals of the AR(3) resemble those of white noise, based on autocorrelation and partial autocorrelation. To confirm that they are independent, we use portmanteau tests. Both Ljung-Box and Box-Pierce tests have a null hypothesis of independence. Both statistics fail to reject the null at a wide range of lags. We vary the lag choice between 10 and 20 (adjusting the degrees of freedom accordingly). Overall, we have evidence that the AR(3) fits the data well. The best candidate model for the second period takes the following structure (standard errors listed below):

$$margin_t = \underset{(0.0044)}{(0.496)} - \underset{(3.5e-0.5)}{(0.0006)}t + \underset{(0.0673)}{(0.4331)}margin_{t-1} + \underset{(0.0725)}{(0.1760)}margin_{t-2} + \underset{(0.0675)}{(0.1268)}margin_{t-3} + \epsilon_t$$

In terms of parameters changing over time, we can see that the coefficient $\hat{\phi}_1$ has dropped in value significantly compared to the first period. This second segment is much less persistent.

3.2.2 Montreal

We remove the linear trend, for which both slope and intercept are significant. The ACF and the PACF of the detrended series indicate a small autoregressive component will provide a decent fit, and that the series is stationary. Again, formal tests for unit roots give mixed results. The ADF test retains the unit root null, albeit with a questionable p-value of 0.09. The Phillips-Perron test

[3] Note that two of the PACFs (AR(2) and ARMA(3,1)) have one lag outside the confidence bounds, but with 95 percent confidence, we expect 1 in 20 lags of white noise residuals to lie outside the bounds.

[4] The R package *forecast* uses the method of exponential smoothing.

again rejects the null, with a p-value of below 0.01.

Next, we fit an AR(2) and find both coefficients are significant. Taking the residuals, and plotting the ACF and PACF, we find almost all lags (one exception) are within the confidence intervals. We formally test for serial correlation as before, using Ljung-Box, Box-Pierce and Breusch-Pagan tests, with maximum lags as high as 15. We find evidence for not rejecting the null; in other words, this is evidence indicating no serial correlation. We conclude that the AR(2) fits the detrended series well. As before, we fit other surrounding candidate models. These indicate, as in the previous period, that the ARMA(1,1) and ARMA(2,1) also provide a decent fit according to the evaluations used above.

To determine which model is best, we use the three information criteria mentioned in the first period analysis: Akaike, Bayesian and Hannan-Quinn. All three criteria indicate that the ARMA(1,1) is the best choice of the three. The model of best fit is given by the following equation:

$$margin_t = \underset{(0.0042)}{(0.485)} - \underset{(3.34e-05)}{(0.00053)}t + \underset{(0.0496)}{(0.876)}margin_{t-1} + \epsilon_t - \underset{(0.0866)}{(0.533)}\epsilon_{t-1}$$

We can see that this model is very similar to the model of best fit for Montreal's first period. Also, it is very different from the AR(3) that best fits Sherbrooke's second period.

3.3 Period 3: 14th April 2009 - 4th November 2014

3.3.1 Sherbrooke

This series uses 290 observations, beginning just after the Great Recession, and continuing until just before the start of U.S. crude oil fracking. There does not appear to be a time trend on this sequence. The autocorrelation and partial autocorrelation functions of the series show signs of both moving average and autoregressive components. There is a sharp decay for the first three lags of the ACF, followed by some persistence until around lag 7. The PACF shows three significant lags, but the latter two are nearly within the confidence interval. This information is useful, but does not indicate exactly where to start.

To avoid missing a model that fits reasonably well, we fit all ARMA models ranging from ARMA(1,1) to ARMA (2,3) (six in total). We use the AIC and BIC to narrow down the selection, and both criterion indicate that the ARMA(2,1) is the best choice of the six models tested. However, we find that several of the models show coefficients that are just below or above one. Although the initial plots indicated otherwise, these coefficients indicate that the series may be non-stationary. To test this, we run the usual ADF, PP and KPSS tests. The ADF and KPSS tests indicate a unit root process, while the PP test provides evidence of stationarity.

