МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ ИМ. ПРОФ. М.А. БОНЧ-БРУЕВИЧА»
(СПбГУТ)

ФАКУЛЬТЕТ

КАФЕДРА

Практическая работа по теме: «Прогнозирование трафика Интернета вещей с помощью параметрических моделей ARIMA»

Дисциплина: «Математическое и программное обеспечение киберфизических систем»
Вариант

Выполнил:

Студент группы ИКПИ-
.

Подпись _____

Принял:

к.т.н., доцент кафедры СС и ПД
Гребенщикова А. А.

Дополнительная информация для практической по МиПоКС (ИКПИ-43)

1. Исследование трафика Интернета вещей в Wireshark

Анализ трафика проводился на основе реальных дампов трафика, в котором присутствует нагрузка интернета вещей (Internet of Things, IoT). Данные интеллектуальной среды, состоящей из устройств интернета вещей, собирались в течение шести месяцев. В число устройств входили камеры, светильники, розетки, датчики движения, бытовая техника и мониторы состояния здоровья. Только часть данных доступна для использования исследовательским сообществом. Помимо устройств типа IoT в сформированных файлах присутствует также трафик устройств, не относящихся к данному типу. На веб-ресурсе представлены дампы трафика за 20 дней, а сами данные имеют форматы ***pcap*** *и* ***csv****.*

Таким образом, перед анализом сетевого трафика было необходимо отфильтровать данные в формате ***pcap*** *по принципу отбрасывания пакетов, физический адрес которых не относился к устройствам типа IoT. Фильтрация нужного набора пакетов проводилась в программе анализа сетевого трафика Wireshark. Для примера на рисунке 1 представлено распределение трафика в течение 10 дней для дальнейшего прогнозирования.*

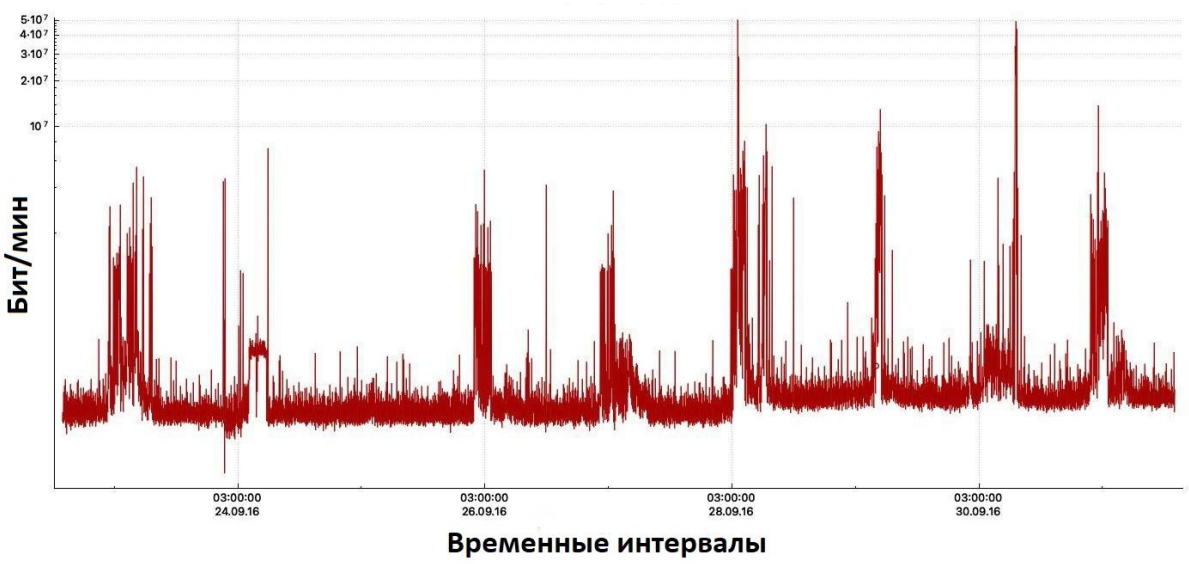


Рисунок 1 – Распределение трафика типа интернета вещей в течение 10 дней

Как показано на рисунке 1, наибольшая интенсивность трафика наблюдается в ночное время, что может быть связано с режимом работы устройств интернета вещей и/или со стремлением уменьшить нагрузку канала в период пиковой нагрузки днем.

