МОДЕЛИ ВРЕМЕННОГО РЯДА: AR(P), MA(Q), ARIMA(P,D,Q). ПРИМЕР ИССЛЕДОВАНИЯ ПОТРЕБЛЕНИЯ НЕФТЕПРОДУКТОВ ВО ФРАНЦИИ.

Доронина А.И.
Финансовый Университет при Правительстве Российской Федерации
Москва, Россия

TIME-SERIES MODELS. THE EXAMPLE OF OIL PRODUCTS CONSUMPTION IN FRANCE.

Doronina A. I.
Financial University under the Government of the Russian Federation
Moscow, Russia

Содержание

Введение

Темы производства и потребления нефтепродуктов становятся все более актуальными по всему миру на фоне проявляющейся угрозы дефицита этих ресурсов и отсутствия заменителей в широком использовании. Показатели взяты с сайта Национального Института Статистических и Экономических исследований (INSEE) Франции [1] . Информация представлена в денежном выражении, в биллионах евро. В данных учитывается ИПЦ, привязка идет к ценам 2005 года. Данные о потреблении взяты за период в 12 лет, а точнее с января 2000 года по август 2012 года. При этом частота наблюдения равна один месяц.

В данной работе будут построены и оценены модели временных рядов ARIMA (и, с учётом сезонной составляющей, SARIMA). Такие модели (интегрируемые модели авторегрессии и модели скользящего среднего) достаточно гибкие и могут описывать множество характеристик временных рядов. В модели авторегрессии каждое значение ряда находится в линейной зависимости от предыдущих значений. Модель скользящего среднего предполагает, что в ошибках модели в предшествующие периоды сосредоточена информация обо всей предыстории ряда. В зависимости от свойств изучаемого показателя, модели ARIMA могут включать в себя сразу обе модели, или каждую по отдельности.

В общем виде модель ARMA(p,q), где p – порядок авторегрессии, q – порядок скользящего среднего, выглядит следующим образом:

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \theta_1 y_{t-1} + \cdots + \theta_q y_{t-q}$$

Если процесс оказывается нестационарным и для приведения его к стационарному виду потребовалось взять несколько разностей, то модель становится ARIMA(p,d,q), где d – порядок разности.

[1] Данные взяты с сайта Национального Института Статистических и Экономических исследований http://www.bdm.insee.fr/bdm2/choixCriteres.action?request_locale=en&codeGroupe=1309

ШАГ 1. Описание исходных данных. Анализ динамики временного ряда.

Ниже графически представлена динамика потребления нефтепродуктов жителями Франции за период с 2000 по 2012 год (см рисунок 1). Отметим сразу, что тренд за весь период – нисходящий, он также отмечен на графике серой линией.

Наивысший уровень потребления нефтепродуктов приходится на февраль 2004. Меньше всего французы потребляли в 2011 году. Как правило, для зимних месяцев наблюдается высокое потребление, для летних – низкое, что легко объясняется погодными условиями.

Рисунок 1 Динамика потребления нефтепродуктов

Отметим, при анализе графика видно, что тренд для потребления нефтепродуктов нисходящий. Это легко объясняется тем, что во Франции активно идет переход к альтернативным источникам энергии.

Продолжим анализ и рассчитаем показатели динамики временного ряда:

- Темпы роста;
- Темпы прироста.

Подобные расчеты позволят провести сопоставление значений величины потребления за разные периоды. Стоит отметить, что расчет указанных показателей может быть произведен как базисным, так и цепным методом.

Были произведены соответствующие расчеты, ниже представлены их графические интерпретации (см рисунок 2-5).

Рисунок 2

Рисунок 3

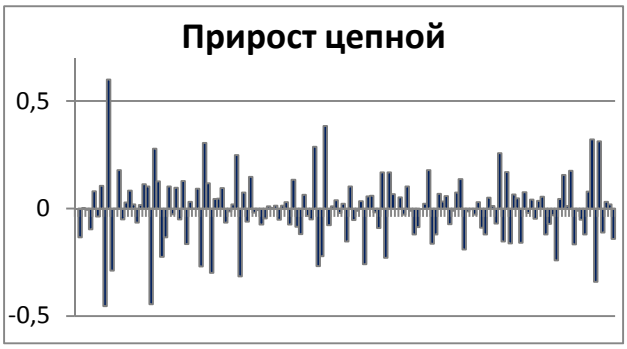


Рисунок4

Рисунок 5

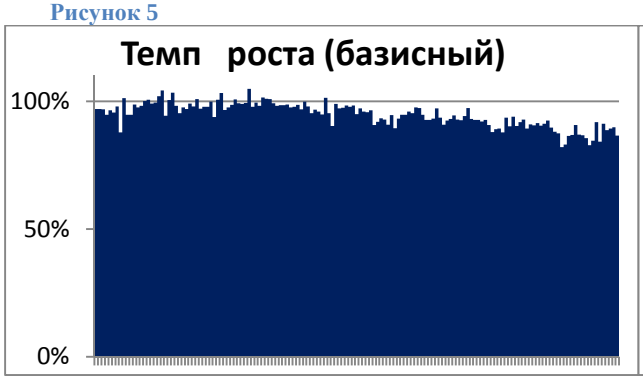


Рисунок 6

Рисунок 7

Отметим, что базисные показатели повторяют динамику переменной. Самый стабильный, а значит и полезный для анализа показатель «Темп роста (цепной)».