Using the three information criteria above, the best recommendation is the ARIMA(1,1,1). We see that the coefficients of the differenced series are within the limits of stationarity and invertibility ($\hat{\phi}, \hat{\theta} < 1$). The ACF and PCF of the model's residuals, which all lie within the confidence intervals, with the exception of one lag in partial autocorrelation. With this in mind, we run portmanteau tests on the residuals to see if these innovations are independent. The Ljung-Box and Box-Pierce test statistics both fail to reject the independent null with p-values of 0.2536 and 0.2969. These results do not change under long lag truncation.

Our best candidate for the third period is represented by the following structure (standard errors reported below values):

$$\Delta margin_t = 0.4052 + \underset{(0.0766)}{(0.4009)} margin_{t-1} + \epsilon_t - \underset{(0.0432)}{(0.8992)} \epsilon_{t-1}$$

Note that this is the first period whose best candidate model is integrated. The other segments showed signs of long memory, but differencing was ineffective. This is evidence that the Great Recession changed the structure of Canadian gas margins. These results show the same model when the breakpoint is moved either forward or backward by one week.

3.3.2 Montreal

The behaviour of this segment draws many parallels to the same segment for Sherbrooke throughout our analysis. Referring to Figure 3, there does not appear to be a time trend on this segment. The autocorrelation and partial autocorrelation functions of the series show signs of both moving average and autoregressive components. There is a sharp decay for the first two or three lags of the ACF, followed by some persistence until around lag 25. The PACF shows two significant lags, but the latter one is nearly within the confidence interval. These results provide some useful information, but do not strongly suggest an appropriate approach to the model selection for this segment.

We fit all ARMA models ranging from ARMA(1,1) to ARMA (2,3) (six in total). AIC and BIC indicate that the ARMA(2,1) is the best choice of the six. However, we find that all of the models show autoregressive coefficients that do not violate non-stationarity. Although the initial plots indicated otherwise, this indicates that the series may be non-stationary. To test this, we run the ADF, PP and KPSS tests. The ADF and KPSS tests indicate a unit root process, while the PP test provides evidence for stationarity.

Using information criteria, the best recommendation is again the ARIMA(1,1,1). We find that the coefficients of the differenced series are within the limits of stationarity and invertibility $(, < 1)$. The residuals all lie within the confidence intervals, with the exception of two lags of partial autocorrelation (which we expect). Portmanteau test statistics both fail to reject the independent null with p-values of 0.4241 and 0.4615, respectively. These results do not change under long lag truncation.

Our best candidate for the third period is represented by the following structure (standard errors reported below values):

$$\Delta margin_t = (0.4130) + \underset{(0.0689)}{(0.2080)} margin_{t-1} + \epsilon_t - \underset{(0.0337)}{(0.9033)} \epsilon_{t-1}$$

Note that this is the same ARIMA ordering for Sherbrooke, although the coefficients are slightly different.

3.4 Period 4: 11th November 2014 - Present

3.4.1 Sherbrooke

This period has 175 observations, beginning just after the start of oil fracking in the U.S. in late 2014. The PACF suggests high persistence as the first autoregressive coefficient is approximately

0.8. We do see smooth decay on the ACF, and only one significant lag on the PACF. These two signs warrant an initial candidate model of AR(1). To qualitatively validate this interpretation, we simulate the ACF and PACF of a theoretical AR(1), both of which are shown in Figure 4. The two sets of plots are nearly identical.

As before, we run unit root tests. The results are confounding. The ADF test retains the null hypothesis of unit root under both possible alternatives (stationary and explosive). P-values are not marginal, as both alternatives give values of around 0.5. However, the PP test rejects the unit root null with a p-value of 0.01 at both long and short lag truncation. Similarly, the KPSS test retains the null of level stationarity with a p-value of 0.07. Between these tests and the autocorrelation structure of the series, we find that most evidence supports stationarity. With this in mind, we proceed to ARMA fitting.

In addition to the AR(1) indicated by the correlogram and partial autocorrelation plots, we also fit an AR(2), an ARMA(1,1) and an ARMA(1,2). Both the AIC and BIC suggest that the AR(2) is the optimal choice. However, the second autoregressive coefficient, $\hat{\phi}_2$, is insignificant (p-value of 0.33). In each of the four models, the only significant coefficient is the AR(1). Next we run a series of likelihood ratio tests to determine whether a model larger than the AR(1) is necessary.

