ISSN 2224-087X. Електроніка та інформаційні технології. 2013. Випуск 3. С. 211–219
Electronics and information technologies. 2013. Issue 3. P. 211–219

УДК 004.415.2-519.246.8

ПОБУДОВА ARIMA МОДЕЛЕЙ ЧАСОВИХ РЯДІВ ДЛЯ ПРОГНОЗУВАННЯ МЕТЕОДАНИХ НА МОВІ ПРОГРАМУВАННЯ R

О. Дзендзелюк, Л. Костів, В. Рабик

Львівський національний університет імені Івана Франка
вул. ген. Тарнавського, 107, 79017 Львів, Україна
RabykV@ukr.net

Розглянуто алгоритм побудови ARIMA моделі часових рядів на основі методології Бокса - Дженкінса. Приведено опис критеріїв, які використовуються для перевірки стаціонарності часових рядів, вибору оптимальної ARIMA моделі та перевірки її адекватності. Основну увагу акцентовано на реалізації цього алгоритму на мові програмування R. Як приклад приведено реалізацію короткочасного прогнозування температури повітря на Чорногірському географічному стаціонарі.

Ключові слова: часові ряди, автокореляційна функція, часткова автокореляційна функція, ARIMA модель, прогнозування, мова прогнозування R.

Прогнозування метеоданих забезпечує важливу інформацію про майбутню погоду. Існує низка методів для прогнозу метеорологічних даних. Для короткочасного прогнозування доцільно використовувати статистичні методи, які базуються на ідентифікації параметрів певних моделей часових рядів [1]. При цьому вважається, що фактори, які впливали на формування погоди в недалекому минулому, будуть діяти і в найближчому майбутньому.

У цій роботі прогнозування метеоданих базується на ідентифікації параметрів ARIMA моделей часових рядів, які задають інформацію про навколишній стан довкілля. Вони добре описують як стаціонарні, так і нестаціонарні часові ряди. Розглянуто алгоритми побудови цих моделей та їхню програмну реалізацію, для якої використовувалася мова програмування R [2]. Причина її використання полягає в наявності статистичних пакетів для моделювання і прогнозування часових рядів та в можливості не тільки програмувати власні функції, але в багатьох випадках покращувати вже існуючі.

Для аналізу та прогнозування були використані дані, записані на метеостанції Чорногірського географічного стаціонару Львівського національного університету імені Івана Франка. Їх початковий аналіз показав, що найбільш важливими параметрами метеорологічних даних є температура і тиск повітря, швидкість і напрям вітру. Тому саме ці параметри (температура, тиск, швидкість і напрям вітру) були вибрані для прогнозування. Вони представляють собою часові ряди. Тривалість таких рядів може складати місяці або й роки. Інші фактори, які впливають на прогноз, не враховуються.

Для опису часових рядів використано математичні моделі, які можуть набувати різних форм. Серед них можна виділити авторегресивні моделі, моделі ковзкого середнього та інтегральні моделі. На їхній основі побудовано моделі авторегресивного ковзкого середнього (ARMA) [1], авторегресивного інтегрованого ковзкого середнього (ARIMA) [1].

Нехай X_t , $t \in T_0$ – це множина спостережень, яка отримується послідовно в часі шляхом вимірювань, а T_0 – множина відліків моментів часу, в які виконано спостереження. Спостереження трактуються як реалізація стохастичного процесу $\{ X_t : t \in T_0\}$ за час $T_0 \in T$. Для процесу $\{X_t\}$ із середнім значенням $E(X_t)=\mu$ і ″білим″ шумом $\{\varepsilon_t\}$ ($E(\varepsilon_t)=0$, $Var(\varepsilon_t)=\sigma^2_\varepsilon$, $Cov(\varepsilon_t, \varepsilon_s)=0$ для $t \neq s$) ARMA-процес порядку (p, q) описують виразом [1]:

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)X_t=\left(1+\sum_{j=1}^{q}\beta_i L^j\right)\varepsilon_t . \quad (1)$$

Цей вираз можна представити у вигляді:

$$A(L)X_t = B(L)\varepsilon_t , \quad (2)$$

де L – це лаговий оператор (оператор часового зсуву – $LX_t=X_{t-1}$, $L^{-1}X_{t-1}=X_t$). Поліноми $A(L)$ і $B(L)$ визначаються наступним чином:

$$\begin{aligned} A(L) &= 1-\alpha_1 L-...-\alpha_p L^p , \\ B(L) &= 1+\beta_1 L+...+\beta_q L^q . \end{aligned} \quad (3)$$

За умови $q=0$ процес $\{X_t\}$ називають авторегресивним процесом AR(p), а для $p=0$ – ковзким середнім процесом MA(q).