В рамках данного пункта дополнительно вам понадобятся все ваши скриншоты с wireshark!!! + не забыть уточнить что для дальнейшей работы формат времени целенаправленно был переведен в “время между пакетами”.

2. Исследование трафика Интернета вещей и апроксимация

При начальном анализе дампов трафика были получены графики плотностей распределения интервалов между пакетами (рисунок 2) и длин пакетов (рисунок 3).

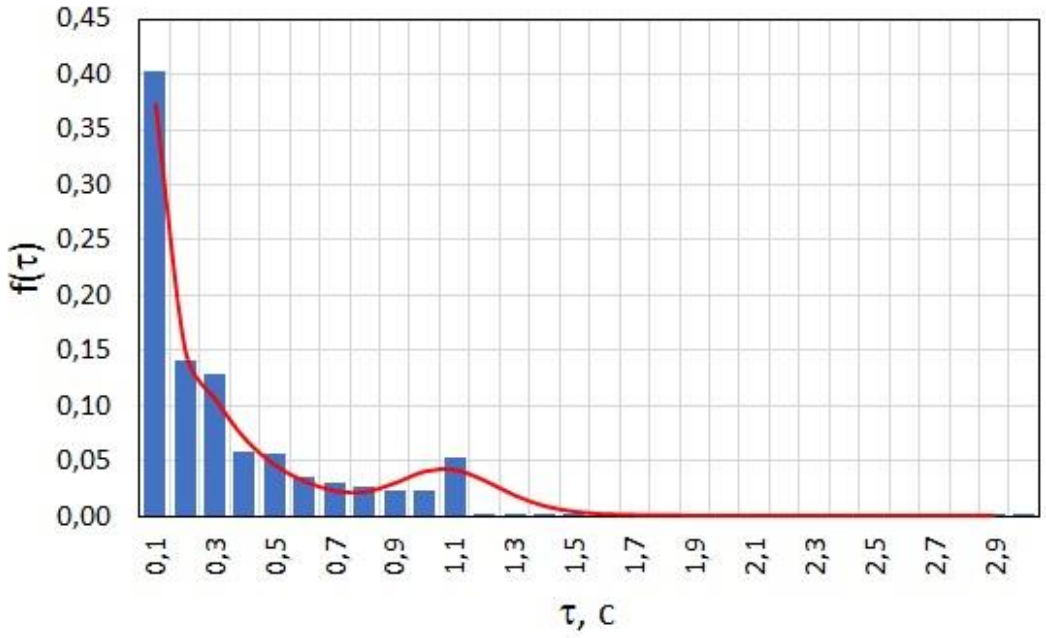


Рисунок 2 – Плотность распределения времени между пакетами трафика интернета вещей

Как видно на рисунке 2, полученное распределение было аппроксимировано и на основе этой аппроксимации была выведена математическая функция вида

$$f(t) = k_i f_i(t) + k_{i+1} f_{i+1}(t) + \ldots + k_m f_m(t) = \sum_{i=1}^{m} k_i f_i(t), \qquad (1)$$

где m>0; $\sum_{i=1}^{m} k_i = 1.$

Таким образом, для плотности распределения, подходящей под данный пример, было подобрано три коэффициента. Несмотря на то, что распределение времени между пакетами на промежутке между 0,5 и 1 с можно идентифицировать как экспоненциальную функцию с коэффициентом k_1=0,15, не рекомендуется игнорировать всплеск на графике при 1,1 с. На промежутках времени τ<0,3 с и τ>0,8 с распределение можно аппроксимировать гамма-функциями с коэффициентами k_2=0,1 и k_3=0,15 соответственно.

НЕОБХОДИМО УКАЗАТЬ ВАШУ ПОЛУЧЕННУЮ ФУНКЦИЮ СО ВСЕМИ КОЭФФИЦИЕНТАМИ ПО ПРИМЕРУ (1 формула).