For each of the other three candidates, we fail to reject the null of AR(1) at all major levels of significance. This is strong evidence that, of the four candidates for period 4, the AR(1) is the best choice. Note that this is again different than each of the other periods. We use portmanteau tests on the residuals. The ACF and PACF of the AR(1) residuals, although not shown, are reasonably well behaved. In particular, all lags lie within the confidence intervals. For the Ljung-Box test, we get a Chi-square statistic of 12.91, with a p-value of 0.163. Similarly, the Box-Pierce test statistic is 12.31, with a p-value of 0.197. Both of these results use a chosen lag choice of 10, but the test response does not change when we alter that. Together, these two tests support the null hypothesis of independence of errors, providing further evidence that the AR(1) produces white noise errors.

To summarize, the candidate model of best fit for this forth period takes the following structure (standard error reported underneath):

$$margin_t = 0.457 + \underset{(0.0441)}{(0.8093)} margin_{t-1} + \epsilon_t$$

This is a much simpler structure than we find in the other three periods. The coefficient value changes slightly if we move the date of break forward or back by one week, but the model structure (AR(1) with drift) does not.

3.4.2 Montreal

We find that this period has a high degree of persistence, similar to the segment for Sherbrooke. The PACF confirms this as the first autoregressive coefficient is approximately 0.7. There is see smooth decay on the ACF, and three significant lags on the PACF. These two signs suggest an initial candidate model of AR(3).

Following the same procedure as for Sherbrooke, we run unit root tests but the results are confounding. The ADF test fails to reject the null hypothesis of unit root under both possible alternatives (stationary and explosive). P-values are not marginal, as both alternatives give values of around 0.5. However, the PP test rejects the unit root null with a p-value of 0.01 at both long and short lag truncation. Similarly, the KPSS test retains the null of level stationarity with a p-value greater than 0.10. Between these tests and the autocorrelation structure of the series, we find that most evidence supports stationarity. We then proceed to ARMA fitting.

In addition to AR(3) indicated by the correlogram and partial autocorrelation plots, we also fit an AR(1), an AR(2), an ARMA(1,1) and an ARMA(1,2). Both the BIC and AIC suggest that the AR(3) is the optimal choice. The first and third autoregressive coefficient are significant at a 99 percent confidence level. However, the second autoregressive coefficient, ϕ_2, is insignificant (p-value of 0.14).

Next we run a series of LR tests to determine whether a model less than the AR(3) is a better fit. For each of the other three candidates, we reject the null of AR(1), ARMA(1,1) and AR(2) at all major levels of significance. This is strong evidence that, of the four candidates for period 4, the AR(3) is the best choice. Note that this is different than each of the other periods and different than the suggested model for Sherbrooke. To check robustness, we use portmanteau tests on the residuals. The ACF and PACF of the AR(3) residuals, are reasonably well behaved. For the Ljung-Box test, we get a Chi-square statistic of 18.03, with a p-value of 0.0348. Similarly, the Box-Pierce test statistic is 17.00, with a p-value of 0.0486. Both of these results use a chosen lag choice of 10, but the results of the test still fail to reject the null as the p-value increases significantly with less or more lags. Together, these two tests fail to reject the null hypothesis of independence of errors, providing further evidence that the AR(3) produces white noise errors. We conclude that this model is the best fit. The process is stated formally in the following equation:

$$margin_t = (0.4744) + \underset{(0.0712)}{(0.5772)} margin_{t-1} + \underset{(0.0831)}{(0.0870)} margin_{t-2} + \underset{(0.0709)}{(0.3355)} margin_{t-3} + \epsilon_t$$

4 Results

To summarize, we fit a separate model on each of the four periods. Each period has a different best candidate model. We find period one is best modeled using an ARMA(1,1), while the second period indicates an AR(3). The third period appears to follow an ARIMA(1,1,1), while the last period follows a simple AR(1). These results are robust to small changes in the timing of structural breaks. Overall, the variance of the entire series appears to decrease over time. The differences between models after each break are evidence of the structural changes we hypothesize. The following table shows the model of best fit by city and period:

Table 2: Models of Best Fit for Sherbrooke and Montreal (By Structural Period)

Period	Sherbrooke	Montreal
1(9/6/1998-8/6/2004)	ARMA(1,1)	ARMA(1,1)
2(8/6/2004-29/7/2008)	AR(3)	ARMA(1,1)
3(14/4/2009-4/11/2014)	ARIMA(1,1,1)	ARIMA(1,1,1)
4(11/11/2014-Present)	AR(1)	AR(3)

To verify that the changing in processes are not just due to high volatility, we plot rolling values of the first autoregressive coefficient over subsamples of size 52. Figure 5 shows these plots. For both cities, the coefficient is almost always significant, and we do see that the structural breaks cause significant changes in the process.

