Текущее домашнее задание (ТДЗ) 12. «Сезонность 1»

БПИ227. Артемьев Александр

1.1) В первом задании проведем анализ периодограммы для исходных данных - числа заключённых браков в России, тыс. Первым шагом заполним следующую вспомогательную таблицу:

i	fi	wi	ai	bi	I(f)	p

Где столбцы обозначают следующее:
i - номер гармоники ($i \in \overline{1,q}, q-$ кол-во гармоник $= \frac{N}{2} = 7$)
f_i - частота(число повторений циклов в единицу времени $f_i = \frac{i}{N}$)
w_i - круговая частота(частоты в полярных координатах $w_i = 2\pi f_i$)
a_i, b_i- амплитуды компонент ($a_i = \frac{2}{N}\sum_{t=1}^{N} y_t cos(w_i t); b_i = \frac{2}{N}\sum_{t=1}^{N} y_t sin(w_i t)$)
$I(f_i)$ - интенсивность($I(f_i) = \frac{N}{2}(a_i^2 + b_i^2)$)
p - период(минимальный интервал времени, необходимый для того, чтобы значения ВР начали повторяться $p_i f_i = 1$)

Теперь, обладая нужными знаниями и формулами нахождения метрик, заполним пустые ячейки и получим следующую таблицу:

i	fi	wi	ai	bi	I(f)	p
1	0,07	0,45	-27,44	-21,18	8409,55	14
2	0,14	0,90	7,33	17,65	2557,88	7
3	0,21	1,35	-10,14	-5,81	955,60	4,67
4	0,29	1,80	3,60	-3,14	159,71	3,5
5	0,36	2,24	3,54	9,69	744,65	2,80
6	0,43	2,69	-0,78	-1,90	29,62	2,33
7	0,50	3,14	3,49	0,00	85,39	2

1.2) Вспомогательные расчёты для каждой гармоники расположены в Excel файле HW12, а здесь

приведём подробные расчёты для i = 1.

Собственно первая гармоника имеет номер $i = 1 \Longrightarrow f_1 = \frac{1}{N} = \frac{1}{14} = 0,07$. Круговая частота $= 2 * \pi * \frac{1}{14} = \frac{\pi}{7} = 0,45$. Теперь рассчитаем a_1 и b_1.

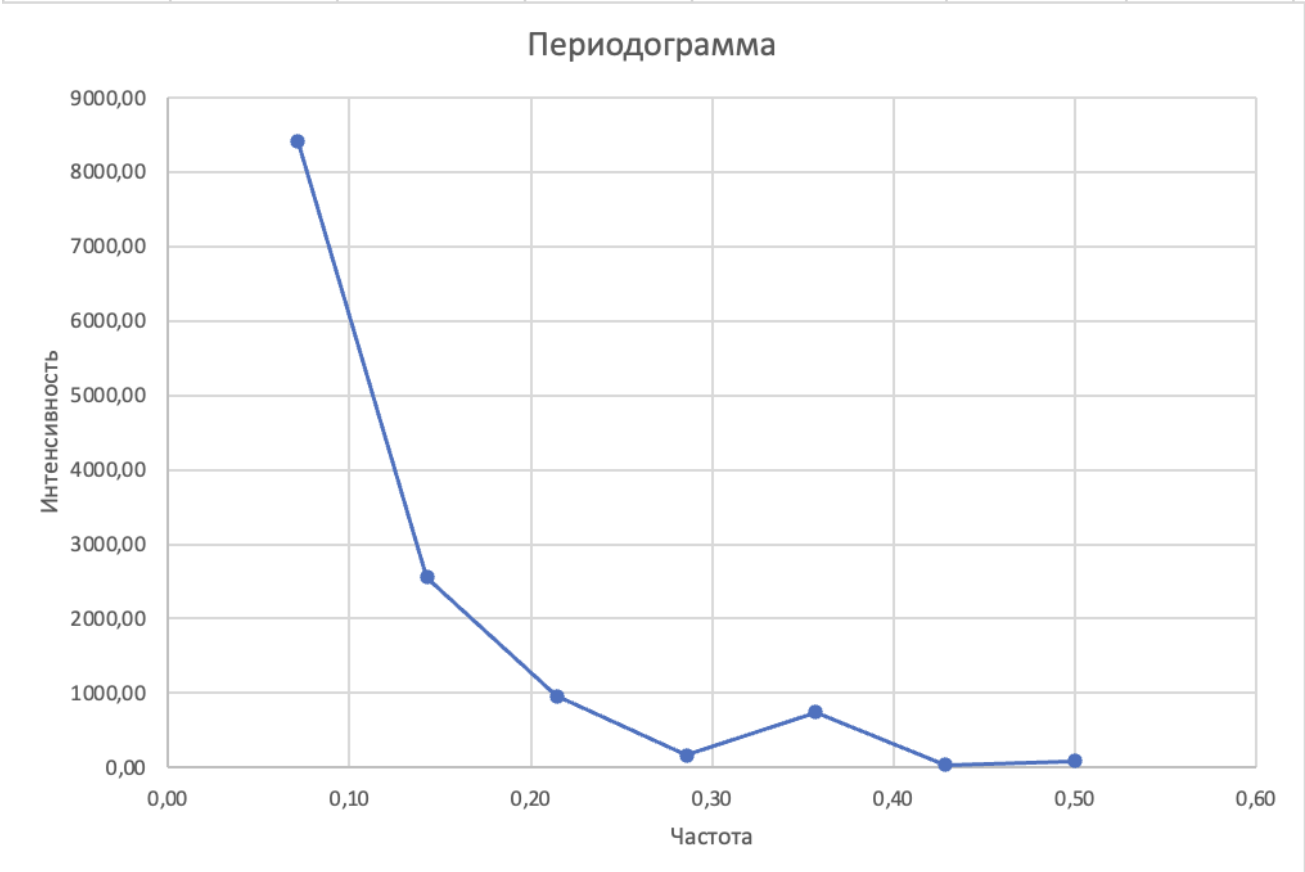
$a_1 = \frac{2}{14}(60,777 * cos(0,45 * 1) + 63,76 * cos(0,45 * 2) + 78,424 * cos(0,45 * 3) + 58,576 * cos(0,45 * 4) + 37,662 * cos(0,45 * 5) + 102,812 * cos(0,45 * 6) + 127,025 * cos(0,45 * 7) + 137,191 * cos(0,45 * 8) + 134,834 * cos(0,45 * 9) + 98,916 * cos(0,45 * 10) + 85,707 * cos(0,45 * 11) + 80,682 * cos(0,45 * 12) + 55,509 * cos(0,45 * 13) + 62,449 * cos(0,45 * 14)) = -27,43646$

$a_1 = \frac{2}{14}(60,777 * sin(0,45 * 1) + 63,76 * sin(0,45 * 2) + 78,424 * sin(0,45 * 3) + 58,576 * sin(0,45 * 4) + 37,662 * sin(0,45 * 5) + 102,812 * sin(0,45 * 6) + 127,025 * sin(0,45 * 7) + 137,191 * sin(0,45 * 8) + 134,834 * sin(0,45 * 9) + 98,916 * sin(0,45 * 10) + 85,707 * sin(0,45 * 11) + 80,682 * sin(0,45 * 12) + 55,509 * sin(0,45 * 13) + 62,449 * sin(0,45 * 14)) = -21,18029.$