ШАГ 2. Исследование тенденции временного ряда

Приступая к следующему этапу, стоит сузить набор данных в силу практического удобства. Если до этого рассматривались помесячные данные за почти 12 лет, то теперь будет взять период с января 2004 по декабрь 2010.При изучении потребления нефтепродуктов такой период видится вполне приемлемым, особенно, если принять во внимание тот факт, что основные изменения в данной сфере начались сравнительно недавно.

На графике ниже изображена динамика потребления нефтепродуктов за выбранный отрезок времени (см рисунок 9).

Рисунок 8

При первоначальной визуальной оценке можно предположить, что наблюдается нисходящий тренд. Однако визуального анализа графика недостаточно для состоятельного вывода. Установить точную картину помогут тесты для проверки наличия тренда во временном ряду.

А именно:

- Метод Фостера-Стюарта;
- Критерий серий.

Применим для начала метод Фостера-Стюарта.

Нулевая гипотеза выглядит так: $H_0: M(y(t)) = a = const$ = > в динамике значений показателя тренд отсутствует. Рассчитаем специальные показатели: D=-13, $\sigma_D = 2{,}831$ (для 84 показателей). Согласно критерию Стьюдента при $\alpha = 0{,}05$ и $\nu = 84 - 1 = 83$, $t_{\text{кр}} = 1{,}998$. Значит, $t_{\text{набл}} = \frac{D}{\sigma_D} = -4{,}591$. Отсюда следует, что гипотеза об отсутствии тренда отклоняется с вероятность ошибки 0,05 и тренд в данных есть.

Критерий серий.

Нулевая гипотеза утверждает, что тренд в ряду отсутствует, если выполняются неравенства:

$$\begin{cases} \tau_{max}(n) < [3{,}3(\lg n + 1)] \\ \nu(n) > \left[\frac{1}{2}\left(n + 1 - 1{,}96\sqrt{n-1}\right)\right] \end{cases}$$

Где $\tau_{max}(n)$ – протяженность самой длинной серии, а $\upsilon(n)$- число серий повторяющихся знаков при сравнении медианы и значений ряда.

В нашем случае гипотеза об отсутствии отвергается, так как не выполняются оба неравенства. Значения параметров представлены в таблице №2:

Таблица 1

Параметр	Значение
Число серий	12
Самая долгая серия	23
t	17,9217
v	33,57178

Таким образом, согласно критерию серий в рассматриваемом временном ряду присутствует тренд.

Теперь стоит аналитически определить тренд. Для этого необходимо оценить существующие модели, такие как:

- Прямолинейная;
- Параболическая;
- Гиперболическая;
- Логарифмическая;
- Экспоненциальная.

В данной работе все модели не будут рассмотрены, для анализа используются только прямолинейная и полиномиальная модели. Прямолинейная модель – универсальная и проста в интерпретации, полиномиальная хороша для описания величин, попеременно возрастающих и убывающих.

Уравнение линейной модели:

$$y = -0{,}0048x + 4{,}4377$$ [2]

Уравнение полиномиальной модели:

$$y = 0{,}00003x2 - 0{,}0072x + 4{,}4721$$ [3]

Были рассчитаны параметры моделей, на основе которых производится их отбор, результаты представлены в таблице 3:

Таблица 2

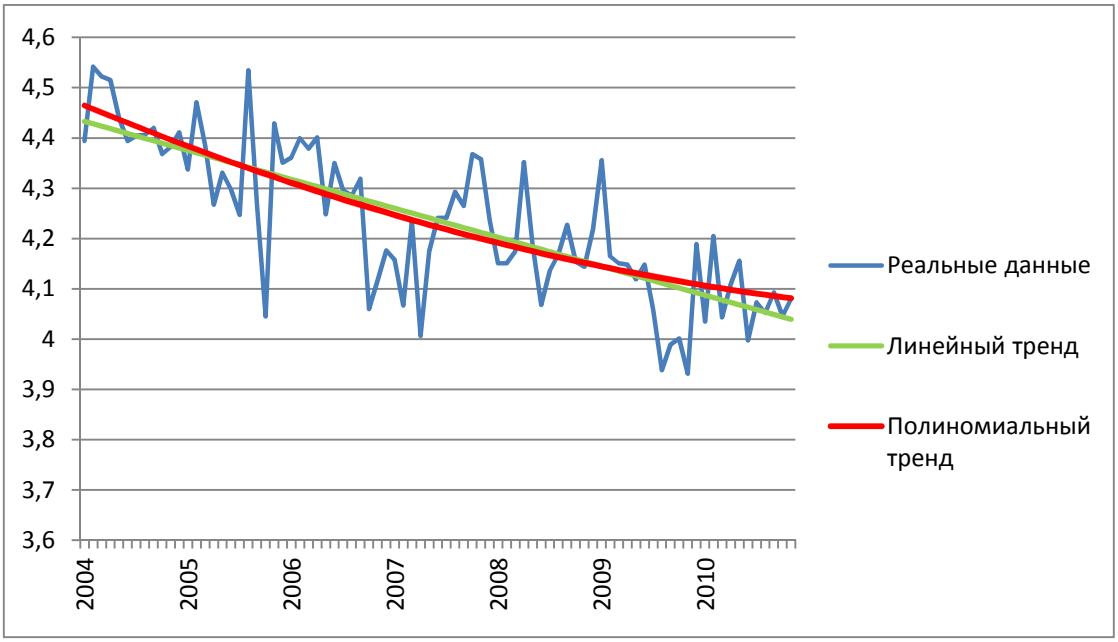
Параметры	Прямолинейная модель	Полиномиальная модель
k	2	2
S	0,09349	0,09253
R	0,6159	0,6258

[2] См. приложение 1
[3] См. приложение 1

Таким образом, получается, что полиномиальная модель – самая адекватная из выбранных, в ней ошибка S ниже, чем в линейной. В то же время коэффициент детерминации R^2 чуть выше, значит, в этой подели потребление в больше степени определяется временем.