5 Conclusion and Implications

Regarding the antitrust implications, we assume a structural break in profit margins would occur in mid-2004 after a newspaper article outlined price-fixing behavior allegations against cartel members. The stochastic nature (coefficients near one) of the first segment does not seem surprising considering that it includes both the pre-collusive (1998 to 2002) and collusive (2002 to 2004) periods. The difference in our model selection for the first and second segments suggests that the break has lead to a difference in predicting the series. The market profit margin in Sherbrooke dropped to due members of the cartel tightening their profit margins to reduce suspicion, whereas Montreal's process did not change.

Regarding macroeconomic implications, we expect a structural break to occur due to the Great Recession. Observing the plotted series and the summary statistics alone, we see that the second segment has a decreasing trend component and a larger variance, whereas the third component is more constant with far less variance (In both Montreal and Sherbrooke). From our empirical analysis and model selection, we conclude that these segments do not fit the same model. Our results suggest that a structural break is present between segments 2 and 3.

The other event occurs from increased supply of crude oil in Quebec primarily caused by the introduction of fracking in the U.S. in the early 2010s. We expect the positive supply shock to lower wholesale and retail prices of gasoline, but the margins of retailers to increase through lower incurred costs. Observing the plotted series and the summary statistics alone, we see that both segments have a constant component and a low variance compared to the other segments in the Sherbrooke series. However, the fourth segment contains a higher mean, suggesting a permanent increase to profit margins following the supply shock. We determine that these segments are not fitted by the same model. This indicates that there is a structural break between segments 3 and 4 in both Sherbrooke and Montreal.

To determine if the break in June 2004 and the decreasing trend post-"whistle blower" is exclusive to Sherbrooke, we model the regular gasoline market of Montreal. Montreal is fit for a control market since it is one of the closest markets to Sherbrooke, there has been little to no evidence of gasoline cartels present and the rack prices for both markets are similar due to their close proximity. We confirm that the structural breaks in margins caused by the Great Recession and the increased supply from oil fracking are not exclusive to the market in Sherbrooke. However, the 2004 investigation had no effect on Montreal.

6 References

Baumeister, Christiane, and Lutz Kilian. "Forty Years of Oil Price Fluctuations: Why the Price of Oil May Still Surprise Us." Journal of Economic Perspectives, vol. 30, no. 1, 2016, pp. 139-160.

Baumeister, Christiane, and Lutz Kilian. ”Lower Oil Prices and the U.S. Economy: Is This Time Different?” Brookings Papers on Economic Activity, vol. 2016, no. 2, 2016, pp. 287-357.

Chassin, Youri. “Quebec’s Energy Reality.” IEDM, Montreal Economic Institute, 16 Apr. 2013, www.iedm.org/43933-quebec-s-energy-reality.

Christiano, Lawrence, et al. ”Understanding the Great Recession.” American Economic Journal: Macroeconomics, vol. 7, no. 1, Jan. 2015, pp. 110-167.

Clark, Robert, and Jean-François Houde. ”The Effect of Explicit Communication on Pricing: Evidence from the Collapse of a Gasoline Cartel.” The Journal of Industrial Economics, vol. 62, no. 2, 2014, pp. 191-228.

Erutku, Can, and Vincent A Hildebrand. ”Conspiracy at the Pump”. The Journal of Law and Economics, vol. 53, no. 1, Feb. 2010, pp. 223-237.

Wang, Jiahui, and Eric Zivot. ”A Bayesian Time Series Model of Multiple Structural Changes in Level, Trend and Variance.” Journal of Business and Economic Statistics, vol. 18, no. 3, July 2000, pp. 374-386.

