Целью данного анализа является исследование временного ряда коэффициентов естественного прироста населения на 1000 человек в Центральном федеральном округе России. Для начала был построен график (Рис. 1), отображающий динамику коэффициентов естественного прироста населения на 1000 человек в Центральном федеральном округе. Визуальный анализ графика позволяет выделить периодические колебания и наличие тренда в данных, что предполагает возможность структурных изменений в ряду.

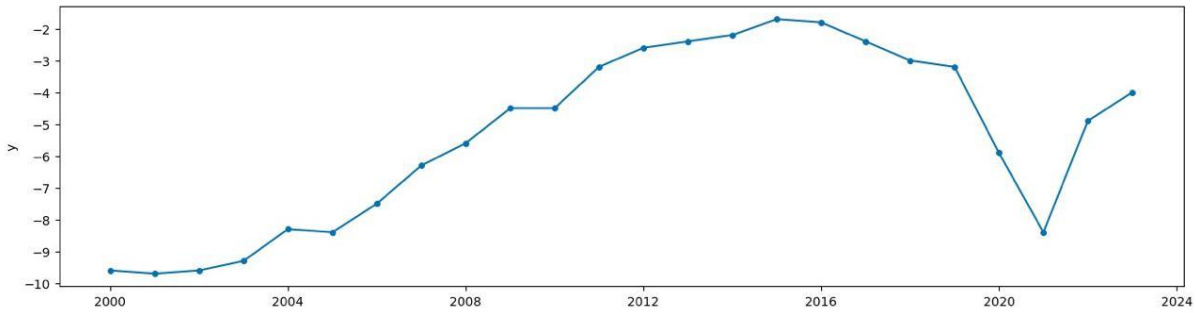

Рис. 1. Коэффициент естественного прироста населения на 1000 человек в ЦФО.

Для анализа стационарности временного ряда были проведены тесты: ADF и KPSS.

- Тест ADF: Этот тест проверяет гипотезу о наличии единичного корня, который свидетельствует о наличии тренда в ряде. В нашем случае p-value теста составил 0.132, что указывает на наличие тренда в данных, так как гипотеза о стационарности не была отвергнута на уровне значимости 10%.

- Тест KPSS: Этот тест проверяет гипотезу о стационарности ряда. Результат теста показал p-value 0.066, что также указывает на отсутствие стационарности ряда, поскольку гипотеза о стационарности не была отвергнута на уровне значимости 5%.

Оба теста свидетельствуют о наличии тренда в ряде, что требует дальнейших преобразований данных для приведения ряда к стационарному виду.

Было принято решение применить метод разностей для удаления тренда. Этот метод заключается в вычитании текущего значения из предыдущего, что позволяет избавиться от долгосрочных тенденций в данных.

После применения первых разностей был проведен повторный тест на стационарность. Результаты тестов показали: p-value = 0.946 в случае ADF, что указывает на сохранение нестабильности в ряду, p-value для KPSS теста составил 0.1, что также подтверждает наличие нестабильности в ряду. Таким образом, ряд остался нестационарным после применения первых разностей, что требует применения вторичных разностей.

Для дальнейшего приведения ряда к стационарному виду были применены вторые разности и снова проведен анализ на стационарность:

- Тест ADF: p-value составил 0.47, что по-прежнему указывает на нестационарность ряда.

- Тест KPSS: p-value составил 0.1, что также не отвергает гипотезу о стационарности.

Так как результаты остались схожими, было решено найти третьи разности для достижения стационарности. После применения третьих разностей ряд стал стационарным, так как p-value для теста ADF составил 0.002, что подтверждает стационарность ряда (гипотеза о наличии тренда отверглась), и p-value для теста KPSS составил 0.1, что подтверждает стационарность ряда. Таким образом, после применения третьих разностей временной ряд стал стационарным и готов для дальнейшего анализа.