Следовательно,, интенсивность $I(f_1) = 7 * (27,43646^2 + 21,18029^2) = 8409,55$

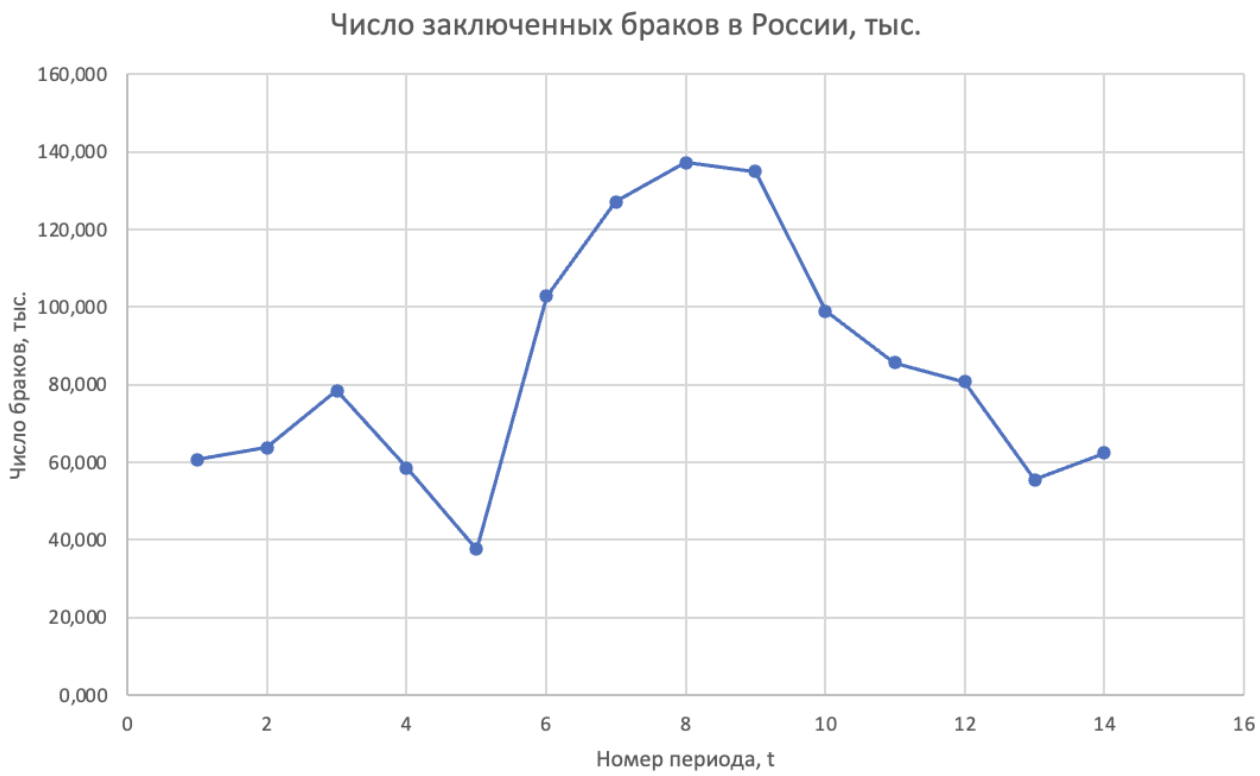
Осталось посчтать период, который равен $\frac{1}{f_1} = \frac{N}{1} = 14$

1.3) Последним шагом построим периодограмму



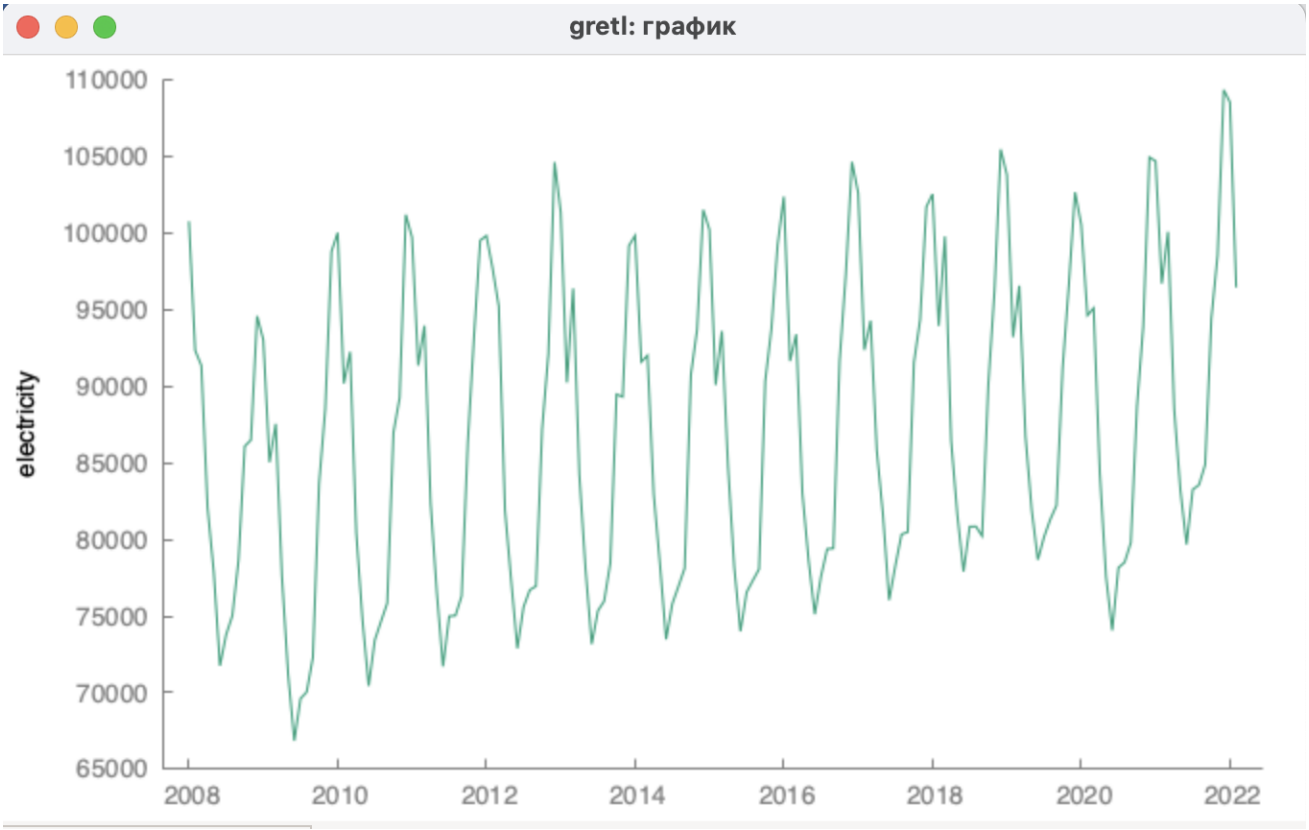
Из анализа построенной периодограммы мы можем сделать следующий вывод: на графике присутствует ярко выраженный пик с наибольшей интенсивностью(8409,55), который соответствует

первой гармонике с частотой $\frac{1}{14}$. Следовательно у нашего ряда присутствует сезонность с периодом p = 14(период для интенсивности пика). То есть сезонность в ряду составляет весь период наблюдений.(Здесь хочется добавить, что максимальная мощность достигается при низкой частоте, а также наблюдается тренд уменьшения интенсивности при увеличении f_i. Возможно в данных есть сильный тренд, который зашумляет сезонность, и на самом деле сезонность слабая. Поэтому посмотрим на график ряда ниже)



В итоге, по графику видно, что сильного тренда нет, и мы можем вернутся к заключению о сезонности с периодом p = 14.