7 Appendix: Figures

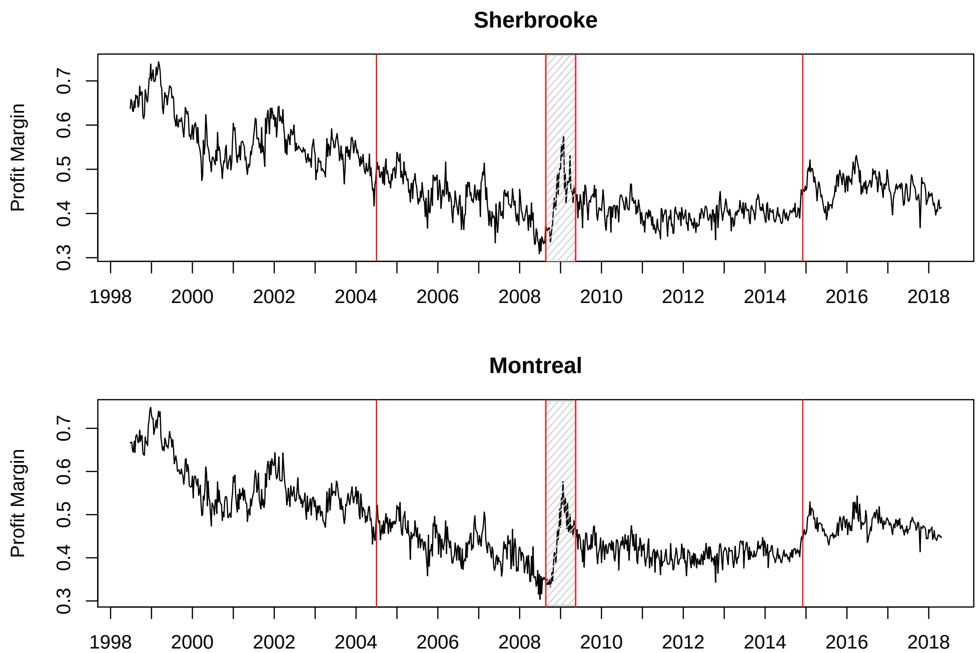

Figure 1: Weekly Canadian Retail Gas Margins For Sherbrooke, QC (9th June 1998 - Present). The three structural breaks are, from left to right: The disclosure of the existence of the cartel (week of June 8, 2004), the great recession (Aug 2008 to 7 April 2009) and the start of U.S. oil fracking (4 Nov 2014).

Figure 2: Weekly Canadian Retail Gas Margins For Sherbrooke, QC (separated by period).