В аналізі та прогнозуванні нестаціонарних часових рядів часто використовують більш загальну модель ARIMA(p,d,q), яку можна трансформувати до авторегресивної моделі AR(p), моделі ковзкого середнього MA(q) або моделі ARMA(p,q). Як модифікація ARMA(p,q)- процесу, ARIMA(p,d,q)- процес – це d- кратне використання оператора скінченних різниць $\Delta=1-L$ до початкового часового ряду $\{X_t\}$. Його описують рівнянням:

$$A(L)\Delta^d X_t = B(L)\varepsilon_t , \quad (4)$$

де d – це порядок різниці (ціле число).

Побудову ARIMA(p,d,q) моделі часового ряду детально розглянуто в роботі Бокса–Дженкінса [1]. Вона складається з таких етапів:

– визначення загального класу моделей;

– вибір моделі (тобто, значень p, d, q) для експериментальної перевірки;

– оцінка параметрів під час експериментальної перевірки моделі (тобто, обчислення параметрів $\alpha_1, \alpha_2, \ldots, \alpha_p; \beta_1, \beta_2, \ldots, \beta_q$);

– діагностика моделі (перевірка того, чи не має досліджуваний часовий ряд властивостей, що суперечать одержаній моделі);

– використання моделі для виконання прогнозу.

У такому підході Бокса–Дженкінса не передбачено конкретної моделі для прогнозування досліджуваного часового ряду. Задається лише загальний клас моделей, що описують часовий ряд і дають змогу у деякий спосіб виражати поточне значення параметра ряду через його попередні значення. Алгоритм сам обере найбільш оптимальну модель для прогнозу. Для його реалізації використовують ітераційний

підхід. У виборі моделі враховують як якісні характеристики, так і кількість її параметрів.

На етапі ідентифікації моделі необхідно виконати перевірку часового ряду на стаціонарність. Для цього найчастіше використовується візуальний аналіз вибіркової автокореляційної (ACF) і часткової автокореляційної (PACF) функцій. Для стаціонарних часових рядів ACF і PACF швидко спадають після декількох перших значень. Якщо ж графіки спадають повільно, то часовий ряд може виявитися нестаціонарним. Нестаціонарні часові ряди можна перетворити в стаціонарні шляхом взяття різниць. Вихідний ряд замінюється рядом різниць. Взяття різниць може повторюватися декілька раз. Число повторень взяття різниць, необхідних для отримання стаціонарної поведінки даних, позначається параметром d. Також на цьому етапі використовуються статистичні тести на наявність одиничного кореня (розширений тест Діки–Фуллера [1] – ADF).

Після отримання стаціонарного ряду досліджується характер поведінки вибіркових ACF і часткової PACF і висуваються гіпотези про значення параметрів p і q. Під час цього формується базовий набір ARIMA – моделей.

На другому етапі виконується оцінка параметрів цих моделей. Для цих цілей найчастіше використовується метод максимальної правдоподібності. Для отримання початкових значень параметрів ARIMA-моделі використовують рівняння Юла–Уокера [1], а для уточнення оцінок параметрів – метод Марквардта. Для кожної з обраних моделей оцінюють її параметри та обчислюють залишки.