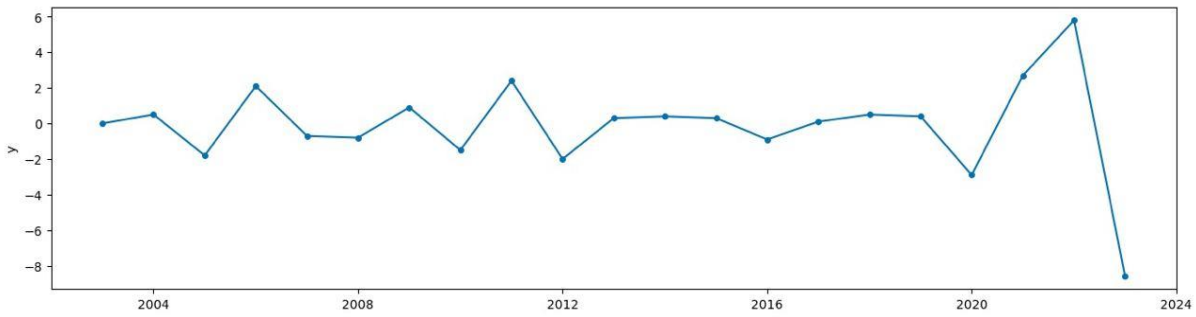

Рис. 2. Третья разность для коэффициента естественного прироста населения на 1000 человек в ЦФО.

Для выбора оптимальных гиперпараметров модели ARIMA (параметры p и q), были построены графики автокорреляции (ACF) и частичной автокорреляции (PACF). Графики ACF и PACF помогают определить, сколько лагов следует использовать для моделей AR (авторегрессии) и MA (скользящего среднего).

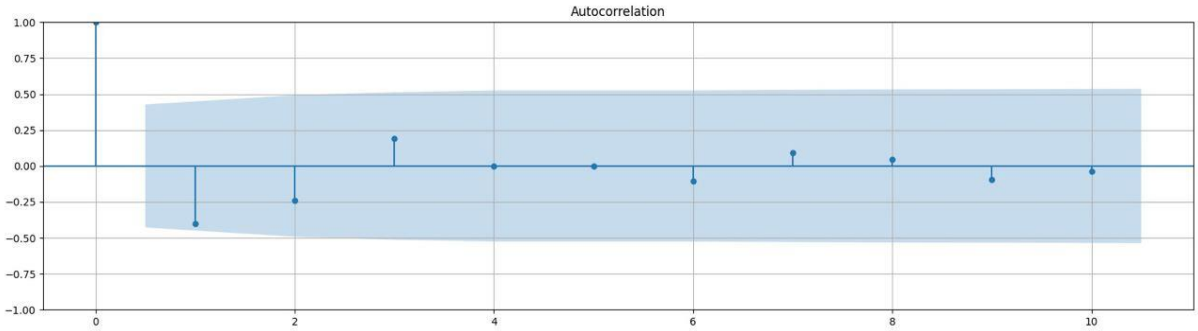

Рис. 3. График ACF для третьей разности.

График автокорреляции показывает, как значения ряда на разных временных интервалах связаны друг с другом. Это помогает определить порядок для модели MA(1). В данном случае по коррелограмме ACF мы предположили, что значимых лагов кроме 0 - нет, поэтому рассмотрим модель MA(1) для третьей разности.

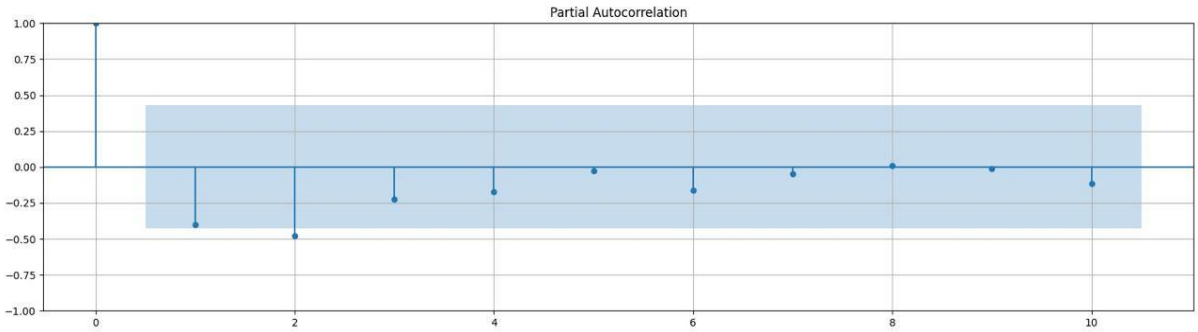

Рис. 4. График PACF для третьей разности.