Дополнительно в рамках данного пункта необходимо привести так же график, где собраны все плотности распределения вместе. Так же указать что оценка производилась по RMSE (PFD и CDF), формулы расчета, отличия и соответствующие выводы согласно оценке аппроксимации.

3. Самоподобие.

Вследствие наличия во временном ряде сильной зависимости значений от предыдущих возникает такое понятие как самоподобие. Таким образом, Интернет-трафик при сглаживании имеет определенную структуру с трендом, на которую стохастически влияют редкие «всплески» пакетов. Такие особенности трафика имеют особе значение и влияние на математические моменты временной последовательности как в локальных масштабах времени, так и на больших размахах.

В традиционных моделях сетевого трафика практически отсутствует свойство долговременной зависимости. Причиной такого явления выступает ярко выраженное сглаживание пульсаций. Именно с приходом концепции единой мультисервисной сети процесс генерации трафика стоит рассматривать с точки зрения степени самоподобия трафика и его долговременной зависимости. Таким образом, теория пуассоновских процессов уходит на второй план, уступая самоподобному процессу, для которого характерно сохранение некоторых статистических особенностей при масштабировании времени.

В рамках данного пункта необходимо дать описания двум методам оценки параметра Херста – дисперсионный анализ и r/s анализ. Привести полученные графики, где будет указана линия тренда и соответствующие ей функция и R^2. Так же указать как по этим данным найти параметр Херста и чему он равен в первом и во втором случае.

Дополнительно в рамках исследования дисперсионного анализа привести результаты по оценке стационарности ряда и обосновать необходимость (или наоборот, почему не меняем исходный ряд) применения разности на основе отклонений математического ожидания и дисперсии.

4. Приведение ряда в эквидистантный вид и поиск портфеля моделей

На основе полученных значений параметров Херста по дисперсионному анализу и rs анализу обосновать выбор периода агрегации и окна (window size).

Дополнительно привести график агрегированного ряда, например:

Для дальнейшего исследования трафика использовались файлы в формате ***csv****, которые содержали такие характеристики, как идентификационный номер, дата, время, физические адреса устройств отправителя и получателя, тип протокола и длина пакетов. Обработка и визуализация такого массива данных осуществлялась с помощью программы,*

написанной на языке Python. Для оптимального анализа в исследуемых файлах удалялись все характеристики, кроме времени поступления и длин пакетов.

В соответствии с методикой, имеющиеся данные, соответствующие временному промежутку в 24 ч с 17:00 23.09 до 17:00 24.09 приводились в эквидистантный вид с различным временем агрегации. На рисунке представлено распределение трафика в течение 24 ч на примере одного из дней наблюдения с периодом агрегации 10 мин.

