УДК 004.9+574

В. Є. Копняк[1]
В. Б. Мокін[1]

ДОСЛІДЖЕННЯ ПРОБЛЕМ ІЗ ГЕТЕРОСКЕДАСТИЧНІСТЮ ДАНИХ МОНІТОРИНГУ ЯКОСТІ АТМОСФЕРНОГО ПОВІТРЯ

[1]Вінницький національний технічний університет

Анотація

Проведено регресійний аналіз прикладу даних з моніторингу якості атмосферного повітря за щодобовими багаторічними даними EcoCity. Аналіз показав, що ряд є стаціонарним, гетероскедастичним і розподіленим за нормальним законом. Порівняння результатів моделювання даних за класичною для гомоскедастичних рядів моделлю ARIMA та моделями для гетероскедастичних рядів даних ARCH, GARCH, EGARCH, APARCH, HARCH при різних параметрах показало, що гетероскедастичні моделі є менш адекватними, аніж модель ARIMA, яка теж дає чималу похибку. Отже усі ці моделі потребують удосконалення для такої задачі.

Ключові слова: гетероскедастичні часові ряди, GARCH, ARIMA, регресійний аналіз, якість атмосферного повітря, EcoCity.

Abstract

A regression analysis of an example of atmospheric air quality monitoring data based on EcoCity daily multi-year data was carried out. The analysis showed that the series is stationary, heteroskedastic and distributed according to a normal law. A comparison of data modeling results using the classic ARIMA model for homoscedastic series and models for heteroskedastic data series ARCH, GARCH, EGARCH, APARCH, HARCH with different parameters showed that heteroskedastic models are less adequate than the ARIMA model, which also gives a considerable error. Therefore, all these models need to be improved for such a task.

Keywords: heteroskedastic time series, GARCH, ARIMA, regression analysis, atmospheric air quality, EcoCity.

Вступ

Якість повітря має прямий вплив на здоров'я людей. Відсутність адекватного аналізу та передбачення може призвести до небезпечного забруднення повітря, що спричиняє ризик для респіраторного здоров'я та інших захворювань. Застосування GARCH дозволяє моделювати й передбачати коливання забруднюючих речовин та їх вплив на якість повітря, що дозволяє вжити необхідних заходів для запобігання небезпечному забрудненню та збереження здоров'я населення.

Велика кількість реальних процесів у складних системах формалізуються як часові ряди. Особливим їх класом є процеси, які зазнають впливу метеорологічних та інших факторів, які мають нестаціонарний характер. Такі процеси характеризуються як гетероскедастичні, тобто процеси зі змінною дисперсією.

Метою даного дослідження є проаналізувати дані громадського моніторингу стану атмосферного повітря на прикладі мережі EcoCity щодо їх можливої гетероскедастичності та адекватність опису цих даних класичними моделями для гомо- та гетероскедастичних процесів.

Розвідувальний аналіз даних

Використаємо дані моніторингу якості атмосферного повітря від громадського проекту EcoCity [1] та їх сервісу «Кабінет дослідника» [2], до якого мають доступ автори, завдяки угоді між EcoCity і ВНТУ. Візьмемо багаторічний (2020-2023 рр.) ряд щодобових даних громадського моніторингу стану атмосферного повітря з мережі EcoCity, зокрема дані показника «PM10» (пил розміром 10 мкм чи менше) зі станції № 650, розташованої у смт Турбів, Вінницька область (рис. 1).

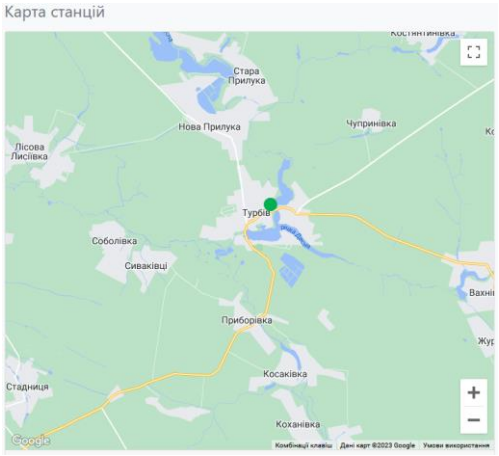

Рисунок 1. Станція «Турбів» системи моніторингу якості повітря у Вінницькій області

Проаналізуємо ряд даних на стаціонарність з використанням тесту Дика-Фуллера. Як видно з рис. 2, ряд є стаціонарним.

```
In [8]:
# Stationarity check
check_stationarity(df['y'])

ADF Statistic: -3.317601
p-value: 0.014110
Critical Values:
        1%: -3.437
        5%: -2.864
        10%: -2.568
Stationary
```

Рисунок 2. Аналіз ряду даних на стаціонарність

Проаналізуємо чи є ряд гетероскедастичним з використанням тесту Шапіро-Вілкса. Як видно з рис. 3, ряд є розподіленим за нормальним законом.

```
In [21]:
arch_test = het_arch(resid, maxlag=lags_max)
shapiro_test = shapiro(st_resid)

print(f'Lagrange mulitplier p-value: {arch_test[1]}')
print(f'Shapiro-Wilks p-value: {shapiro_test[1]}')

Lagrange mulitplier p-value: 0.8696570031254455
Shapiro-Wilks p-value: 2.6230008358657576e-27
```