Представим графически сравнение трендов (см рисунок 10).

Рисунок 9



ШАГ 3. Проверка на стационарность

Итак, в данных присутствуют нисходящий тренд и слабая сезонность. Это может быть объяснено климатическими условиями, а также тем, что Франция – одна из тех стран, которые начали переходить на альтернативные источники энергии. В ходе прошлого анализа было выявлено, что модель тренда – полиномиальная, и лучше всего описывает данные аддитивная модель, но заметим, что проверку на адекватность она не прошла.

Стационарный ряд – это ряд, чье поведение в настоящем и будущем совпадает с поведением в прошлом, т.е. на свойства не влияет изменение начала отсчёта времени. Определить, стационарен ли ряд, можно по виду автокорреляционной функции (ACF) и частной автокорреляционной функции (PACF) и путем проведения теста Дики-Фуллера.

Анализ автокорреляционной функции (ACF) и частной автокорреляционной функции (PACF)

Ниже представлены графики функций ACF и PACF (см рисунок 2 и 3). Красными пунктирынми линиями на графиках отмечен критический интервал $[-\frac{2}{\sqrt{n}}; \frac{2}{\sqrt{n}}]$, в пределах которого значения ACF и PACF считаются не отличающимися от нуля.

Рисунок 110. Автокорреляция (ACF)

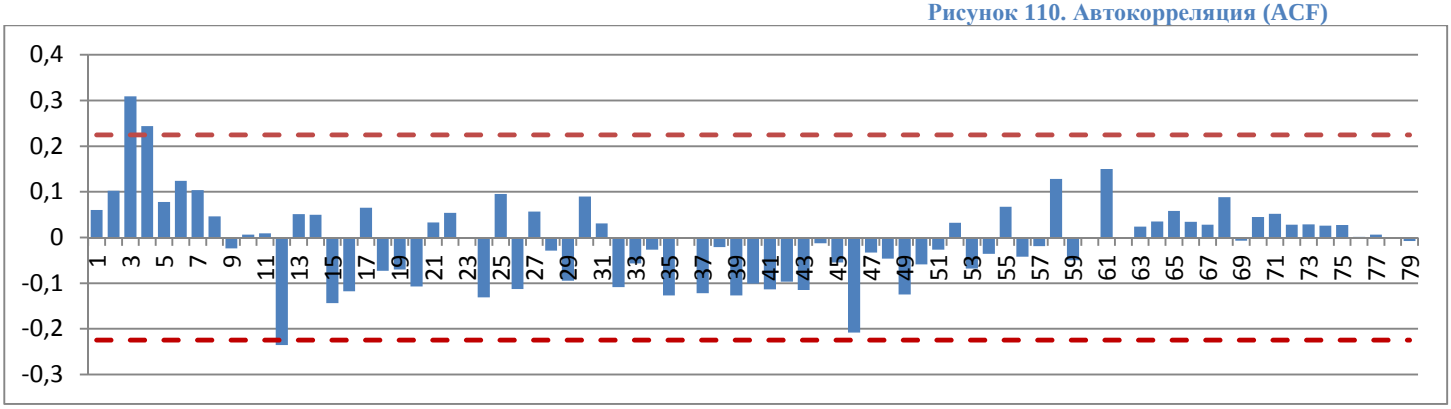
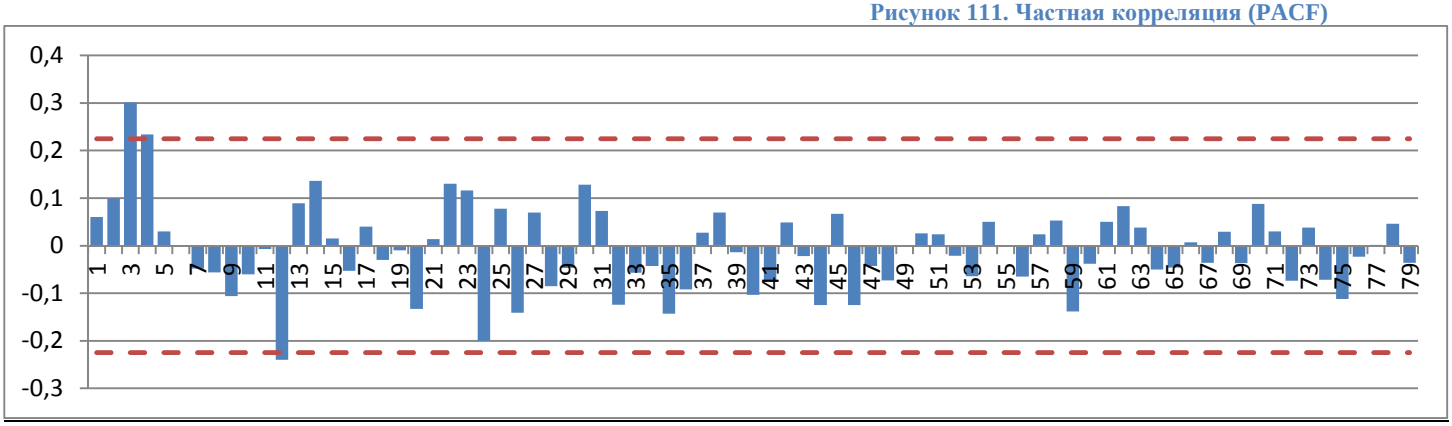