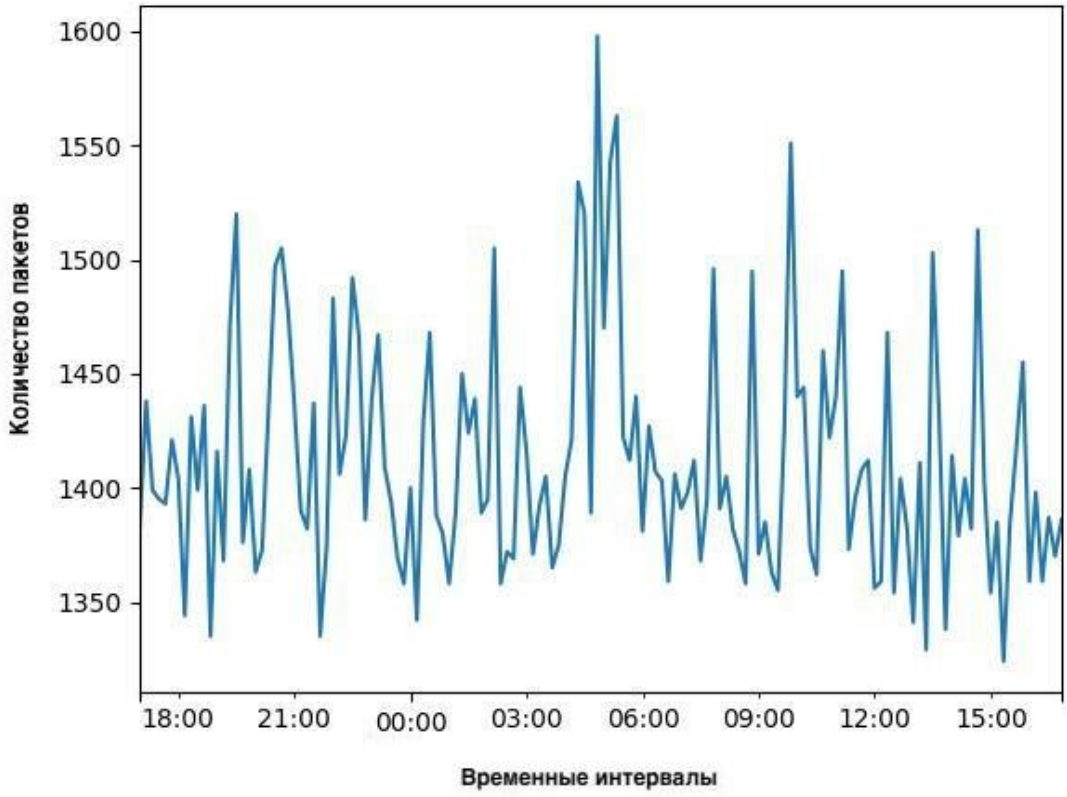
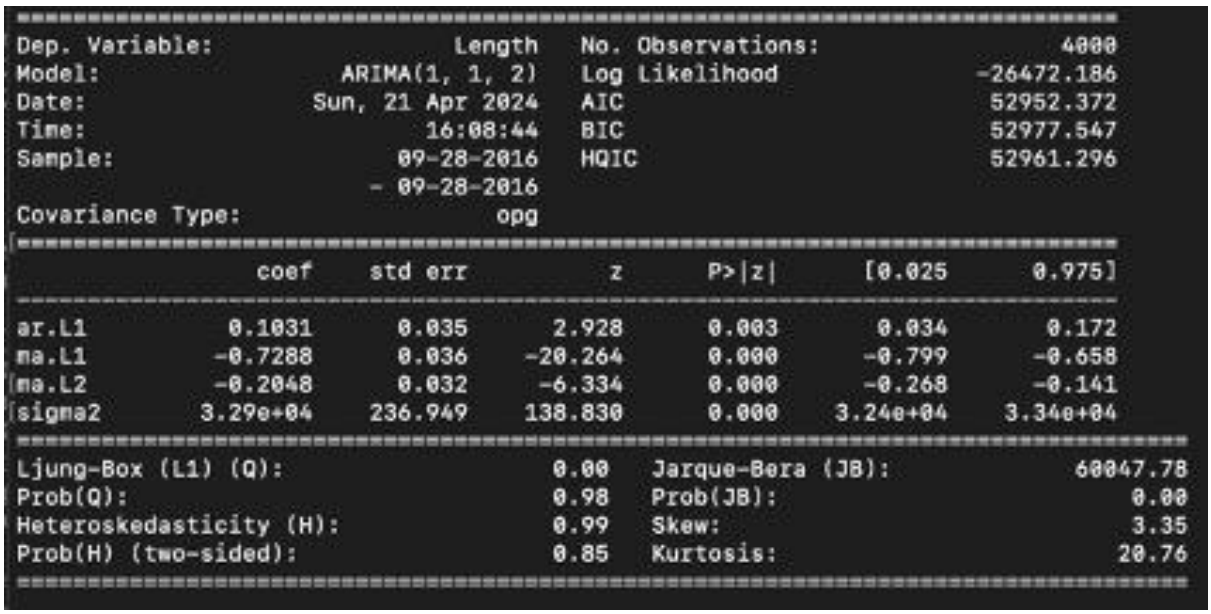