2.1) Теперь перейдем к исследованию следующих предоставленных данных, а именно - Объем потребления электроэнергии в РФ (тыс. МВт·ч)) (2008-2022) . Для начала опишем исходные данные:



Показатель: Объём потребления электроэнергии в Российской Федерации, измеренный в тысячах мегаватт-часов (тыс. МВт·ч).

Период наблюдений: с января 2008 года по февраль 2022 года.

Представлено 170 наблюдений, помесячные данные из рассматриваемого периода.

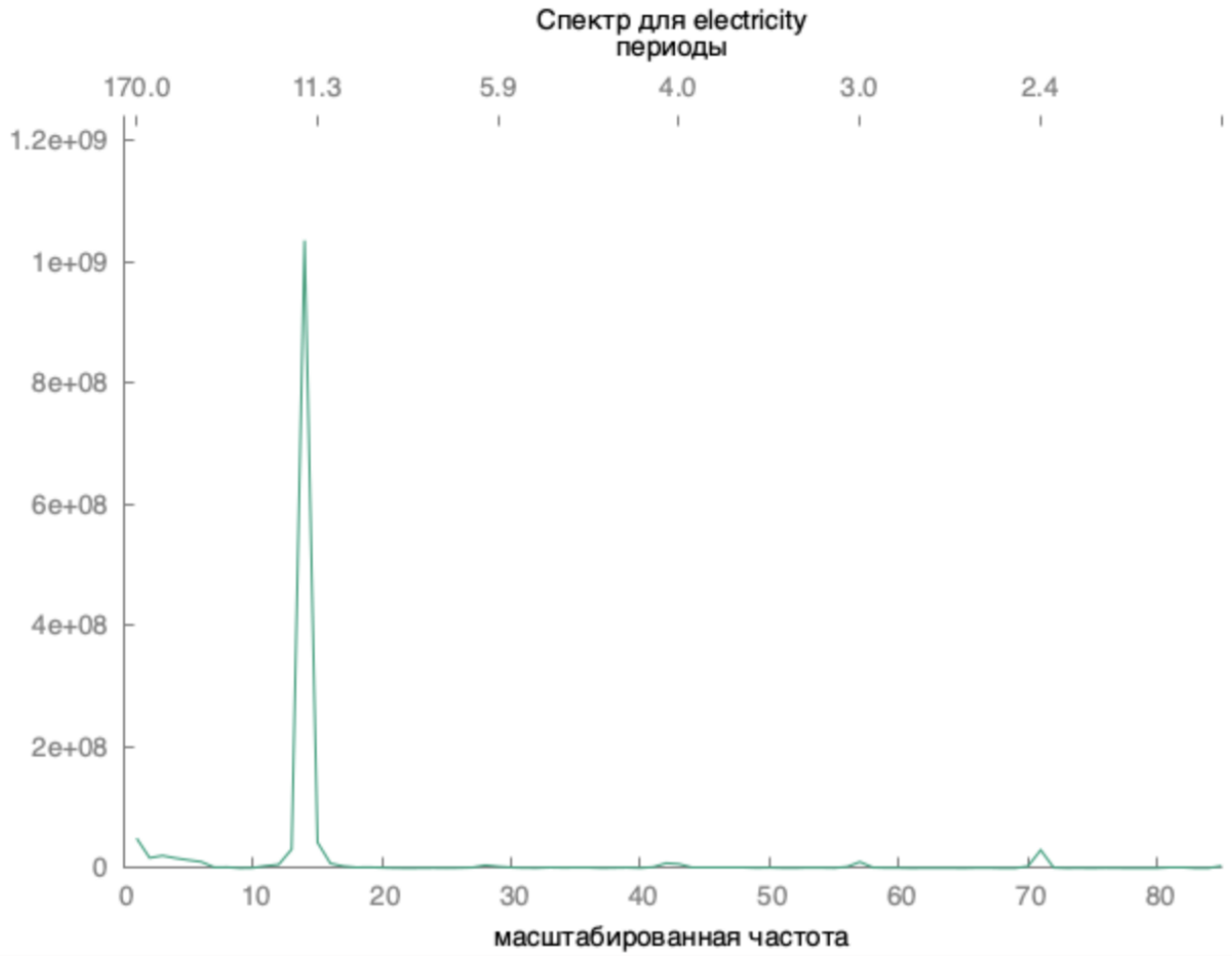
Чётко выраженная сезонность: почти каждый год наблюдается похожая структура — рост и падение потребления.

Наличие слабого тренда вверх: общее потребление электроэнергии за годы постепенно увеличивается (хотя есть некоторые падения в отдельных годах).

Небольшие всплески и аномалии(например, около 2009 и 2020 годов, что вероятно связано с кризисом 2008 года и пандемией 2020 года).

2.2) Следующий шаг - исследование данных на сезонность.

Построим и интерпретируем периодограмму:

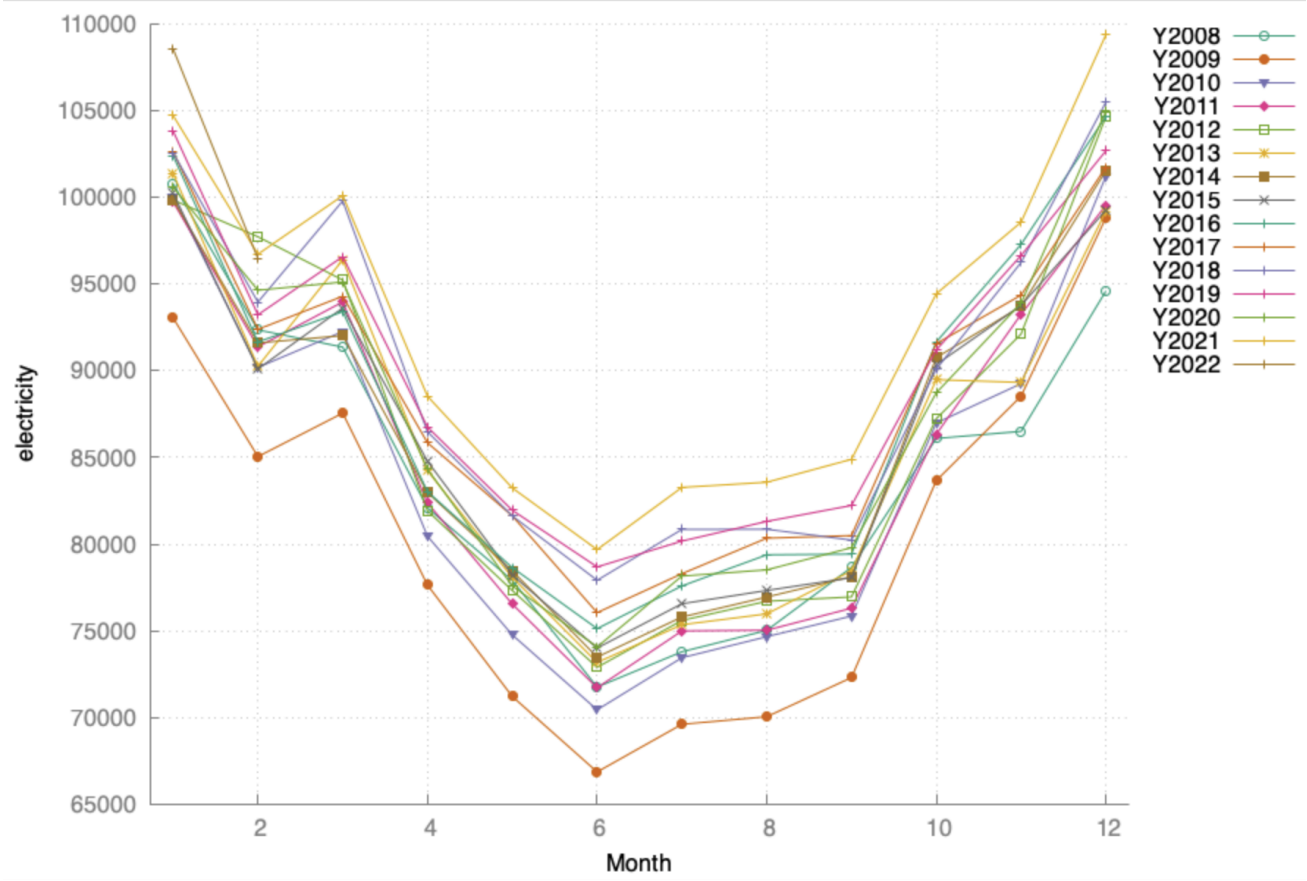


На графике периодограммы присутствует сильный пик, с наибольшей интенсивностью. Чтобы определить период сезонности, обратимся к значениям масштабированной частоты в интервале от 10 до 20(где и находится наш пик):

Периодограмма для electricity
Количество наблюдений = 170

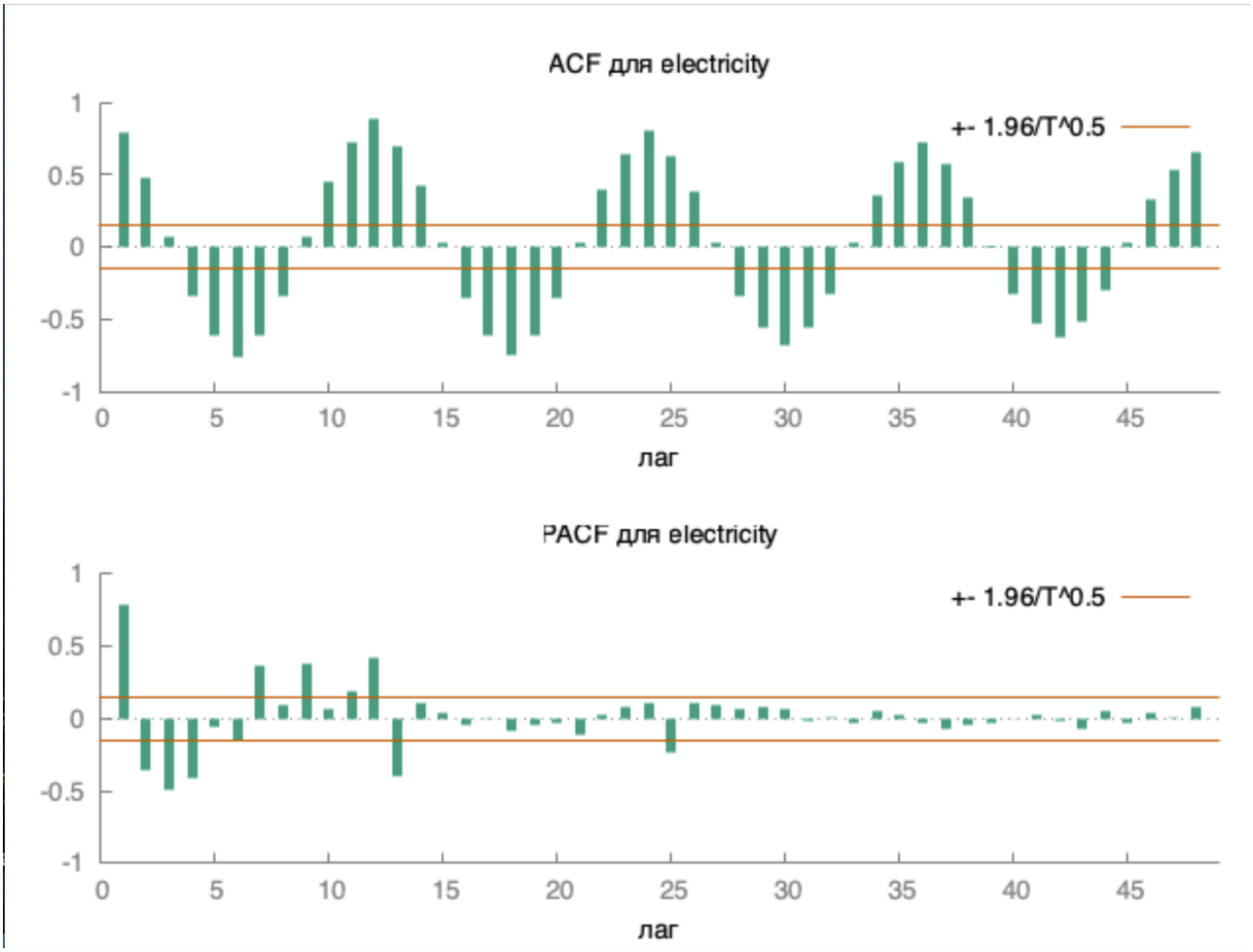
Омега	Масштаб. Частота	Периоды	Спектрал. частота
0.03696	1	170.00	4.9071e+07
0.07392	2	85.00	1.6871e+07
0.11088	3	56.67	2.0471e+07
0.14784	4	42.50	1.6479e+07
0.18480	5	34.00	1.3273e+07
0.22176	6	28.33	1.0365e+07
0.25872	7	24.29	1.3142e+06
0.29568	8	21.25	1.5330e+06
0.33264	9	18.89	68166
0.36960	10	17.00	4.2717e+05
0.40656	11	15.45	3.5593e+06
0.44352	12	14.17	5.9259e+06
0.48048	13	13.08	3.1065e+07
0.51744	14	12.14	1.0351e+09
0.55440	15	11.33	4.2433e+07
0.59136	16	10.62	7.7807e+06
0.62832	17	10.00	3.0201e+06
0.66528	18	9.44	1.1581e+06
0.70224	19	8.95	1.4457e+06
0.73920	20	8.50	7.7436e+05

Таким образом Спектрал. частота достигается на 14 масштаб. частоте со значением 1.0351e+09, причем период = 12.14. То есть для нашего временного ряда наблюдается годовая сезонность(12 месяцев). Теперь используем еще один способ, как можно убедиться в наличии сезонности - используя график сезонной волны:



Сезонная волна наглядно показывает устойчивый годовой цикл в потреблении электроэнергии. Во всех годах регулярно повторяется чёткая структура с пиковой нагрузкой в холодный период и минимумом в тёплые месяцы, что подтверждает ярко выраженную годовую сезонность. Сравнивая рассматриваемые года, видно, что амплитуда сезонных колебаний постепенно растёт на фоне слабо выраженного восходящего тренда: как зимние максимумы становятся выше, так и летние минимумы тоже слегка поднимаются. Отдельные годы (например, 2008 и 2020) заметно отклоняются вниз от общей картины, что, вероятно, связано с экономическим кризисом и пандемией.