График частичной автокорреляции помогает выявить количество лагов для модели AR (авторегрессия). В данном случае по коррелограмме PACF мы предположили, что значимый лаг второй, поэтому рассмотрим модель AR(2) для третьей разности.

Для оценки точности прогноза данных был выполнен процесс разделения временного ряда на обучающую и тестовую выборки, результат отображен на Рис. 5. Для тестовой выборки были использованы последние три года данных. Это позволило провести проверку качества прогноза и оценить результаты модели.

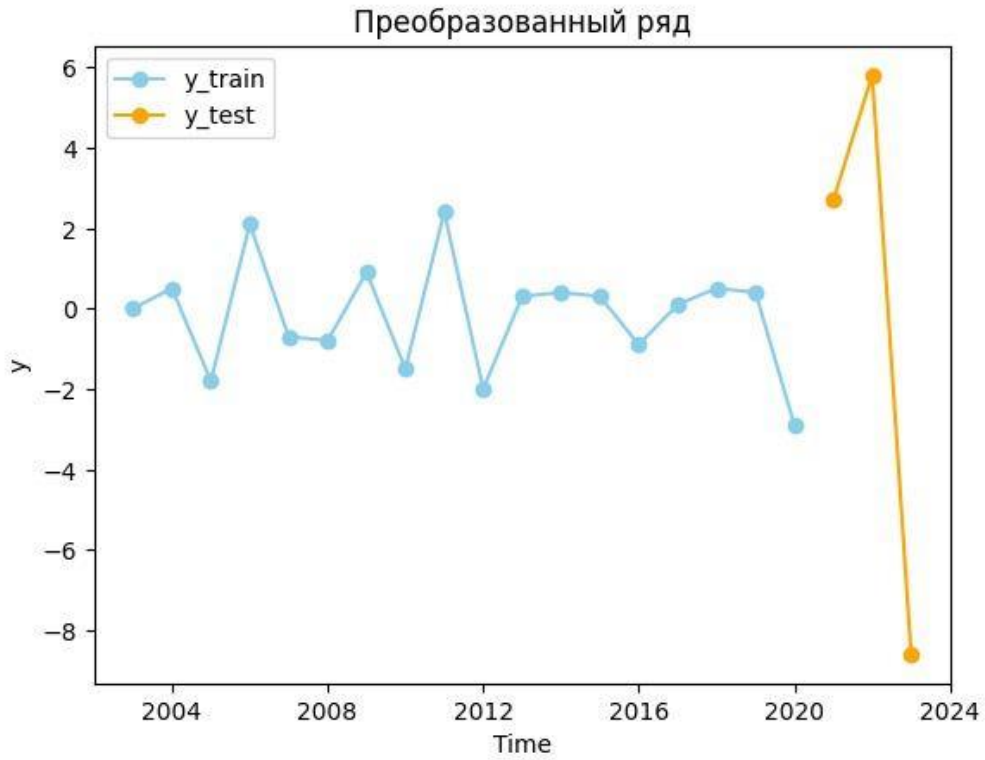

Рис. 5. Третья разность, разбиение на тестовую и тренировочную выборки.

Следующим шагом является выбор наилучшей модели для прогнозирования. В данном случае были рассмотрены три основные модели для третьей разности (параметр I = 3):

AR(2) - авторегрессионная модель второго порядка, которая использует два предыдущих значения временного ряда для предсказания текущего значения (Результаты приведены на Рис. 6).

MA(1) - модель скользящего среднего первого порядка, которая использует одно предыдущее значение ошибки прогноза для расчета текущего значения (Результаты приведены на Рис. 8).

ARMA(1, 1) - комбинированная модель, которая использует как авторегрессию (AR), так и скользящее среднее (MA) для предсказания (Результаты приведены на Рис. 10).