Рисунок – Распределение трафика на временных интервалах с периодом агрегации 10 мин в первый день наблюдений

Продемонстрировать полученный портфель моделей ARIMA и привести статистику для выбранных моделей на основе статистически значимых параметров. Пример статистики:



Dep. Variable:	Length	No. Observations:	4000
Model:	ARIMA(1, 1, 2)	Log Likelihood	-26472.186
Date:	Sun, 21 Apr 2024	AIC	52952.372
Time:	16:08:44	BIC	52977.547
Sample:	09-28-2016	HQIC	52961.296
	- 09-28-2016		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1031	0.035	2.928	0.003	0.034	0.172
ma.L1	-0.7288	0.036	-20.264	0.000	-0.799	-0.658
ma.L2	-0.2048	0.032	-6.334	0.000	-0.268	-0.141
sigma2	3.29e+04	236.949	138.830	0.000	3.24e+04	3.34e+04

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	60047.78
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	0.99	Skew:	3.35
Prob(H) (two-sided):	0.85	Kurtosis:	20.76

Рисунок – Результаты по модели ARIMA(1,1,2)

5. Оценка прогноза

Необходимо продемонстрировать полученный график прогноза для ваших моделей ARIMA на 5 шагов и соответствующие оценки прогноза согласно MAPE и SER, сделать выводы.

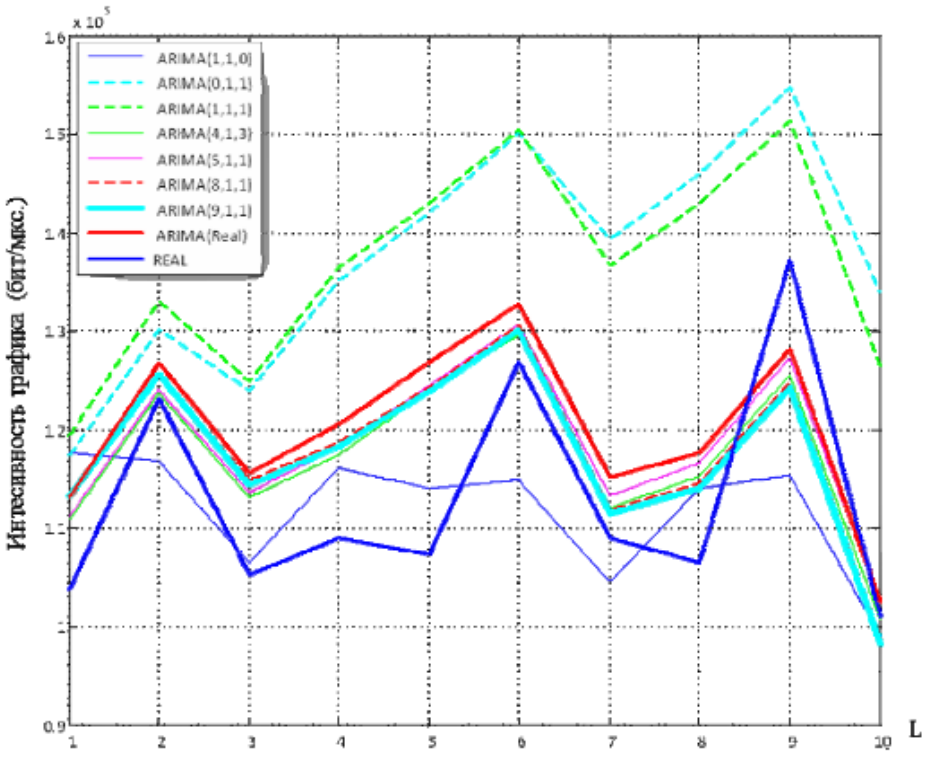


Рис. 6. Прогноз значений трафика для одного из случайных участков временного ряда

Для оценки точности прогноза используется ряд стандартных показателей.