Для перевірки кожної з отриманих моделей на адекватність аналізується її ряд залишків. У адекватної моделі ряд залишків повинен бути подібним на ″білий″ шум. Також для перевірки гіпотези про те, що спостережувані дані є реалізацією ″білого″ шуму, використовується Q – статистика. Q – статистика Льюнга – Бокса визначається наступним чином [1]:

$$Q = n(n+2)\sum_{k=1}^{M}\frac{r_k^2}{n-k}, \tag{5}$$

де n – об’єм вибірки, M – кількість лагів, що тестуються, r_k – коефіцієнти ACF і має асимптотичний розподіл $\chi^2_{1-\alpha,M}$ (квантиль χ^2 - розподілу рівня 1-α з M ступенями вільності (α = 0,05 – рівень значущості)). Якщо $Q < \chi^2_{1-\alpha,M}$ - то приймається гіпотеза про відсутність автокореляції до M- го порядку в досліджуваному ряді залишків.

Якщо в результаті перевірки декілька моделей є адекватними спостережуваним даним, то при кінцевому виборі враховуються фактори: підвищення точності; зменшення числа параметрів моделі. Ці вимоги об’єднані в критеріях Акаіке і Шварца, які побудовані на принципі штрафів за додаткові параметри моделі. Інформаційний критерій Акаіке [1]:

$$AIC = \ln\left(\hat{\sigma}^2\right) + \frac{2\left(p+q+1\right)}{n}, \tag{6}$$

де $\hat{\sigma}^2$ - очікувана вибіркова дисперсія.

Байєсівський інформаційний критерій (критерій Шварца) [1]:

$$BIC = \ln\left(\hat{\sigma}^2\right) + \frac{\left(p+q+1\right)\ln(n)}{n}. \tag{7}$$

Перший доданок в виразах (6) і (7) представляє собою штраф за велику дисперсію, а другий – штраф за використання додаткових змінних. Кращою серед декількох ARIMA – моделей вважається модель з меншою величиною AIC, BIC.

З допомогою отриманої моделі можна побудувати точний і інтервальний прогнози на K кроків вперед. Для оцінки точності прогнозу використовуються стандартні критерії [3]:

- середня абсолютна процентна похибка (MAPE):

$$MAPE = \frac{100}{K}\sum_{i=1}^{K}\left|\frac{x_i - \hat{x}_i}{x_i}\right|,\% , \tag{8}$$

де x_i - спостережуване значення; $\hat{x}_i$ - значення прогнозу; K – інтервал прогнозу. Якщо MAPE<10%, то прогноз реалізований з високою точністю, при 10%<MAPE<20% - прогноз добрий, при 20%<MAPE<50% - прогноз задовільний, при MAPE>50% - прогноз незадовільний.

- середня процентна похибка (MPE):

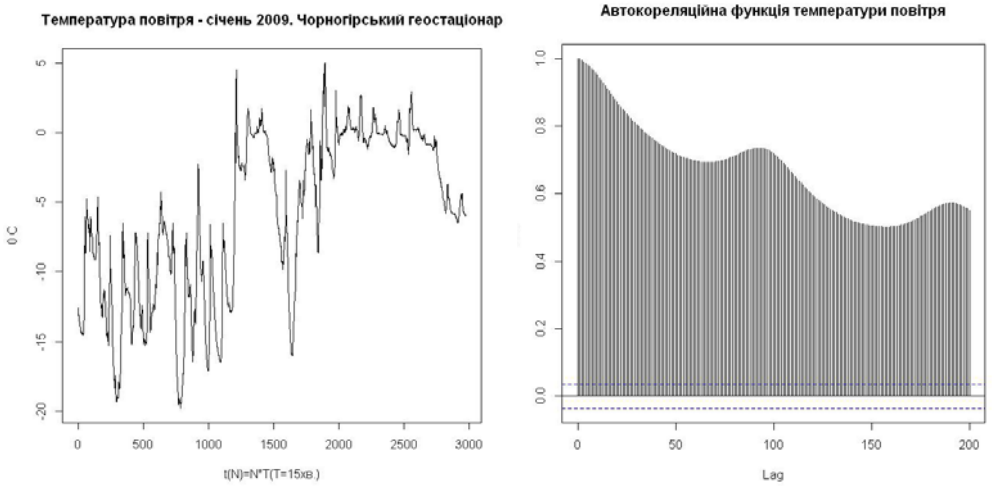
$$MPE = \frac{100}{K}\sum_{i=1}^{K}\frac{x_i - \hat{x}_i}{x_i},\% , \tag{9}$$

яка дозволяє визначити зміщення отриманого прогнозу. Якщо отримана модель є незміщеною, то MPE < 5%. Якщо в результаті розрахунків отримується велике від'ємне значення, то модель є з послідовним переоцінюванням. Якщо ж отримано велике додатне число, то модель - з послідовним недооцінюванням.