Следующий способ исследования на сезонность - анализ коррелограммы:



На графиках ACF/PACF отчетливо проявляется годовая сезонность. Это следует из периодических всплесков автокорреляции на лагах 12, 24, 36, 48. Также между ними меняется знак (отрицательные промежуточные пики около 6, 18, 30 и т.д.)

2.3) Теперь смоделируем сезонность используя фиктивные переменные, предполагая наличие линейного тренда. Мы строим 11 фиктивных переменных d_i, для которых будет выполняться:

$$d_i = \begin{cases} 1, \text{ если наблюдение принадлежит i месяцу} \\ 0, \text{ иначе} \end{cases}$$

Значения i пробегают от 2 до 12. Таким образом наша модель будет иметь вид $Y_t = a+bt+\sum_{i=2}^{12} c_i d_i$, где для i месяца($i \in \overline{2,12}$) $Y_t = a + bt + c_i d_i$ и для 1-ого месяца $Y_t = a + bt$. Одна переменная специально исключена, чтобы избежать дополнительной мультиколинеарности. Получили следующую модель:

gretl: модель 1

Файл Правка Тесты Сохранить Графики Анализ LaTeX

```
Модель 1: МНК, использованы наблюдения 2008:01-2022:02 (T = 170)
Зависимая переменная: electricity

             коэффициент   ст. ошибка   t-статистика   p-значение
  ---------------------------------------------------------------
  const       96958.9      568.364        170.6        3.81e-180  ***
  dm2         -8884.20     714.528        -12.43       3.84e-25   ***
  dm3         -6731.95     727.272         -9.256      1.50e-16   ***
  dm4        -17523.5      727.227        -24.10       1.28e-54   ***
  dm5        -22879.7      727.195        -31.46       1.10e-69   ***
  dm6        -27309.5      727.176        -37.56       2.41e-80   ***
  dm7        -24673.2      727.169        -33.93       3.53e-74   ***
  dm8        -23852.4      727.176        -32.80       3.73e-72   ***
  dm9        -22754.6      727.195        -31.29       2.30e-69   ***
  dm10       -12317.5      727.227        -16.94       2.91e-37   ***
  dm11        -8448.85     727.272        -11.62       6.58e-23   ***
  dm12          375.061    727.330          0.5157     0.6068
  time           51.7004     3.06259       16.88       4.09e-37   ***

Среднее завис. перемен  86915.26   Ст. откл. завис. перем  10033.55
Сумма кв. остатков      6.01e+08   Ст. ошибка модели        1956.799
R-квадрат               0.964666   Исправ. R-квадрат        0.961965
F(12, 157)              357.1900   P-значение (F)           2.7e-107
Лог. правдоподобие     -1522.899   Крит. Акаике             3071.797
Крит. Шварца            3112.563   Крит. Хеннана-Куинна     3088.339
параметр rho            0.634505   Стат. Дарбина-Уотсона    0.707261
обратите внимание на сокращенные обозначения статистики

Исключая константу, наибольшее p-значение получено для переменной 12 (dm12)
```

Анализируя метрики построенной модели, стоит отметить:

Все фиктивные переменные получились значимыми, кроме 12-ой дамми-переменной для декабря. dm12 незначима (p=0.61): уровень в декабре близок к январскому.

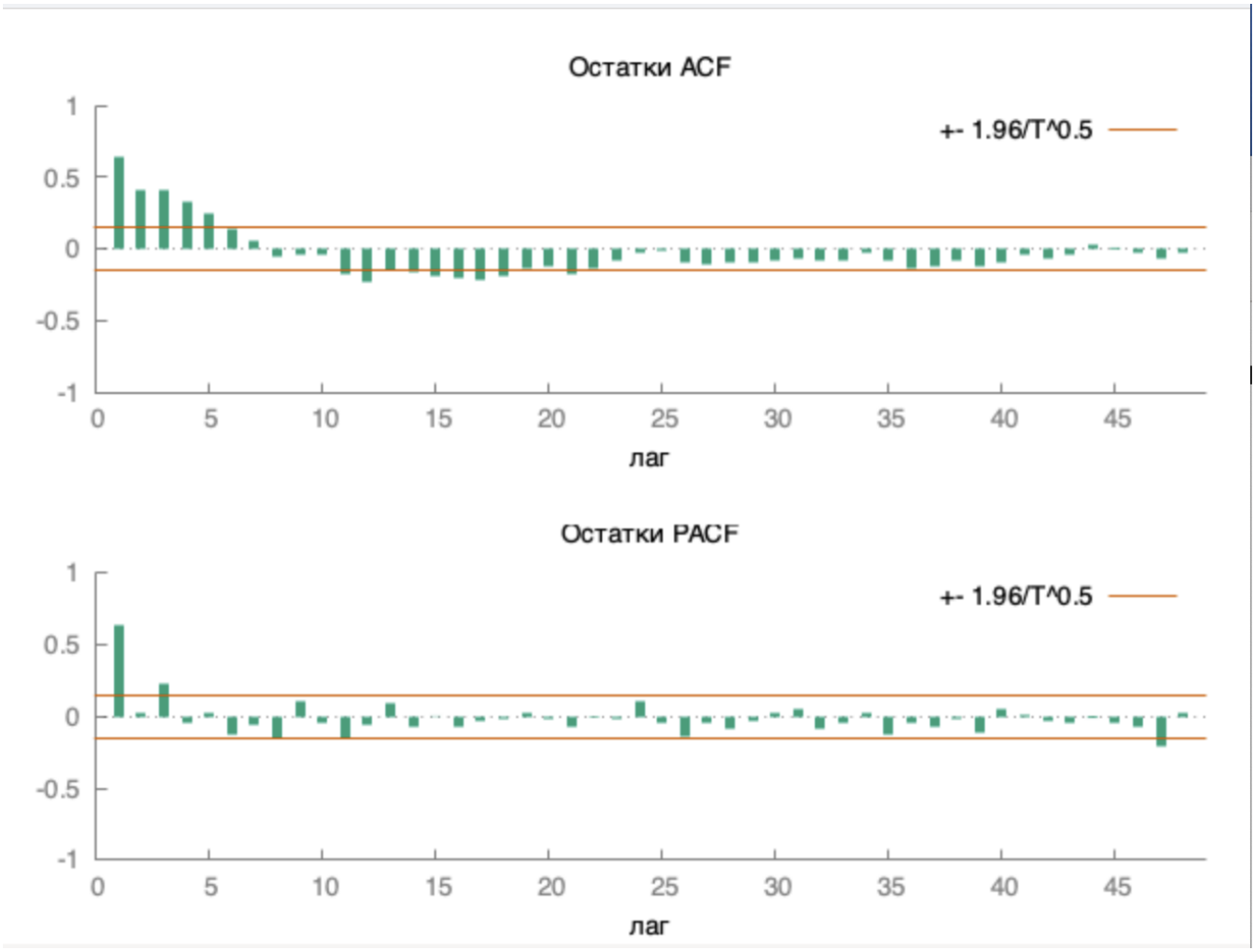
Отклонение потребления электроэнергии для dm2-dm11 по сравнению с первым месяцем более чем на 8000 процентных пунктов меньше(причем для летних месяцев доходит до 20000).