SARIMAX Results

Dep. Variable:		y	No. Observations:			18
Model:	ARIMA(2, 0, 0)		Log Likelihood			-22.764
Date:	Thu, 24 Apr 2025		AIC			53.527
Time:	18:58:50		BIC			57.089
Sample:	01-01-2003		HQIC			54.018
	- 01-01-2020					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0639	0.090	-0.708	0.479	-0.241	0.113
ar.L1	-1.1674	0.341	-3.419	0.001	-1.837	-0.498
ar.L2	-0.6231	0.426	-1.461	0.144	-1.459	0.213
sigma2	0.6679	0.279	2.395	0.017	0.121	1.214

Ljung-Box (L1) (Q):	2.38	Jarque-Bera (JB):	0.29
Prob(Q):	0.12	Prob(JB):	0.87
Heteroskedasticity (H):	3.15	Skew:	-0.31
Prob(H) (two-sided):	0.19	Kurtosis:	3.02

Рис. 6. Результаты модели ARIMA(2,3,0).

После построения моделей был произведен процесс прогнозирования. Для этого использовались данные обучающей выборки, и модель была применена для предсказания значений на тестовой выборке. Прогнозы для тестовой выборки были получены и визуализированы на графиках, чтобы сравнить фактические значения с прогнозами. Графики были построены для следующих моделей на данных третьих разностей: Рис. 7 - AR(2), Рис. 9 - MA(1), Рис. 11 - ARMA(1, 1).

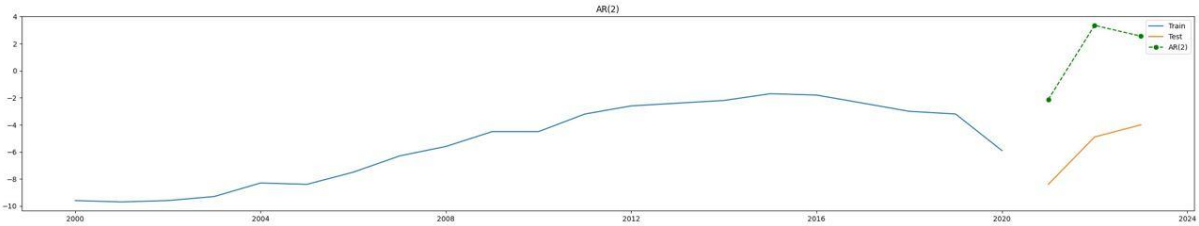

Рис. 7. Прогнозное значение, ARIMA(2,3,0).

SARIMAX Results

Dep. Variable:	y	No. Observations:	18
Model:	ARIMA(0, 0, 1)	Log Likelihood	-23.229
Date:	Thu, 24 Apr 2025	AIC	52.458
Time:	18:59:32	BIC	55.129
Sample:	01-01-2003	HQIC	52.826
	- 01-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0556	0.056	-0.998	0.318	-0.165	0.054
ma.L1	-1.0000	1094.821	-0.001	0.999	-2146.809	2144.809
sigma2	0.6568	719.109	0.001	0.999	-1408.771	1410.084

Ljung-Box (L1) (Q):	5.03	Jarque-Bera (JB):	0.94
Prob(Q):	0.02	Prob(JB):	0.62
Heteroskedasticity (H):	1.76	Skew:	-0.56
Prob(H) (two-sided):	0.51	Kurtosis:	3.09

Рис. 8. Результаты модели ARIMA(0,3,1).

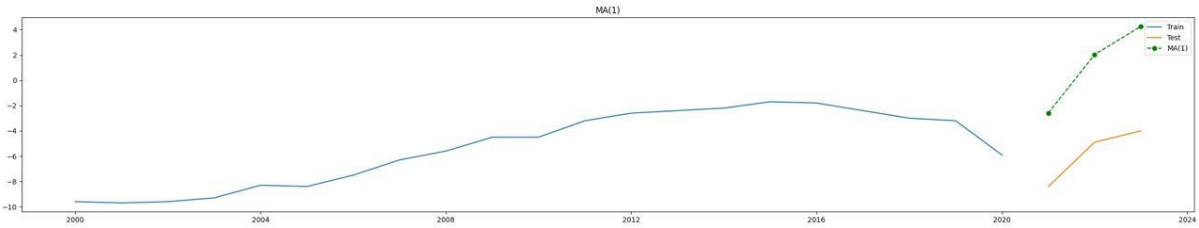

Рис. 9. Прогнозное значение, ARIMA(0,3,1).