Алгоритм Бокса – Дженкінса дозволяє виконувати достатньо точний короткочасний прогноз. Але необхідно відмітити, що не існує простого способу корекції параметрів ARIMA моделі для нових даних. Модель потрібно періодично повністю перебудовувати або вибирати зовсім нову модель.

Побудова ARIMA моделі для прогнозування метеорологічних даних виконувалася з допомогою мови програмування R (версія 3.01) [2]. Також при прогнозуванні використовувалися функції, які входять до складу статистичного пакету для прогнозування (package 'forecast') [4, 5]. Для реалізації ARIMA моделі були взяті дані температури повітря за січень місяць 2009 р на Чорногірському геостаціонарі. Це дозволило отримувати ARIMA модель без врахування сезонної складової. Температура повітря вимірювалися метеостанцією з інтервалом $t = 15$ хв. Довжина досліджуваних часових рядів за даний місяць складає n=2976. Аналогічний підхід використовується і для прогнозування тиску повітря, швидкості і напрямку вітру. Прогноз виконувався на 6 год., 12 год. та одну добу.

Графік температури повітря за січень місяць 2009 р. приведено на рис. 1,а, де час вимірювання визначається як $t_N = N*15$, хв. Середня температура повітря за місяць склала $-5{,}9^0$ С. Вигляд ACF для температури приведено на рис. 1,б. Для побудови графіків ACF і PACF та отримання їх значень використовувалися функції "acf()" і "pacf()" в R. Для отримання фактичних значень ACF і PACF необхідно встановити "plot=FALSE" в функціях "acf()" і "pacf()".



а) б)

Рис. 1. а) - графік температури повітря за січень 2009 р. на Чорногірському геостаціонарі; б) – ACF температури повітря

З графіка температури повітря видно, що часовий ряд є нестаціонарний (дані не прив'язані до якогось постійного рівня). Також з графіка ACF температури повітря видно, що вона спадає повільно. Тому знаходимо перші різниці, які видаляють тренд з початкових даних. Диференціювання часових рядів виконуємо за допомогою функції "diff ()" в R. Отриманий часовий ряд (рис. 2) представляє собою коливання навколо нульового рівня, а ACF (рис. 3, а) та PACF (рис. 3, б) для отриманого ряду перших різниць швидко загасають. Середнє значення перших різниць температури повітря за місяць складає $0{,}002^0$ С.

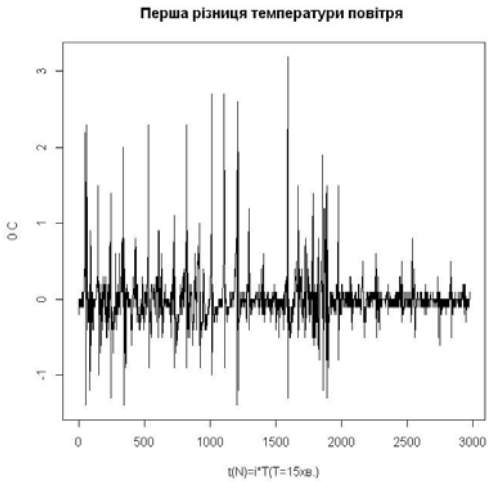
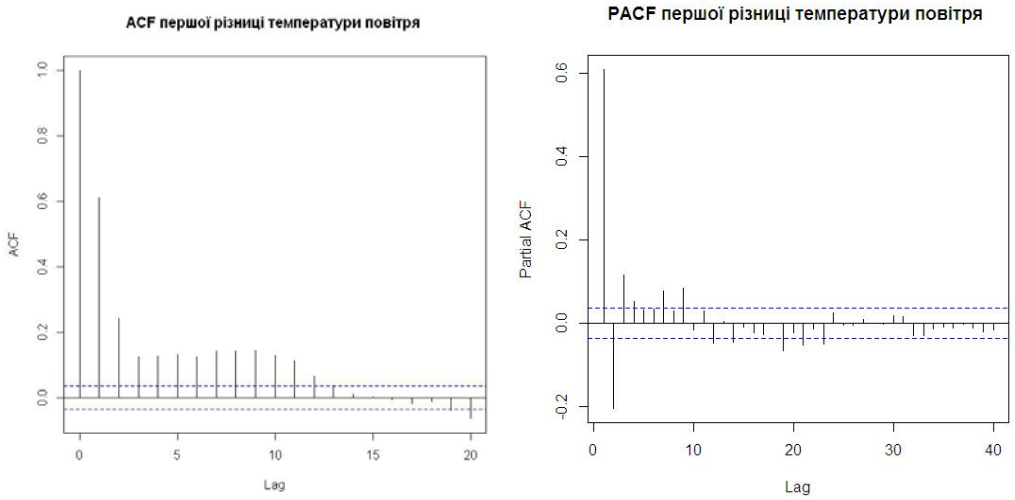