Рисунок 111. Частная корреляция (PACF)



При визуальном анализе графиков видно, что ряд не является стационарным. Автокорреляция не стабильна, имеются выбросы. Графики отражают сезонность, но она довольно слабая.

Тест Дики-Фулера

Суть Дики-Фуллера состоит в том, что необходимо проверять нулевую гипотезу о наличии единичного корня в уравнении:

$$\mathbf{y_t = \alpha y_{t-1} + \varepsilon_t}$$

Есть альтернативная гипотеза: $\boldsymbol{\alpha<1}$. Взяв первую разность, можно получить следующее уравнение:

$$\mathbf{y_t - y_{t-1} = (\alpha - 1) y_{t-1} + \varepsilon_t}$$

Или

$$\mathbf{\Delta y_t = \beta y_{t-1} + \varepsilon_t}$$

Тогда

$$\mathbf{H_0: \beta=0,\ H_1: \beta<0}$$

После нахождения оценки $\widehat{\boldsymbol{\beta}}$, вычисляют статистику $\mathbf{t_{набл} = \widehat{\beta}/S(\widehat{\beta})}$. Гипотеза принимается и ряд признается нестационарным, если $\mathbf{t_{набл} > t_{крит}}$.

Для исследуемых данных $\mathbf{t_{набе}=-3.179024}$, что означает, что ряд является стационарным на уровне значимости 5% (см приложение). Результат теста не сходится с визуальным анализом, однако тест предпочтительнее.

ШАГ 4. Оценка моделей ARIMA

Вид модели ARIMA определяется по коррелограмме. В данном случае затухающий график ACF и выбросы на первых нескольких лагах PACF свидетельствуют о том, что это модель авторегрессии. Также выброс на двенадцатом лаге в PACF показывает наличие 12-звенной сезонности, а значит, необходимо оценить модель SARIMA(p,d,q,)(Ps,Ds,Qs), где p – порядок авторегрессии, d – порядок разности, q – порядок скользящего среднего, P_s – порядок сезонной авторегрессии, D_s – порядок сезонной разности, Q_s – сезонный параметр скользящего среднего.

Составление модели происходило на основе анализа графиков (рисунки 12 и 13) и сравнивания возможных моделей. Так была оценена модель SARMA(4;1;4)(12;0;0). При добавлении константы и сезонной компоненты адекватность модели снижалась (рос коэффициент Шварца).

Таблица 3. Параметры модели ARMA(4,0)