SARIMAX Results

Dep. Variable:	y	No. Observations:	18
Model:	ARIMA(1, 0, 1)	Log Likelihood	-19.987
Date:	Thu, 24 Apr 2025	AIC	47.974
Time:	18:59:48	BIC	51.535
Sample:	01-01-2003	HQIC	48.465
	- 01-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0427	0.029	-1.459	0.145	-0.100	0.015
ar.L1	-0.6594	0.284	-2.325	0.020	-1.215	-0.104
ma.L1	-0.9999	520.788	-0.002	0.998	-1021.726	1019.727
sigma2	0.4206	219.039	0.002	0.998	-428.887	429.729

Ljung-Box (L1) (Q):	2.67	Jarque-Bera (JB):	0.95
Prob(Q):	0.10	Prob(JB):	0.62
Heteroskedasticity (H):	2.15	Skew:	-0.55
Prob(H) (two-sided):	0.37	Kurtosis:	2.79

Рис. 10. Результаты модели ARIMA(1,3,1).

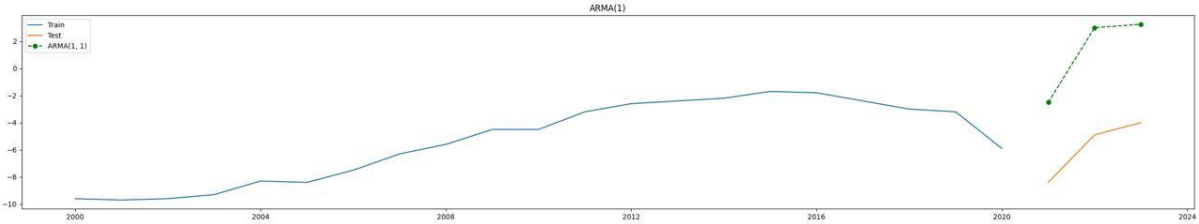

Рис. 11. Прогнозное значение, ARIMA(1,3,1).

Для каждой модели были рассчитаны следующие метрики: RMSE (корень из средней квадратичной ошибки, который измеряет точность прогноза), AIC (информационный критерий, который используется для выбора модели, балансируя точность прогноза и сложность модели), BIC (информационный критерий, аналог AIC, но с более строгим штрафом за сложность модели). Результаты расчетов отражены на Рис. 12.

Method	RMSE	AIC	BIC
AR(2)	7.08	53.527033	57.088520
MA(1)	7.07	52.457706	55.128821
ARMA(1, 1)	7.06	47.973938	51.535425

Рис. 12. Оценка качества моделей.

Модель ARIMA(1,3,1) показала наименьшие значения AIC и BIC, что свидетельствует о ее оптимальности для данного временного ряда. Также, её RMSE также является минимальным, что подтверждает её хорошую точность в прогнозировании. На основе этих результатов, ARIMA(1,3,1) является наиболее подходящей моделью для данного временного ряда на основе предсказательного характера. Результаты прогнозирования отражены на Рис. 13, хоть прогноз сильно завышен, но сохраняется общая тенденция, что можно увидеть из графика.

	Дата	Реальные значения	Прогноз
2021-01-01	2021-01-01	-8.4	-2.5
2022-01-01	2022-01-01	-4.9	3.0
2023-01-01	2023-01-01	-4.0	3.3

Рис. 13. Прогнозное значение лучшей моделей.