$time = 51.7004$, говорит о среднем месячном росте почти на $51,7$тыс.МВт·ч.

С точки зрения оценки R^2 модель объясняет 96% дисперсии, что является хорошим результатом.

F-stat = 357.19 при p-value <0.05, то есть в целом модель значима.

Теперь оценим модель на адекватность, изучая ее остатки:



Для первых 5 лагов, значение автокорреляции выходит за доверительный интервал, то есть автокорреляция значима отличается от 0. То есть остатки зависят друг от друга и модель неадекватна.

```
Распределение частот для residual, наблюдения 1-170
количество столбцов = 13, среднее = 9.33035e-12, ст. откл. = 1956.8

       интервал           середина   частота    отн.     инт.

            < -4091.5   -4576.8         3      1.76%    1.76%
   -4091.5 - -3120.7    -3606.1         9      5.29%    7.06% *
   -3120.7 - -2150.0    -2635.4        11      6.47%   13.53% **
   -2150.0 - -1179.2    -1664.6        13      7.65%   21.18% **
   -1179.2 - -208.50    -693.87        38     22.35%   43.53% ********
   -208.50 -  762.24     276.87        39     22.94%   66.47% ********
    762.24 -  1733.0     1247.6        32     18.82%   85.29% ******
    1733.0 -  2703.7     2218.3        14      8.24%   93.53% **
    2703.7 -  3674.5     3189.1         6      3.53%   97.06% *
    3674.5 -  4645.2     4159.8         4      2.35%   99.41%
    4645.2 -  5615.9     5130.6         0      0.00%   99.41%
    5615.9 -  6586.7     6101.3         0      0.00%   99.41%
          >=  6586.7     7072.0         1      0.59%  100.00%

Нулевая гипотеза - нормальное распределение:
Хи-квадрат(2) = 5.576 p-значение 0.06156
```

Второй этап проверки на адекватность модели - тест Харке-Бера остатков на нормальность. Для нашего $p - value = 0.062 > 0.05$ мы не отклоняем нулевую гипотезу о том, что остатки распределены нормально. Но значение p-value близко к граничному.

По итогу у построенной модели:

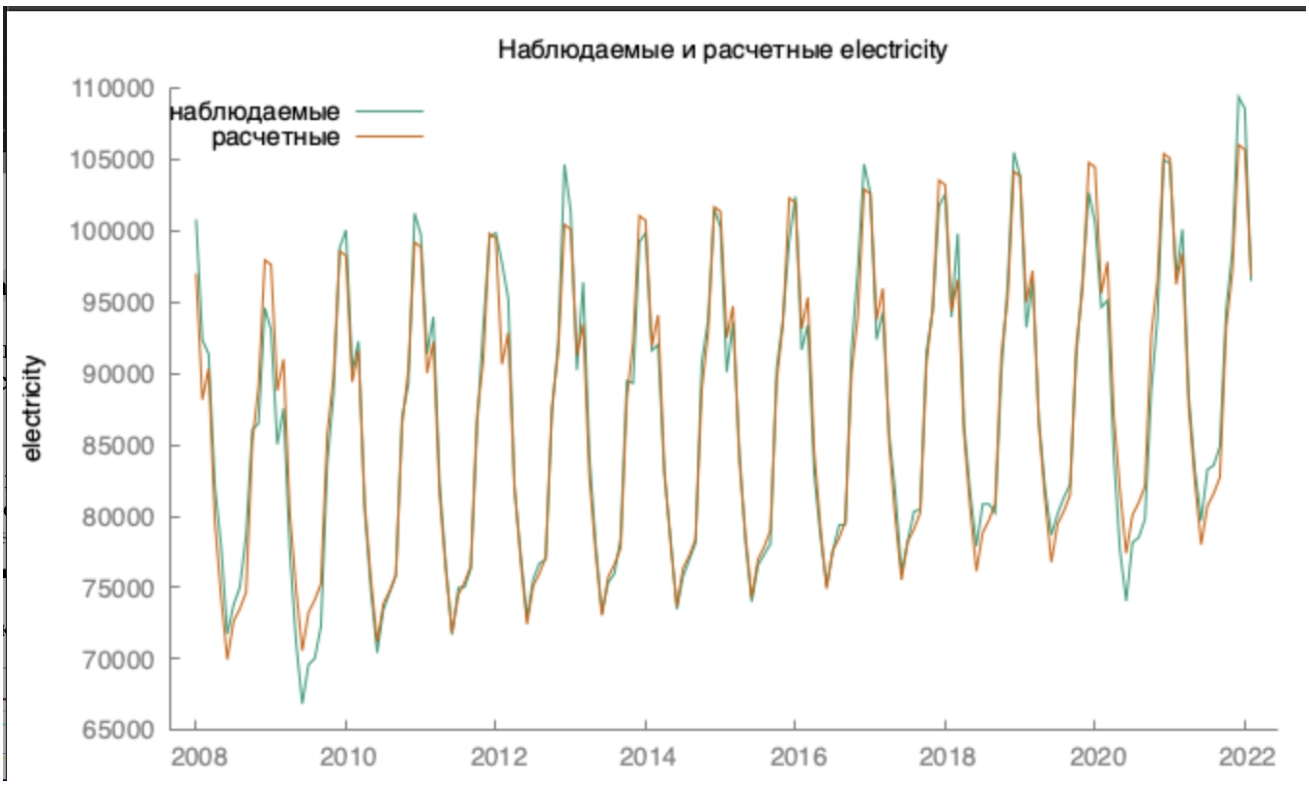
Высокие значения R^2 и F-статистики;

Значимость ключевых переменных;

Остатки почти нормально распределены;

Единственная слабость — высокая автокорреляция остатков , что свидетельствует о необходимости доработки модели

Модель с сезонными фиктивными переменными хорошо подходит для анализа данных с ярко выраженной сезонностью, как в случае потребления электроэнергии. Она позволяет: учитывать различия между месяцами; выявить тренд потребления во времени; объяснить значительную часть изменения показателя. В целом это видно и на сравнении с нашим временным рядом:



2.4) Теперь также построим модель с гармоническими переменными. Добавим 4 переменные: 2 для первой гармоники($a_1 cos(\frac{2\pi t}{12}); b_1 sin(\frac{2\pi t}{12})$) и 2 для второй гармоники($a_2 cos(\frac{2\pi t}{6}); b_2 sin(\frac{2\pi t}{6})$)(аналогично предполагая наличие линейного тренда). Таким образом, мы получим модель $Y_t = a + bt + a_1 cos(\frac{2\pi t}{12}) + b_1 sin(\frac{2\pi t}{12}) + a_2 cos(\frac{2\pi t}{6}) + b_2 sin(\frac{2\pi t}{6})$. Получили следующую модель:

Файл Правка Тесты Сохранить Графики Анализ LaTeX

```
Модель 2: МНК, использованы наблюдения 2008:01-2022:02 (T = 170)
Зависимая переменная: electricity

             коэффициент   ст. ошибка   t-статистика   p-значение
  ---------------------------------------------------------------
  const      82392.9       457.065        180.3        1.83e-190  ***
  time          51.2526      4.63755       11.05       1.42e-21   ***
  cos1       11923.4       321.703         37.06       1.29e-81   ***
  sin1        4891.75      321.703         15.21       3.85e-33   ***
  cos2        1005.35      322.433          3.118      0.0022     ***
  sin2         506.376     320.601          1.579      0.1162

Среднее завис. перемен  86915.26   Ст. откл. завис. перем  10033.55
Сумма кв. остатков      1.44e+09   Ст. ошибка модели       2963.622
R-квадрат               0.915337   Исправ. R-квадрат       0.912756
F(5, 164)               354.6187   P-значение (F)          5.90e-86
Лог. правдоподобие     -1597.174   Крит. Акаике            3206.348
Крит. Шварца            3225.162   Крит. Хеннана-Куинна    3213.982
параметр rho           -0.065410   Стат. Дарбина-Уотсона   2.098094
обратите внимание на сокращенные обозначения статистики

Исключая константу, наибольшее p-значение получено для переменной 17 (sin2)

Тест на нормальное распределение ошибок -
  Нулевая гипотеза: ошибки распределены по нормальному закону
  Тестовая статистика: Хи-квадрат(2) = 0.977525
  p-значение = 0.613385
```

Анализируя метрики построенной модели, стоит отметить:

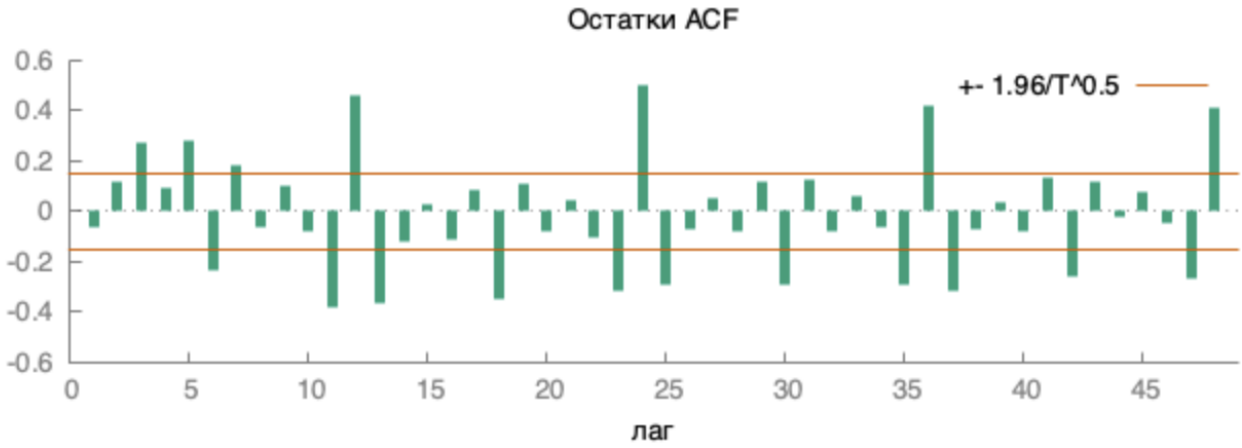
Все переменные получились значимыми, кроме 4-ой = sin2.

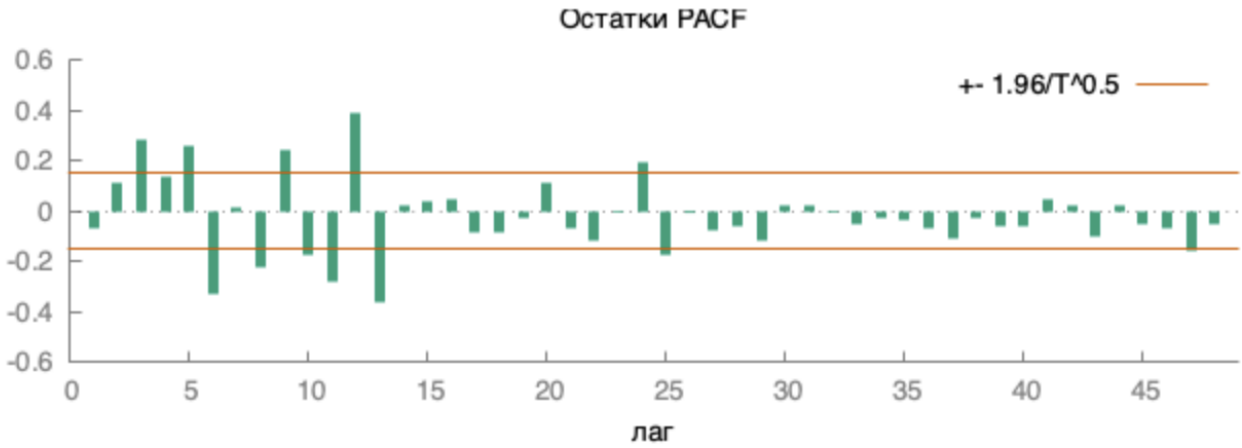
$time = 51.25$, говорит о среднем месячном росте почти на $51,25$тыс.МВт·ч.

С точки зрения оценки R^2 модель объясняет 91% дисперсии, что является также хорошим результатом, но хуже чем первая модель с фиктивными переменными.

F-stat = 354.62 при p-value <0.05, то есть в целом модель значима.