Рисунок 3. Перевірка ряду з допомогою тестів Шапіро-Вілкса та Лагранжа

Проаналізуємо чи є ряд розподіленим за нормальним законом з використанням тесту множника Лагранжа. Як видно з рис. 3, ряд є гетероскедастичним, оскільки нульова гіпотеза про його гомоскедастичність не підтверджується. Отже, варто застосовувати так звані GARCH-моделі [3, 4]:

- ARCH(q) – це модель ARIMA, де для опису змін дисперсії в часі використовується авторегресія (AR), модель з авторегресійною умовною гетероскедастичністю; метод ARCH моделює дисперсію на кроці часу як функцію залишкових помилок від середнього процесу (наприклад, нульового середнього);

- GARCH(p, q) – це базова GARCH-модель (узагальнена ARCH), де для опису змін дисперсії в часі використовується ARMA і де p - порядок моделі ARCH (Autoregressive Conditional Heteroskedasticity), що визначає залежність дисперсії від попередніх квадратів власне часового ряду;

- IGARCH(p, q) – це Integrated GARCH, додає можливість моделювання диференційованої дисперсії, тобто залежність дисперсії від попередніх змін дисперсії, а не самої дисперсії; вона може бути корисною, коли спостерігається деградація в часовому ряді;

- EGARCH(p, q) – це Exponential GARCH, яка дозволяє моделювати асиметрічні зміни дисперсії, оскільки використовує логарифм варіантів змін;

- TGARCH(p, q) – це Threshold GARCH, дозволяє моделювати залежність дисперсії від порогового рівня; використовує два окремі рівня змінності: один для перевищень вищого порогу інший для значень нижче порогу;

- APARCH модель (Asymmetric Power ARCH) – використовується для моделювання архімедійної волатильності в часових рядах, де зміна волатильності може бути асиметричною та залежати від попередніх відхилень;

- HARCH модель (Heterogeneous ARCH) – використовується для моделювання гетероскедастичної волатильності в часових рядах, де зміна волатильності може залежати від попередніх відхилень та може виявляти неоднорідність в структурі волатильності.

В якості метрики пропонуються інформаційний критерій Акаїке (AIC) та Байсівській інформаційний критерій (BIC), які характеризують компроміс між точністю моделі та її складністю, дозволяють порівнювати альтернативні моделі та вибирати найкращу модель з мінімальним значенням критерію.

Результати дослідження

Авторами розроблено програму на Python, яка для заданого часового ряду (показник PM10 на станції «Турбів», але можна взяти й іншу станцію та показник на ній) автоматично ідентифікує модель ARMA (ARIMA з $d = 0$, оскільки ряд є стаціонарним) та зазначені вище GARCH-моделі з використанням відомих бібліотек [5]. Були проведені різні дослідження параметрів:

1. Різні типи GARCH-моделей ("Constant", "Zero", "AR", "ARX", "HAR", "HARX") не впливають на результат, тому зупинились на параметрі "Constant".
2. Різні параметри авторегресії та ковзного середнього (p, q) дають ті самі досить малі оптимальні значення і таку саму похибку, тому тюнінг проводився тільки для комбінацій значень 1 і 2.
3. Були проаналізовані різні метрики – AIC та BIC, але висновки за ними щодо оптимальної моделі є майже однаковими.

Результати наведені на рис. 4.

model_name	AIC	BIC	params
ARIMA	7566.678075	7591.301137	[2, 0, 1]
APARCH	11106.665251	11136.207023	[2, 1]
APARCH	11106.890662	11131.508805	[1, 1]
EGARCH	11113.299779	11137.917922	[2, 1]
EGARCH	11114.027736	11133.722250	[1, 1]
EGARCH	11115.299779	11144.841550	[2, 2]
EGARCH	11116.027735	11140.645878	[1, 2]
APARCH	11117.322161	11146.863933	[1, 2]
APARCH	11120.382787	11154.848187	[2, 2]
HARCH	11125.437300	11140.208186	[1, 1]
HARCH	11125.437300	11140.208186	[1, 2]
ARCH	11125.443673	11140.214559	[1, 1]
ARCH	11125.443673	11140.214559	[1, 2]
ARCH	11125.869508	11145.564023	[2, 1]
ARCH	11125.869508	11145.564023	[2, 2]
HARCH	11125.869516	11145.564031	[2, 1]
HARCH	11125.869516	11145.564031	[2, 2]
GARCH	11125.880888	11145.575402	[1, 1]
GARCH	11127.855593	11152.473736	[1, 2]
GARCH	11127.880890	11152.499034	[2, 1]
GARCH	11129.615347	11159.157118	[2, 2]

Рисунок 4. Результат тюнінгу моделей ARIMA та GARCH-моделей
для показника «PM10» станції № 650 (Турбів) мережі EcoCity з рис. 1