Далее мы рассмотрели поиск по сетке, задав в качестве параметров p, d, q; при этом p и q принимало значение от 0 до 3, а d - параметр отвечающий за интеграцию - 0 или 3 (в остальных случаях мы не можем говорить о стационарности).

	p	d	q	AIC	BIC	RMSE
7	0	3	3	42.6	46.1	5.8
14	1	3	2	44.3	47.9	5.4
15	1	3	3	44.3	48.7	5.7
22	2	3	2	44.4	48.8	6.0
6	0	3	2	44.6	47.3	5.5
30	3	3	2	44.8	50.2	5.8
23	2	3	3	45.1	50.5	5.7
18	2	0	2	45.2	51.4	2.8
19	2	0	3	46.1	53.4	3.4
31	3	3	3	46.6	52.9	5.6
29	3	3	1	47.3	51.7	5.3
25	3	0	1	47.8	54.0	3.1
21	2	3	1	48.5	52.0	5.6

Рис. 14. Подбор параметров.

На Рис. 14 отражена таблица результатов подбора параметров для модели ARIMA, результаты отсортированы по наименьшему AIC, но при этом, видим, что модель ARIMA(2,0,2) несильно отличается по AIC, BIC, но при этом значение RMSE значительно меньше, чем для аналогичных моделей по AIC.

Будем использовать модель ARIMA(2,0,2) для прогнозирования на ближайшие пять лет, результаты отражены на Рис. 15 и Рис.16.

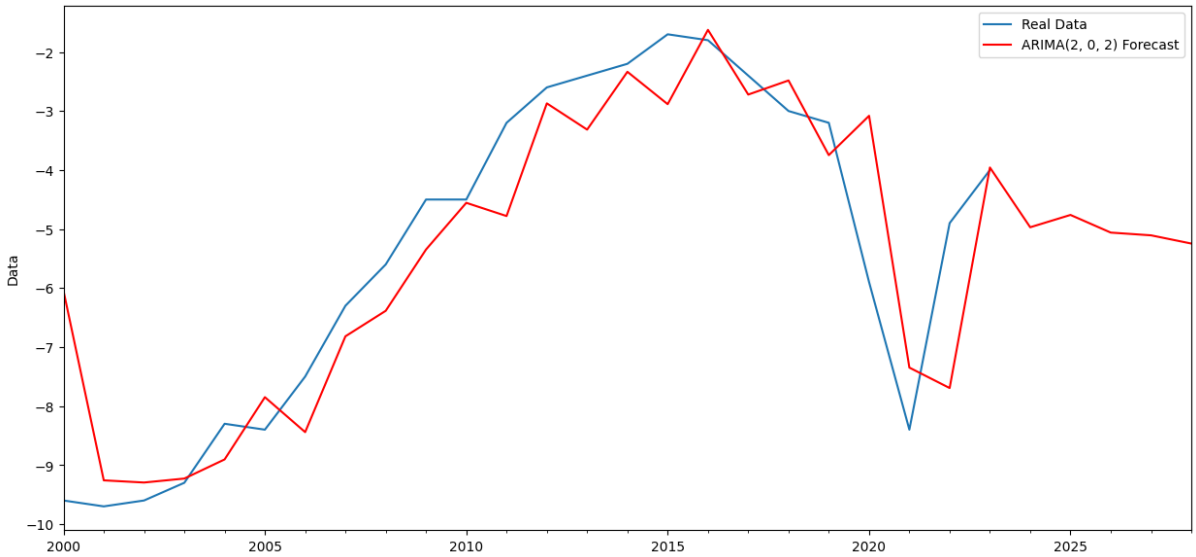

Рис. 15. График прогнозов, ARIMA(2,0,2).

	predicted_mean
2024-01-01	-4.970002
2025-01-01	-4.761112
2026-01-01	-5.060261
2027-01-01	-5.106812
2028-01-01	-5.246103

Рис. 16. Прогнозное значение, ARIMA(2,0,2).

На Рис. 17 отражены результаты метрик качества модели: MAE и MSE показывают, что модель не имеет чрезмерных отклонений, а MAPE ниже 20% подтверждает, что модель делает достаточно приемлимые прогнозы для данного временного ряда.

```
MAE: 0.9
MSE: 1.6
MAPE: 19.3%
```

Рис. 17. Оценка качества модели.