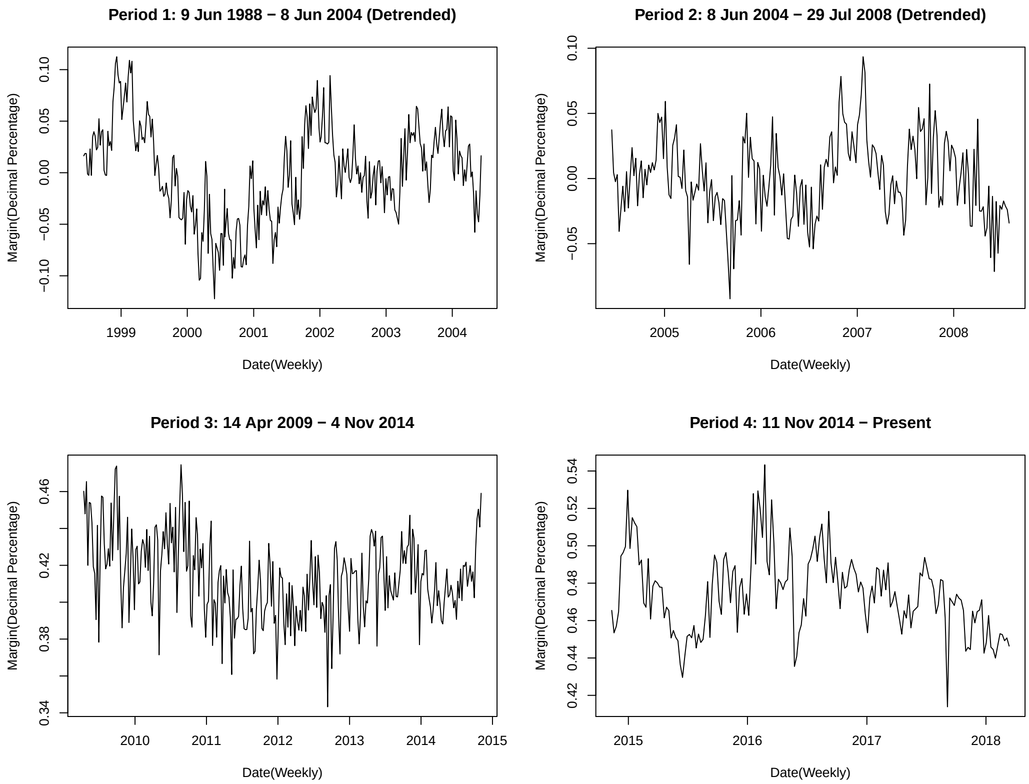

Figure 3: Weekly Canadian Retail Gas Margins For Montreal, QC (separated by period). Selected series are detrended when necessary to allow for better ARIMA fitting.

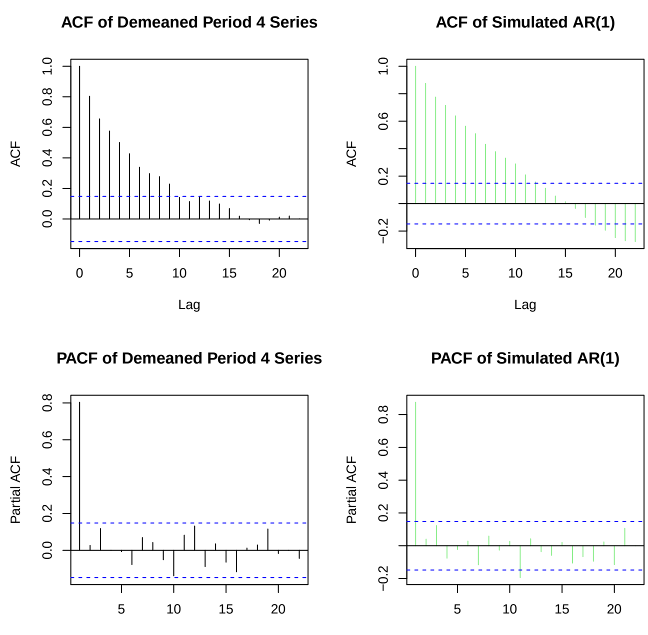

Figure 4: Autocorrelation and partial autocorrelation functions of the demeaned period 4 series and a theoretical AR(1) series with $\phi_1 = 0.8$. Notice that the two are nearly identical.

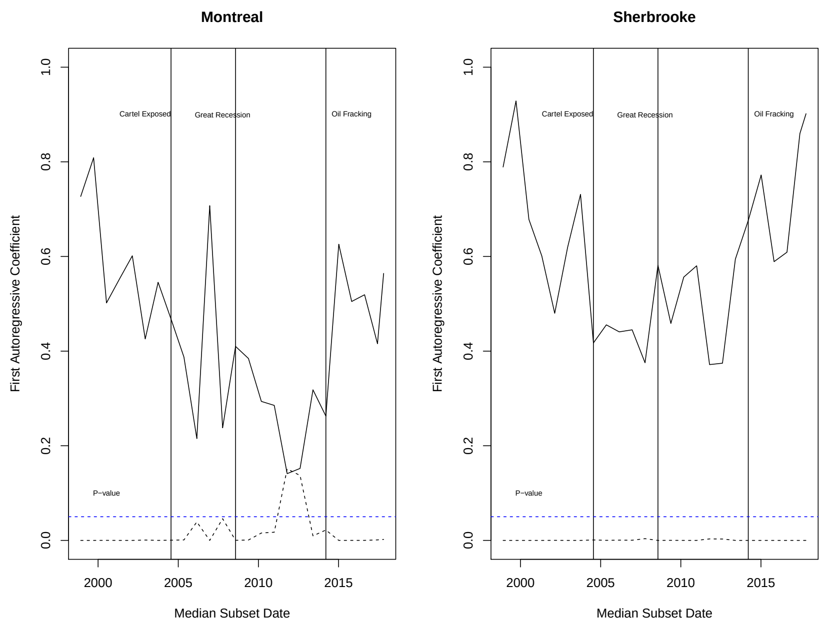

Figure 5: First Autoregression Coefficient ($\hat{\phi}_1$) Values and P-values for Rolling Subsets. Beginning at the first observation, we take subsamples of size 52, with an overlap of 10 (e.g. 1-52, 42-94, 84-136,...). Confidence Intervals represent $\alpha = 0.05$.