Рис. 2. Графік перших різниць температури повітря за січень 2009 р.



а) б)

Рис. 3. а) - ACF перших різниць температури повітря;
б) – PACF перших різниць температури повітря

Перші 20 значень ACF і PACF перших різниць температури повітря приведені в табл. 1.

Таблиця 1.
Значення ACF і PACF перших різниць температури повітря

№ лагу	0	1	2	3	4	5
ACF	1,000	0,610	0,244	0,125	0,126	0,130
PACF		0,610	-0,204	0,117	0,052	0,033
№ лагу	6	7	8	9	10	11
ACF	0,121	0,138	0,146	0,166	0,140	0,110
PACF	0,035	0,077	0,029	0,084	-0,016	0,030
№ лагу	12	13	14	15	16	17
ACF	0,059	0,032	0,008	-0,004	-0,006	-0,016
PACF	-0,048	0,006	-0,045	-0,008	-0,023	-0,028

Отримуємо модель виду ARIMA(p,1,q). Виконуємо оцінку параметрів p і q моделі ARMA(p,q), яка складається з моделей AR(p) і MA(q). Для цього найпростіше скористатися PACF і ACF, відповідно. Якщо вибіркова ACF швидко відсікається, а PACF експоненціально прямує до нуля, то в моделі повинні бути присутні доданки MA(q). Якщо ж вибіркова PACF швидко відсікається, а ACF прямує до нуля, то в моделі повинні бути присутні доданки AR(p). У випадку, якщо ACF і PACF прямують до нуля, то в модель включаються доданки обох типів. Порядок моделі AR(p) відповідає номеру останнього ненульового коефіцієнта PACF, а моделі MA(q) - номеру останнього ненульового коефіцієнта ACF.

Вважатимемо, що значення ACF (рис. 3, а) та PACF (рис. 3, б) відповідно після 13-го і 4-го лагів рівні нулю. Вибір оптимальної моделі серед ряду моделей ARMA(p,q) (p≤4, q≤13) виконувався методом перебору, використовуючи інформаційні критерії AIC,

BIC та очікувану вибіркову дисперсію $\hat{\sigma}^2$. Перевірка адекватності моделі виконувалася з допомогою Q – статистика Льюнга – Бокса та аналізу отриманого ряду залишків на подібність до ″білого″ шуму. Зокрема для ряду залишків температури повітря вираховувалося значення p-value та середнє значення $\overline{m}$. Результати вибору оптимальної моделі часового ряду температури повітря зведені в табл. 2.

Обчислення інформаційних критеріїв AIC, BIC та очікуваної вибіркової дисперсії $\hat{\sigma}^2$ виконувалося за допомогою функції "arima ()", а Q – статистики Льюнга – Бокса – з допомогою функції "Box.test ()" в R.

Таблиця 2.