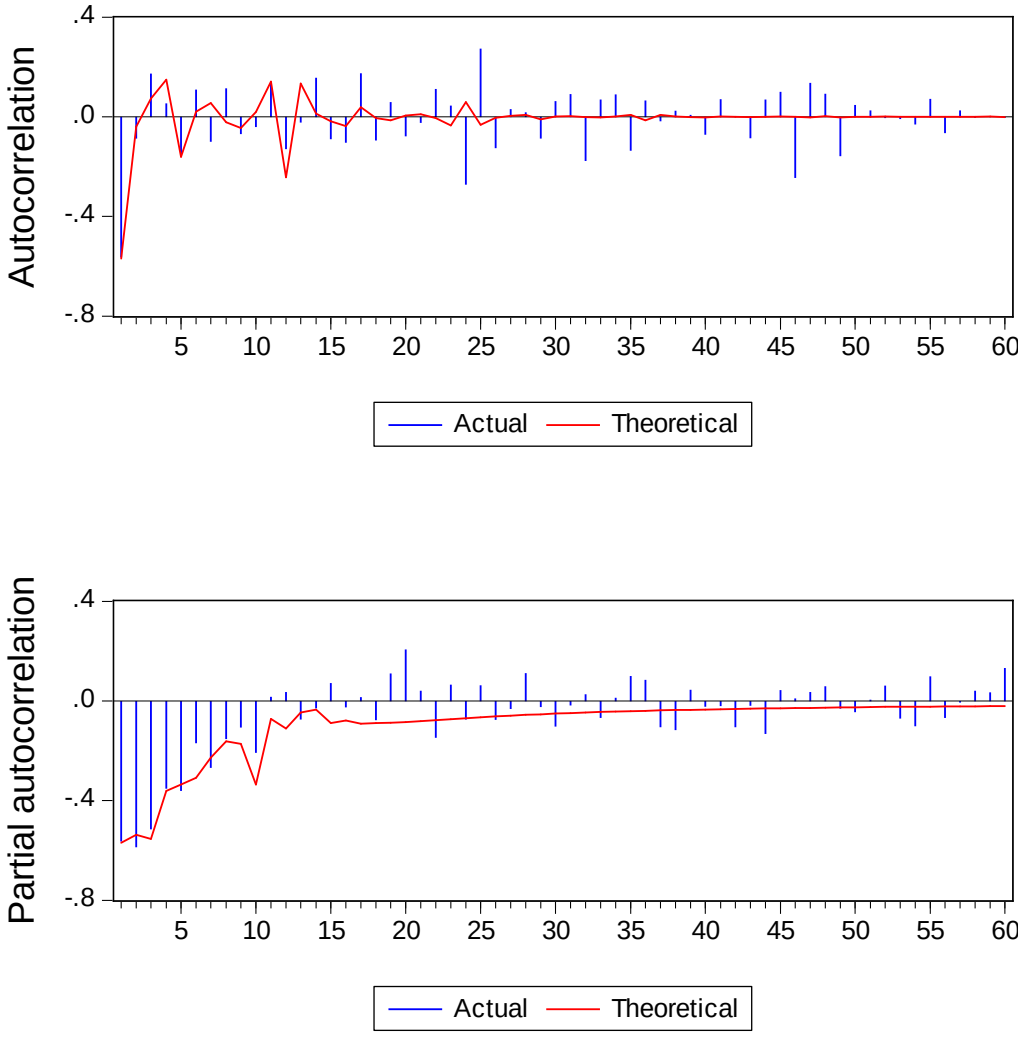
Модель SARMA (4,4)(12,0,0) **Зависимая переменная: потребление нефтепродуктов во Франции с 2004 по 2010 годы**				
Переменная	**Коэффициент**	**Стд Ошибка**	**t-статисика**	**Вероятость**
AR(1)	0.178706	0.124346	1.437173	0.1563
AR(2)	1.019413	0.118253	8.620642	0.0000
AR(3)	0.237542	0.125961	1.885836	0.0646
AR(4)	-0.436016	0.125345	-3.478534	0.0010
SAR(12)	-0.202323	0.102127	-1.981099	0.0526
MA(1)	-0.153507	0.032221	-4.764114	0.0000
MA(2)	-1.262086	0.037290	-33.84479	0.0000
MA(3)	-0.140595	0.027007	-5.205820	0.0000
MA(4)	0.935625	0.025021	37.39394	0.0000
Критерий Шварца **-1,439**				

В таблице выше (см таблицу 1) видно, что все коэффициенты значимы на уровне 0,1 или 10%.

ШАГ 5. Проверка модели на адекватность

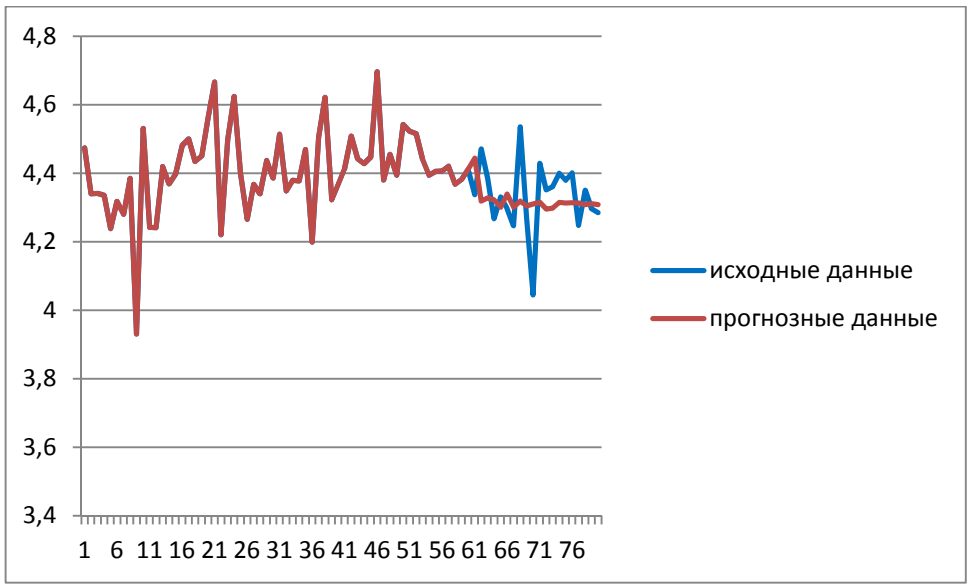
Рассмотрим коррелограммы, отражающие поведение теоретической автокорреляционной и частной автокорреляционной функций (см рисунок 4)

Рисунок 12. Коррелограммы, теоретическое и действительное значения



На рисунке видно, теоретическая функция автокорреляции сильно отличается от данной. В то время как графики фактической и теоретической частной корреляции довольно похожи.