Теперь оценим модель на адекватность, изучая ее остатки:





При малых лагах (1–2) автокорреляция невелика, но: начиная с 3-го лага, значительная автокорреляция (особенно лаги 3, 5, 6, 12, 24 и др.), то есть проблема с автокорреляцией остатков остается, как и у 1ой модели и модель нельзя считать адекватной. Далее исследуем остатки на нормальность:

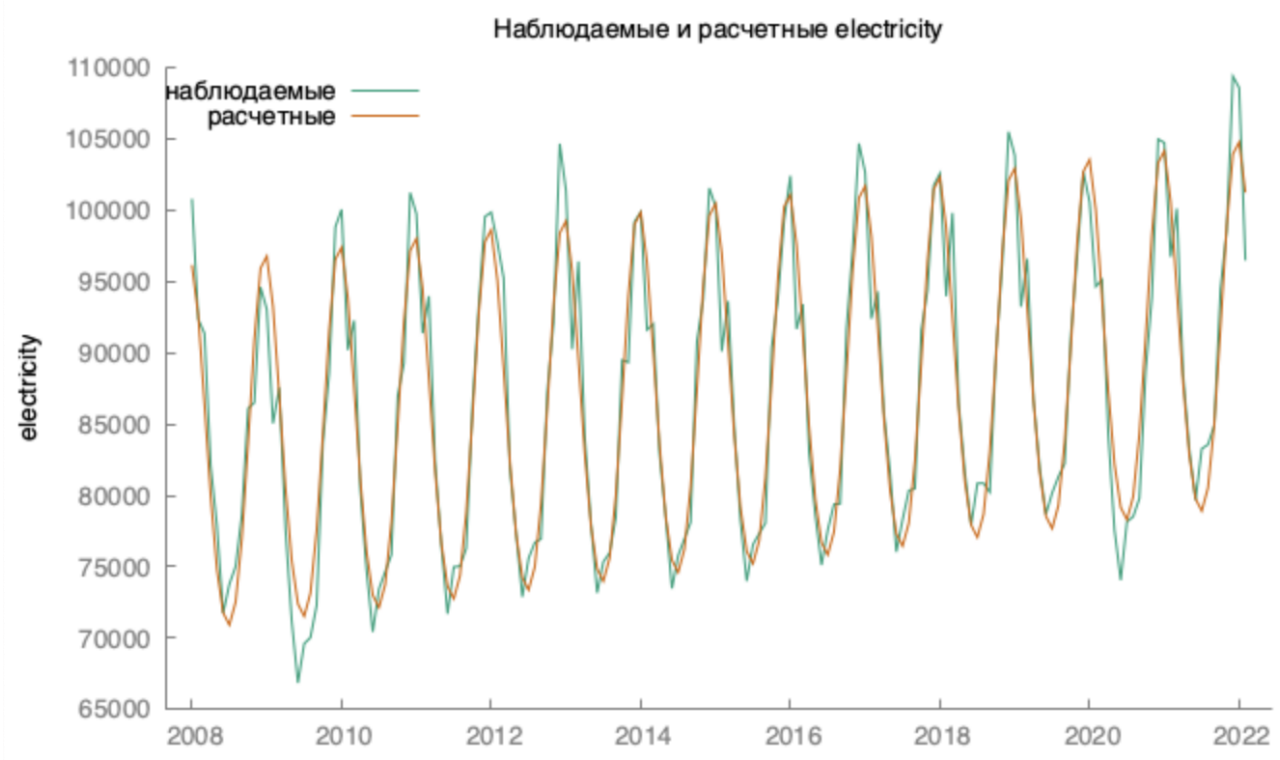
```
Распределение частот для residual, наблюдения 1-170
количество столбцов = 13, среднее = 2.33687e-11, ст. откл. = 2963.62

       интервал          середина   частота    отн.     инт.

         < -7547.8   -8190.0         1      0.59%    0.59%
   -7547.8 - -6263.4   -6905.6         1      0.59%    1.18%
   -6263.4 - -4979.1   -5621.3         8      4.71%    5.88% *
   -4979.1 - -3694.7   -4336.9        11      6.47%   12.35% **
   -3694.7 - -2410.3   -3052.5        15      8.82%   21.18% ***
   -2410.3 - -1126.0   -1768.2        20     11.76%   32.94% ****
   -1126.0 -  158.38   -483.81        25     14.71%   47.65% *****
    158.38 -  1442.7    800.56        35     20.59%   68.24% *******
    1442.7 -  2727.1    2084.9        30     17.65%   85.88% ******
    2727.1 -  4011.5    3369.3        12      7.06%   92.94% **
    4011.5 -  5295.8    4653.6         5      2.94%   95.88% *
    5295.8 -  6580.2    5938.0         5      2.94%   98.82% *
          >=  6580.2    7222.4         2      1.18%  100.00%

Нулевая гипотеза - нормальное распределение:
Хи-квадрат(2) = 0.978 p-значение 0.61338
```

Значение p-value(0.61) аналогично больше уровня значимости и мы не отклоняем нулевую гипотезу о нормальности остатков. В итоге вид у полученной модели также схож с моделью с фиктивными переменными:



Для построенной модели с гармоническими переменными и проведенным анализом можно сделать следующие выводы: Модель достаточно хорошо описывает зависимость потребления от времени и сезонности (высокий R^2, значимые коэффициенты). Однако проблема автокорреляции остатков указывает на неполное объяснение сезонных/периодических эффектов — возможно, нужны дополнительные лаги.

Таким образом: Модель с фиктивными переменными строилась по принципу создания 11 dummy-переменных (dm2...dm12), где январь выступал базовым месяцем (все dummy=0). Уравнение имело вид $Y_t = a + bt + \sum_{i=2}^{12} c_i d_i$. Такая модель объясняет почти 96% дисперсии ряда, все месячные dummy (кроме dm12) и тренд time статистически значимы. Однако в остатках наблюдалась сильная автокорреляция, а нормальность проверка Харке–Бера показала p=0.062 (остатки близки к нормальным).

Гармоническая модель описана двумя парами тригонометрических функций: $sin1 = sin(2\pi * time/12), cos1 = cos(2\pi * time/12)$ (первая гармоника), $sin2 = sin(4\pi * time/12), cos2 = cos(4\pi * time/12)$ (вторая гармоника), плюс та же переменная тренда time. Уравнение: $Y_t = a + bt + a_1 cos(\frac{2\pi t}{12}) + b_1 sin(\frac{2\pi t}{12}) + a_2 cos(\frac{2\pi t}{6}) + b_2 sin(\frac{2\pi t}{6})$. . Эта модель объясняет около 91.5% дисперсии, основные параметры sin1–cos2 значимы, sin2 незначим. Остатки нормально распределены, но сохраняют автокорреляцию на многих лагах.

В сравнении:

– По точности подгонки (R^2) и устранению автокорреляции выигрывает модель с фиктивными месяцами.

– По количеству параметров и плавному описанию сезонности (синусоиды вместо жёстких шагов между месяцами) — гармоническая модель.

– Модель с dummy-переменными лучше фиксирует отдельные аномалии в конкретные месяцы, гармоническая — обобщает повторяющиеся волны.

В обоих случаях для дальнейшего улучшения модели целесообразно убрать оставшуюся автокорреляцию.

2.6) Давайте рассчитаем прогноз по 2ой модели на следующий месяц, то есть на март 2022 года, что равносильно t = 171 наблюдению.

$y_t = a + bt + a_1 cos(\frac{2\pi t}{12}) + b_1 sin(\frac{2\pi t}{12}) + a_2 cos(\frac{2\pi t}{6}) + b_2 sin(\frac{2\pi t}{6}) = 82392.9 + 51.2526 * 171 + 11923.4 * cos(\frac{\pi * 171}{6}) + 48491.75 * sin(\frac{\pi * 171}{6}) + 1005.35 * cos(\frac{\pi t}{3}) + 506.376 * sin(\frac{\pi * 171}{3}) = 91157.0946 + 11923.4 * 0.008109 + 4891.75 * 1 - 1005.35 * 1 + 506.376 * 0.01622 = 95148.395$

Итог: прогноз потребления электроэнергии на март 2022 по гармонической модели составляет примерно 95148.395 тыс.МВт·ч.