Результати вибору оптимальної моделі температури повітря

		q=1	*q*=2	*q*=3	*q*=4	*q*=5	*q*=6	...	*q*=13
p=1	AIC	108,13	110,75	77,17	78,63	69,96	71,96	...	44,31
	BIC	126,92	134,74	107,16	114,62	111,94	119,94	...	134,28
	$\hat{\sigma}^2$	0,0606	0,0606	0,0598	0,0598	0,0596	0,0596	...	0,0588
	p-val	3,5E-12	2,5E-12	1,56E-5	6,89E-5	0,13E-3	0,13E-3	...	0.912
	$\overline{m}$	0,99E-3	0,1E-2	0,73E-3	0,73E-3	0,72E-3	0,72E-3	...	0,71E-3
p=2	AIC	110,87	110,00	78,88	70,77	63,41	65,05	...	40,66
	BIC	134,86	141,99	114,87	112,76	111,39	119,03	...	136,63
	$\hat{\sigma}^2$	0,0606	0,0606	0,0598	0,0596	0,0595	0,0594	...	0,0583
	p-val	3,1E-12	3,9E-12	1,77E-5	0,15E-2	0,11E-1	0,0095	...	0,9942
	$\overline{m}$	0,10E-2	0,10E-2	0,73E-3	0,73E-3	0,73E-3	0,73E-3	...	0,71E-3
p=3	AIC	71,00	71,6	48,94	63,75	36,06	36,88	...	48,10
	BIC	100,99	107,58	90,92	111,73	90,04	96,86	...	150,07
	$\hat{\sigma}^2$	0,0597	0,0597	0,0592	0,0594	0,0589	0,0588	...	0,0588
	p-val	7,47E-5	6,40E-5	0,75E-3	0,0143	0,2170	0,2658	...	0,9247
	$\overline{m}$	0,73E-3	0,72E-3	0,73E-3	0,73E-3	0,76E-3	0,79E-3	...	0,70E-3
p=4	AIC	71,81	66,95	68,88	65,27	36,83	23,12	...	48,47
	BIC	107,77	108,93	116,87	119,25	96,81	89,09	...	156,44
	$\hat{\sigma}^2$	0,0597	0,0596	0,0595	0,0595	0,0588	0,0585	...	0,0583
	p-val	6,70E-5	0,78E-3	0,85E-3	0,0172	0,2640	0,9407	...	0,8927
	$\overline{m}$	0,72E-3	0,72E-3	0,72E-3	0,74E-3	0,79E-3	0,83E-3	...	0,71E-3

Оптимальна модель часового ряду температури повітря (табл. 2) - ARIMA(4,1,6). Для неї отримано найменші значення інформаційних критеріїв AIC=23,12 та BIC=89,09. Перевірка на адекватність цієї моделі підтверджує, що часовий ряд залишків температури повітря з ймовірністю *p*=0,94 подібний на ″білий″ шум. Гістограма часового ряду залишків разом із згенерованим нормальним розподілом приведені на рис. 4.

Коефіцієнти ARIMA(4,1,6) моделі:

$\alpha_1 = 0.0706;\ \alpha_2 = 1.5985;\ \alpha_3 = 0.0216;\ \alpha_4 = -0.8239.$

$\beta_1 = 0.6608;\ \beta_2 = -1.4071;\ \beta_3 = -1.1531;\ \beta_4 = 0.4748;\ \beta_5 = 0.6058;\ \beta_6 = 0.1768.$

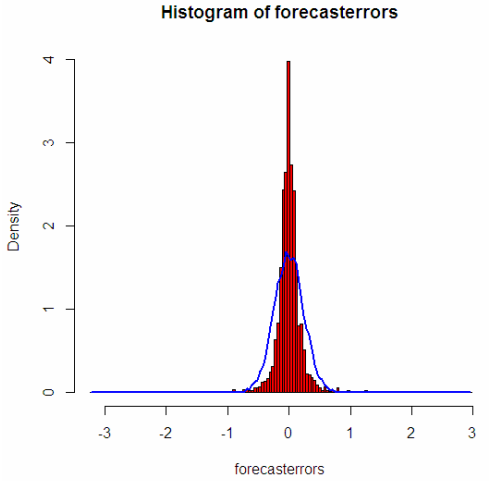


Рис. 4. Вигляд гістограми часового ряду залишків температури повітря

Прогнозування часового ряду температури повітря виконувалося з допомогою функції ″forecast.Arima ()″ в R. Результати короткочасного прогнозу температури на 24 часових інтервали (t = 6 год.) і виміряні метеостанцією значення температури приведені в табл. 3.