Средняя абсолютная процентная ошибка (MAPE):

$$MAPE = \frac{100\%}{L}\sum_{t=1}^{L}\left|\frac{X_t - \hat{X}_t}{X_t}\right|, \qquad (5)$$

где X_t – реальное значение, $\hat{X}_t$ - прогнозное значение, L – интервал прогноза. Если $MAPE < 10\%$, то прогноз сделан с высокой точностью, $10\% < MAPE < 20\%$ - прогноз хороший, $20\% < MAPE < 50\%$ - прогноз удовлетворительный, $MAPE > 50\%$ - прогноз плохой.

Отношение сигнала к шуму (SER):

$$SER = 10\lg\left(\frac{\sum_{t=1}^{L} X_t^2}{\sum_{t=1}^{L}\left(X_t - \hat{X}\right)^2}\right). \qquad (6)$$

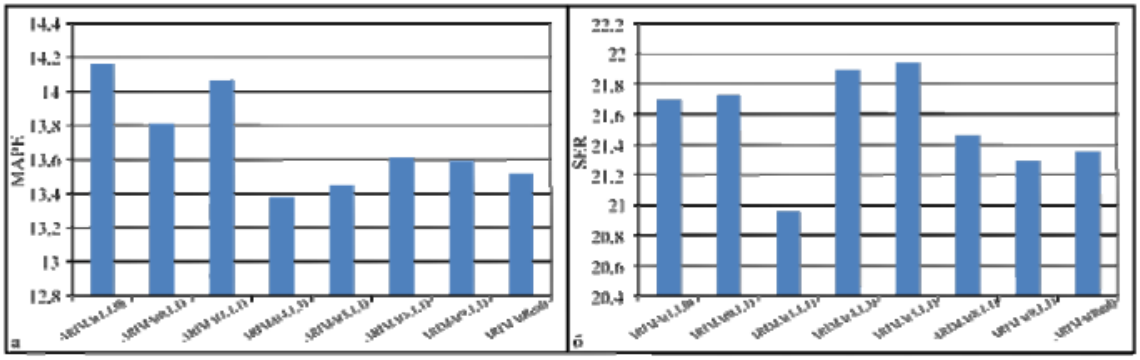


Рис. 7. Среднее значение коэффициентов MAPE и SER для прогноза на 1 шаг вперед

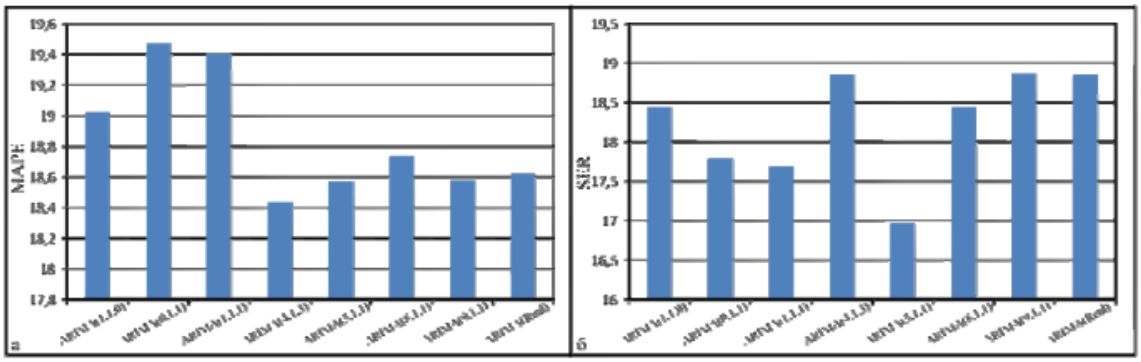


Рис. 8. Среднее значение коэффициентов MAPE и SER для прогноза на 2 шага вперед

6. Оценка остатков ARIMA

Модели ARIMA (p,d,q) представляют особый класс нестационарных моделей, которые считаются однородными и находятся в статистическом равновесии. ***Модель ARIMA определяется уравнением (2)*** *и такая модель является обобщённой, т.к. включает в себя в качестве частных случаев следующие модели: авторегрессионные, скользящего среднего, смешанные модели авторегрессии-скользящего среднего и интеграция всех трёх соответственно.*

$$\Delta^d Y_t = \varphi_1 \Delta^d Y_{\{t-1\}} + \ldots + \varphi_p \Delta^d Y_{\{t-p\}} + \varepsilon_t + \theta_1 \varepsilon_{\{t-1\}} + \ldots + \theta_q \varepsilon_{\{t-q\}} \qquad (2)$$