Як видно на рис. 4, моделювання даних за класичною для гомоскедастичних рядів моделлю ARIMA та моделями для гетероскедастичних рядів даних ARCH, GARCH, EGARCH, APARCH, HARCH при різних параметрах показало, що гетероскедастичні моделі є менш адекватними, аніж модель ARIMA, яка теж дає чималу похибку. Крім того, проаналізовано значущість ідентифікованих параметрів. Для більшості моделей з рис. 4 хоча б якийсь із параметрів, а то – й усі, мають $\alpha > 0{,}05$ (рис. 5а), тобто нульова гіпотеза про адекватність тих моделей не є статистично значущою. Адекватно ідентифікувалась тільки одна модель EGARCH(1,1) (рис. 5б), але вона теж має чималу похибку. Отже усі ці моделі є недостатньо адекватними і потребують удосконалення для поставленої задачі.

Constant Mean - GARCH Model Results

Dep. Variable:	pct_change	R-squared:	0.000
Mean Model:	Constant Mean	Adj. R-squared:	0.000
Vol Model:	GARCH	Log-Likelihood:	-5558.81
Distribution:	Normal	AIC:	11129.6
Method:	Maximum Likelihood	BIC:	11159.2
		No. Observations:	1016
Date:	Tue, Jun 20 2023	Df Residuals:	1015
Time:	19:45:48	Df Model:	1

Mean Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	12.5664	1.894	6.635	3.243e-11	[8.854, 16.278]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	74.8492	217.829	0.344	0.731	[-3.521e+02,5.018e+02]
alpha[1]	2.0143e-17	2.909e-02	6.924e-16	1.000	[-5.702e-02,5.702e-02]
alpha[2]	0.0158	2.838e-02	0.558	0.577	[-3.979e-02,7.146e-02]
beta[1]	0.3562	0.255	1.397	0.163	[-0.144, 0.856]
beta[2]	0.6053	0.243	2.491	1.274e-02	[0.129, 1.082]

а)

Constant Mean - EGARCH Model Results

Dep. Variable:	pct_change	R-squared:	0.000
Mean Model:	Constant Mean	Adj. R-squared:	0.000
Vol Model:	EGARCH	Log-Likelihood:	-5553.01
Distribution:	Normal	AIC:	11114.0
Method:	Maximum Likelihood	BIC:	11133.7
		No. Observations:	1016
Date:	Tue, Jun 20 2023	Df Residuals:	1015
Time:	19:45:57	Df Model:	1

Mean Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	14.0839	2.008	7.014	2.314e-12	[10.148, 18.019]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	5.7315	1.184	4.840	1.296e-06	[3.411, 8.052]
alpha[1]	0.3342	0.146	2.284	2.238e-02	[4.739e-02, 0.621]
beta[1]	0.2949	0.144	2.045	4.083e-02	[1.230e-02, 0.578]

б)

Рисунок 5. Статистична значущість параметрів та структури регресій GARCH-моделей: а) приклад моделі GARCH(2,2), яка не є статистично значущою; б) модель EGARCH(1,1), усі параметри якої є статистично значущими

Висновки

Проведено регресійний аналіз прикладу даних з моніторингу якості атмосферного повітря за щодобовими багаторічними даними EcoCity на прикладі однієї станції та певного показника на ній (№ 650 «Турбів», показник «PM10»).

Аналіз за рядом тестів показав, що ряд є стаціонарним, гетероскедастичним і розподіленим за нормальним законом. Порівняння результатів моделювання даних за класичною для гомоскедастичних рядів моделлю ARIMA та моделями для гетероскедастичних рядів даних ARCH, GARCH, EGARCH, APARCH, HARCH при різних параметрах показало, що гетероскедастичні моделі є менш адекватними, аніж модель ARIMA, яка теж дає чималу похибку. Крім того, усі ці моделі, окрім однієї, мають статистично не значущі коефіцієнти, що теж доводить їх неадекватність. Отже усі ці моделі потребують удосконалення для такої задачі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Eco-City Громадський моніторинг стану якості повітря [Електронний ресурс] – Режим доступу до ресурсу: https://eco-city.org.ua/.

2. Eco-City Кабінет дослідника [Електронний ресурс] – Режим доступу до ресурсу: https://archive.eco-city.org.ua/.

3. Jason Brownlee. How to Model Volatility with ARCH and GARCH for Time Series Forecasting in Python. August 24, 2018. Machine Learning Mastery. Time Series. https://machinelearningmastery.com/develop-arch-and-garch-models-for-time-series-forecasting-in-python

4. Mokin Vitalii. Kopniak Volodymyr. Kaggle Notebook “Air State Analysis: Comparison of ARIMA and GARCH models”, 2023, https://www.kaggle.com/code/vbmokin/air-state-analysis-comparison-of-arima-and-garch.

Копняк Володимир Євгенович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, vkopnyak@gmail.com

Мокін Віталій Борисович **–** д-р техн. наук, проф., завідувач кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: vbmokin@vntu.edu.ua

Kopniak Volodymyr Y. – Postgraduate student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, vkopnyak@gmail.com

Mokin Vitalii B. **–** Dr. Tech. Sciences, Prof., Head of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: vbmokin@vntu.edu.ua