Таблиця 3.

Виміряні значення температури повітря та їх прогноз для h=24 (t = 6 год)

i	1	2	3	4	5	6	7	8
$x(t_i)$	-5,90	-5,90	-6,00	-6.20	-6.20	-6.40	-6.50	-6.50
$\overline{x}(t_i)$	-5.89	-5.90	-5.91	-5.93	-5.94	-5.96	-5.97	-5.98
i	9	10	11	12	13	14	15	16
$x(t_i)$	-6.40	-6.40	-6.30	-6.30	-6.30	-6.40	-6.40	-6.50
$\overline{x}(t_i)$	-5.99	-6.00	-6.01	-6.01	-6.02	-6.01	-6.01	-6.01
i	17	18	19	20	21	22	23	24
$x(t_i)$	-6.60	-6.70	-6.70	-6.60	-6.50	-6.50	-6.50	-6.50
$\overline{x}(t_i)$	-6.01	-6.00	-5.99	-5.99	-5.98	-5.97	-5.96	-5.96

1. *Бокс Дж.* Анализ временных рядов. Прогноз и управление: Вып. 1 (Пер. с англ. А. А. Левшина) / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. – 406 с.
2. *Venables W. N., Smith D. M.* An Introduction to R [Електронний ресурс]. Режим доступу: http://www.e-booksdirectory.com/details.php?ebook=1791
3. *Льюис К.Д.* Методы прогнозирования экономических показателей / Пер. с англ. и предисл. Е.З. Демиденко. М.: Финансы и статистика. (1986). 133 с.
4. Package ‘forecast’ [Електронний ресурс]. Режим доступу: http://cran.r-project.org/web/packages/forecast/forecast.pdf

5. *R. J. Hyndman, Y. Khandakar.* Automatic Time Series Forecasting: The forecast Package for R [Електронний ресурс]. Режим доступу: http://www.jstatsoft.org/v27/i03/paper

BUILDING ARIMA TIME SERIES MODELS FOR WEATHER DATA PREDICTING USING R PROGRAMMING LANGUAGE

O. Dzendzelyuk, L. Kostiv, V. Rabyk

Ivan Franko National University of Lviv, Faculty of Electronics
Tarnavskogo Str. 107, UA - 79017 Lviv, Ukraine
RabykV@ukr.net

The algorithm of ARIMA time series model construction based on the Box-Jenkins methodology is reviewed. The description of the criteria used for the verification of time series stationary mode, selection of the optimal ARIMA model and its reliabiliy testing is discussed. The main attention is focused on the implementation of this algorithm in R programming language. As an example, the implementation of short-term forecasting of ambient temperature in Chornogirska geographical station was done.

Keywords: time series, autocorrelation function, partial autocorrelation function, ARIMA model, forecasting, R programming language.

ПОСТРОЕНИЕ ARIMA МОДЕЛИ ВРЕМЕННЫХ РЯДОВ ДЛЯ ПРОГНОЗИРОВАНИЯ МЕТЕОДАННЫХ НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ R

О. Дзендзелюк, Л. Костив, В. Рабык

Львовский национальный университет имени Ивана Франко
ул. ген. Тарнавского, 107, 79017, Львов, Украина
RabykV@ukr.net

Рассмотрен алгоритм построения ARIMA модели временных рядов на основе методологии Бокса - Дженкинса. Приведено описание критериев, используемых для проверки стационарности временных рядов, выбора оптимальной ARIMA модели и проверки ее адекватности. Основное внимание акцентировано на реализации этого алгоритма на языке программирования R. Как пример приведено реализацию кратковременного прогнозирования температуры воздуха на Черногорском географическом стационаре.

Ключевые слова: временные ряды, автокорреляционная функция, частная автокорреляционная функция, ARIMA - модель, прогнозирование, статистический пакет R.

Стаття надійшла до редколегії 24.06.2013
Прийнята до друку 10.07.2013