Эффективная итеративная процедура построения моделей для описания зависимости наблюдаемых временных рядов состоит из трёх этапов:

1. Идентификация. Подразумевает использование уже имеющихся данных о временном ряде, чтобы подобрать соответствующий и оптимальный класс моделей для последующей оценки.

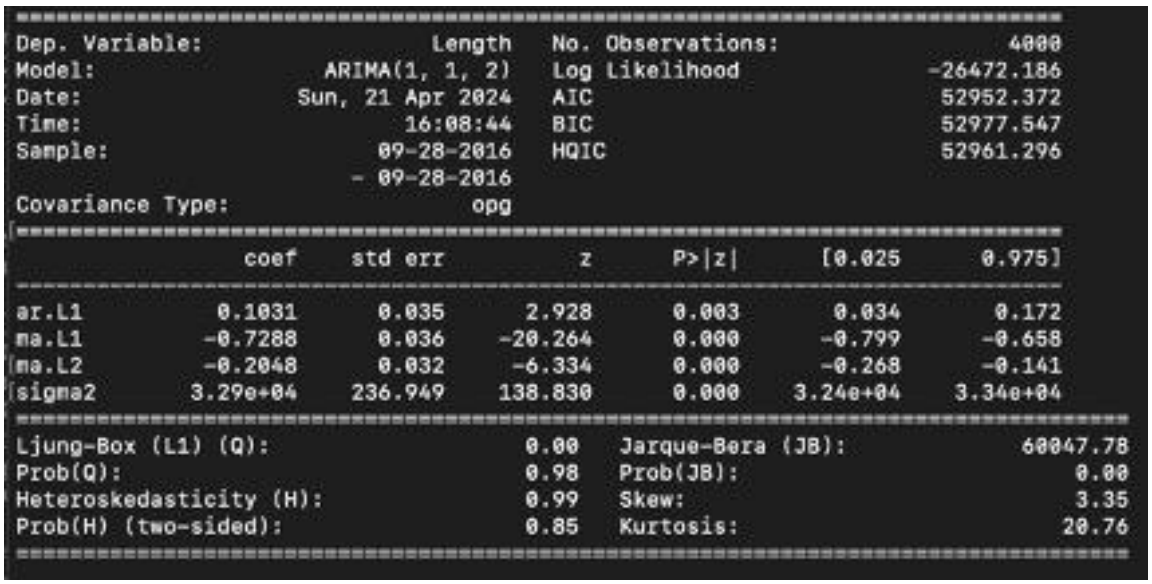
2. Оценка. Подразумевает выявление сопутствующих параметров для построения модели.

3. Диагностика. Подразумевает проверку модели на соответствие имеющимся данным, чтобы выявить несоответствия и определить оптимальные пути улучшения модели.

Для подтверждения адекватности полученной модели необходимо убедиться, что ряд остатков представляет собой случайную компоненту. Таким образом, чтобы оптимально провести анализ остатков сформированной модели используется Q-тест Льюнг–Бокса для проверки гипотезы на наличие автокорреляции в данных. При H0 (нулевая гипотеза об отсутствии корреляции) статистика Q-теста асимптотически имеет распределение Хи-квадрат. Значения p выше 0,05 указывают на принятие нулевой гипотезы о точности модели при уровне значимости 95%.

Наиболее известный тест для проверки гипотезы о нормальном распределении остатков – статистический тест Харке-Бера (Jarque-Bera test). Также необходимо убедиться в наличии условной гетероскедастичности с помощью теста множителей Лагранжа (ARCH LM-тест).