Рисунок 13. Прогноз



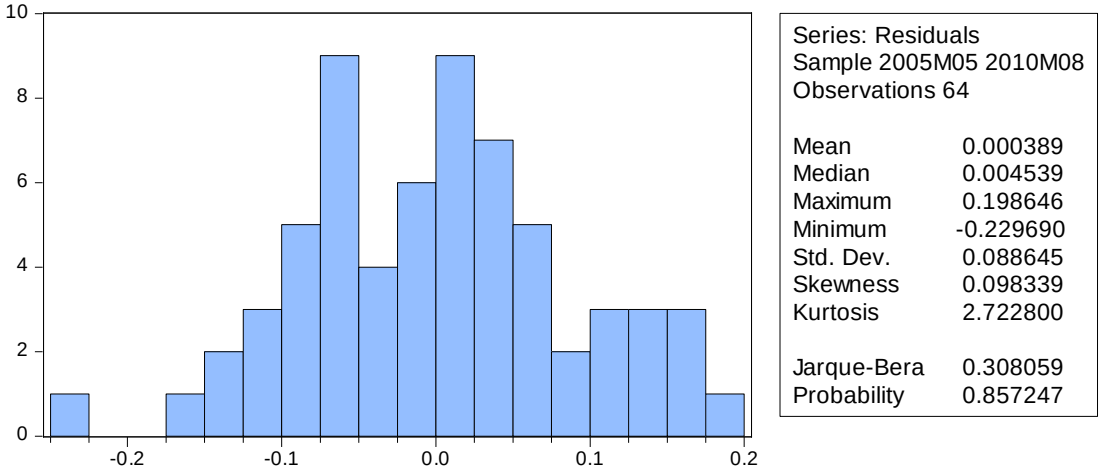
На графике выше (см рисунок 5) изображены исходные данные и прогнозные значения на основе выбранной модели SARMA(4;1;4)(12;0;0). Как можно заметить прогнозу свойственна низкая точность.

ШАГ 6. Тест Жарка-Бера

Далее проведем тест Жарка-Бера:

$$\boldsymbol{JB = \frac{n}{6}\left(A_s^2 + \frac{(K_x - 3)^2}{4}\right) \sim \chi^2(2);\ A_s = \frac{\sum_{i=1}^{n} e_i^3}{n\sigma_s^3};\ K_s = \frac{\sum_{i=1}^{n} e_i^4}{n\sigma_s^4}}$$

Для этого также построим гистограмму по имеющимся данным.

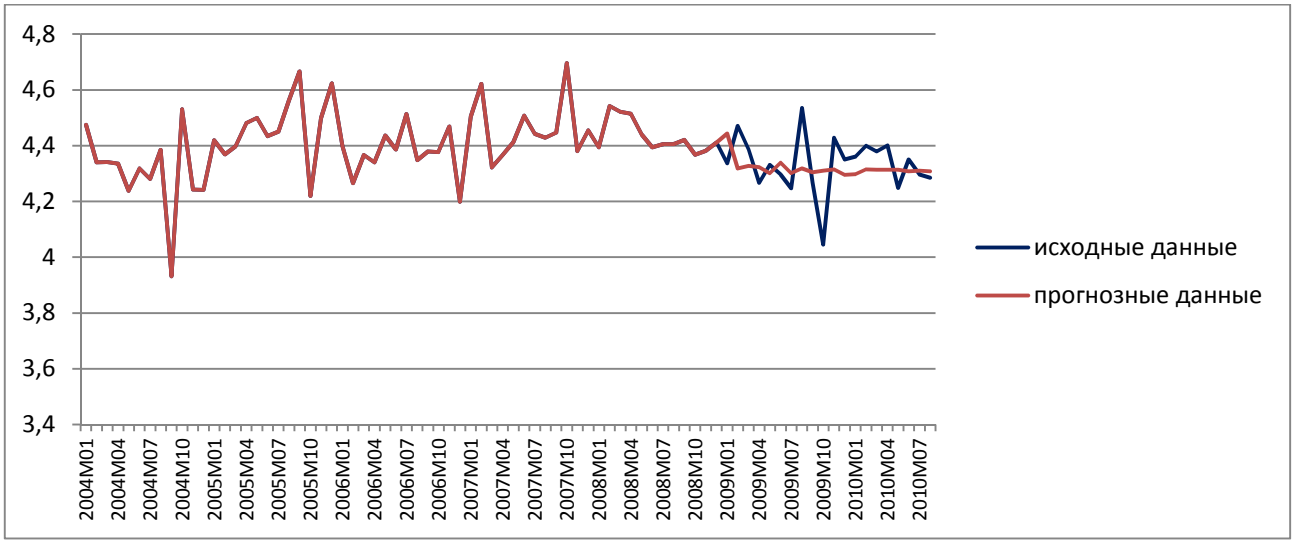


В данном случае $\boldsymbol{JB = 0,308 > \chi^2(0,05;2) = 5,99}$.Значит, гипотеза о нормальности остатков не отвергается. Однако судя по графику, утверждение о нормальности данных является спорным.

Вывод

Таким образом, в ходе анализа была построена модель SARMA(4;1;4)(12;0;0). Были выявлены стационарность и сравнительно наибольшая эффективность. Однако построенный прогноз не показал себя довольно эффективным. В таком случае в заключение построим прогноз на 2010 год (см рисунок 6)

Рисунок 14. Прогноз на год

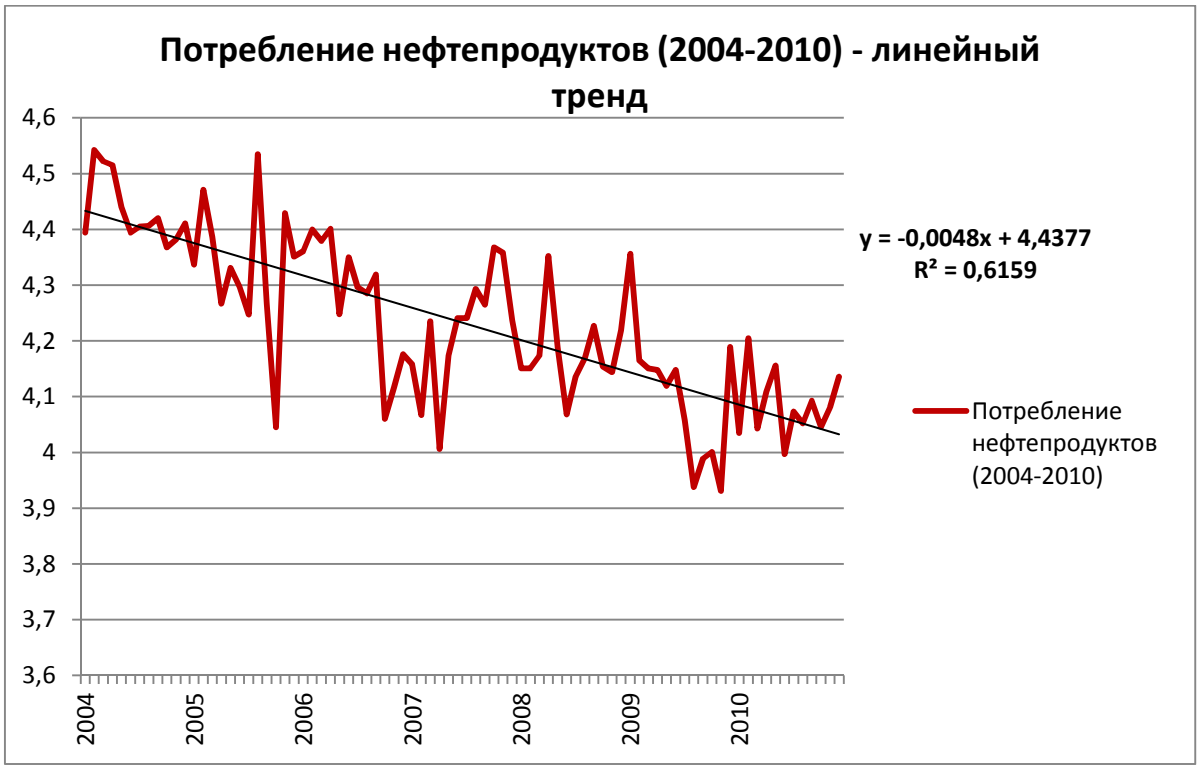


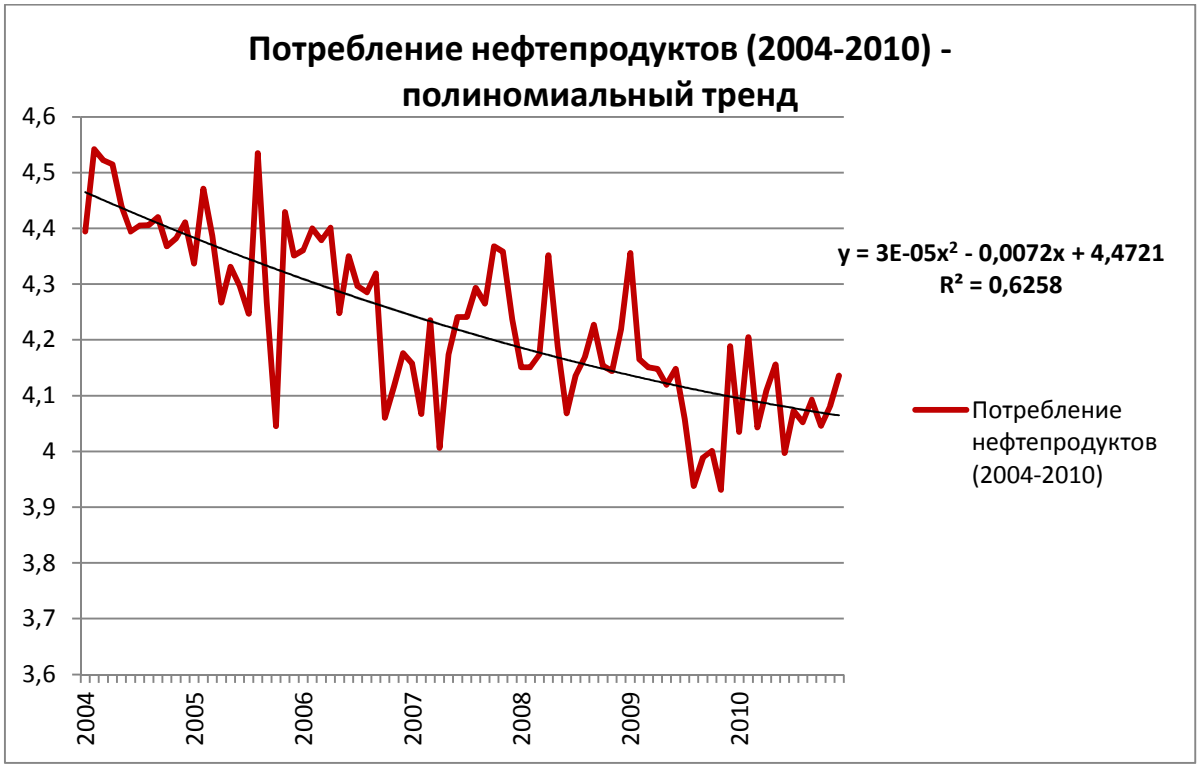
В данном случае прогноз не отразил резкости изменения переменной, однако, направление колебаний совпадает с исходными данными. Также близки по величине значения на конец периода, значит, тренд отражен адекватно.