В рамках данного пункта необходимо по наилучшей модели из предыдущего пункта: записать в виде функции (2) и вынести в таблицу результаты по остаткам + построить график АКФ + нормальность распределения (график и пример кода приведены ниже) - на основе всего сделать выводы.



Dep. Variable:		Length	No. Observations:			4000
Model:		ARIMA(1, 1, 2)	Log Likelihood			-26472.186
Date:		Sun, 21 Apr 2024	AIC			52952.372
Time:		16:08:44	BIC			52977.547
Sample:		09-28-2016	HQIC			52961.296
		- 09-28-2016				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1031	0.035	2.928	0.003	0.034	0.172
ma.L1	-0.7288	0.036	-20.264	0.000	-0.799	-0.658
ma.L2	-0.2048	0.032	-6.334	0.000	-0.268	-0.141
sigma2	3.29e+04	236.949	138.830	0.000	3.24e+04	3.34e+04
Ljung-Box (L1) (Q):		0.00	Jarque-Bera (JB):			60047.78
Prob(Q):		0.98	Prob(JB):			0.00
Heteroskedasticity (H):		0.99	Skew:			3.35
Prob(H) (two-sided):		0.85	Kurtosis:			20.76

Рисунок – Результаты по модели ARIMA(1,1,2)

Таблица – Оценка остатков модели ARIMA(1,1,2)

	ARIMA (1,1,2)	
	Статистика	p-значение
Тест Льюинга-Бокса	0.0	0.98
Тест Харке-Бера	60047.78	0.00
ARCH-LM тест остатков (p-значение)	0.99	0.85
AIC	52952.372	

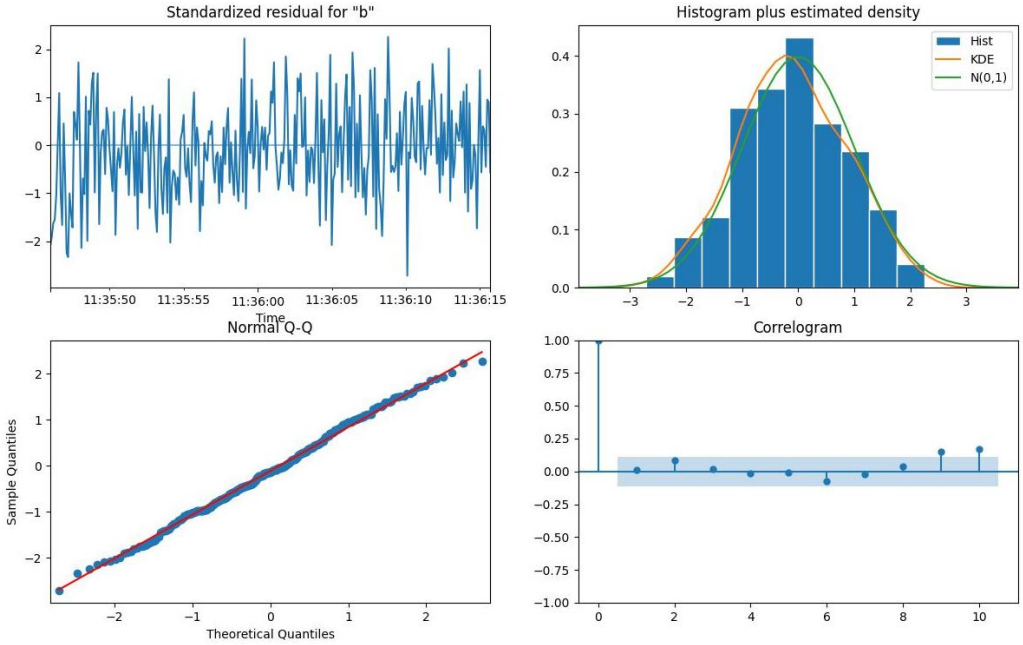


Рисунок – Результаты диагностики по остаткам модели ARIMA(1,1,2)

Вывести графики диагностики можно по примеру:

```
model = ARIMA(train, order=(1,1,2))
stats_mdl = model.fit()
stats_mdl.plot_diagnostics(figsize = (15, 10))
plt.show()
```