Список литературы:

1. Магнус Я.Р. Эконометрика: Начальный курс: Учебное пособие/ Я.Р.Магнус, П.К. Катышев, А.А.Пересецкий. - М.: Дело, 2005. - 503с.
2. Айвазян С.А. Методы эконометрики: учебник. – М.: Магистр: ИНФРА-М, 2010.
3. Айвазян С.А. Прикладная статистика. Основы эконометрики. Изд. 2 – е. Т. 2 – М.: ЮНИТИ,2001.
4. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: учебное пособие, Вузовский учебник, 2007.
5. Сайта Национального Института Статистических и Экономических исследований http://www.bdm.insee.fr/bdm2/choixCriteres.action?request_locale=en&codeGroupe=1309

Приложение 1





Приложение 2

Null Hypothesis: Y has a unit root
Exogenous: Constant
Lag Length: 2 (Fixed)

		t-Statistic	Prob.*
Augmented **Dickey-Fuller** test statistic		-3.179024	0.0251
Test critical values:	1% level	-3.517847	
	5% level	-2.899619	
	10% level	-2.587134	

*MacKinnon (1996) one-sided p-values.

Приложение 3

Dependent Variable: D(Y)
Method: Least Squares
Date: 12/19/12 Time: 22:54
Sample (adjusted): 2005M06 2010M08
Included observations: 63 after adjustments
Convergence achieved after 18 iterations
MA Backcast: 2005M02 2005M05

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.162291	0.414500	-0.391535	0.6969
AR(2)	0.369723	0.419001	0.882392	0.3815
AR(3)	0.348843	0.343531	1.015464	0.3144
AR(4)	0.324409	0.172139	1.884580	0.0649
SAR(12)	-0.276359	0.115088	-2.401289	0.0198
MA(1)	-0.770015	0.428270	-1.797968	0.0778
MA(2)	-0.678956	0.706697	-0.960745	0.3410
MA(3)	0.252731	0.616332	0.410057	0.6834
MA(4)	0.196282	0.348860	0.562639	0.5760

R-squared	0.587934	Mean dependent var	-0.003413
Adjusted R-squared	0.526888	S.D. dependent var	0.150613
S.E. of regression	0.103597	Akaike info criterion	-1.565059
Sum squared resid	0.579543	Schwarz criterion	-1.258897
Log likelihood	58.29936	Hannan-Quinn criter.	-1.444644
Durbin-Watson stat	2.032115		

Dependent Variable: D(Y)
Method: Least Squares
Date: 12/19/12 Time: 22:57
Sample (adjusted): 2005M06 2010M08
Included observations: 63 after adjustments
Convergence achieved after 141 iterations
MA Backcast: OFF (Roots of MA process too large)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.001316	0.002432	-0.540920	0.5908
AR(1)	-0.526297	1.077917	-0.488253	0.6274
AR(2)	-0.264593	0.535161	-0.494418	0.6231
AR(3)	0.010428	0.530909	0.019642	0.9844
AR(4)	0.144801	0.241549	0.599468	0.5514
SAR(12)	-0.248870	0.117841	-2.111914	0.0394
MA(1)	-0.618607	1.101765	-0.561469	0.5768
MA(2)	-0.497375	1.539291	-0.323119	0.7479
MA(3)	-0.193063	0.992726	-0.194478	0.8465
MA(4)	-0.097579	0.625944	-0.155890	0.8767

R-squared	0.697062	Mean dependent var	-0.003413
Adjusted R-squared	0.645619	S.D. dependent var	0.150613
S.E. of regression	0.089660	Akaike info criterion	-1.840967
Sum squared resid	0.426062	Schwarz criterion	-1.500787
Log likelihood	67.99046	Hannan-Quinn criter.	-1.707172
F-statistic	13.55035	Durbin-Watson stat	2.307048
Prob(F-statistic)	0.000000		

Dependent Variable: D(Y,2)
Method: Least Squares
Date: 12/19/12 Time: 22:58
Sample (adjusted): 2005M07 2010M08
Included observations: 62 after adjustments
Convergence achieved after 19 iterations
MA Backcast: 2005M03 2005M06

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-1.119256	0.905085	-1.236631	0.2217
AR(2)	-0.788885	1.176637	-0.670458	0.5055
AR(3)	-0.451115	0.470531	-0.958736	0.3420
AR(4)	-0.089890	0.367086	-0.244873	0.8075
SAR(12)	-0.243165	0.112813	-2.155465	0.0357
MA(1)	-0.842579	0.904598	-0.931441	0.3558
MA(2)	-0.570918	0.747516	-0.763754	0.4484
MA(3)	0.226798	1.190412	0.190521	0.8496
MA(4)	0.186924	0.806922	0.231650	0.8177

R-squared	0.854610	Mean dependent var	0.000887
Adjusted R-squared	0.832665	S.D. dependent var	0.256572
S.E. of regression	0.104955	Akaike info criterion	-1.537091
Sum squared resid	0.583823	Schwarz criterion	-1.228314
Log likelihood	56.64983	Hannan-Quinn criter.	-1.415858
Durbin-Watson stat	1.999077		