Аналіз власного часового ряду
з дисципліни
”Нелінійні часові ряди”
Студента 2 курсу магістратугри
групи ”Статистика”

Горбунов Даніел

23 грудня 2022 р.

Зміст

1 Вступ.

Дана робота присвячена аналізу часового ряду, що має ефект сезонності. В аналізі пропонується дослідити адитивну модель $X = T+W$, де T – невипадкова функція та W – стаціонарний процес. Побудовано прогноз за цією моделлю на декілька днів вперед.

2 Хід роботи.

2.1 Опис даних.

Досліджуються дані [1] про середню температуру за добу (в шкалі Фаренгейт) у місті Київ, з 1995 по 2020 роки. Першочергово потрібно було провести обробку вхідних даних для подальної роботи. З цього можна перелічити таке, як: "заміщення" невипадкових викидів у часовому ряді, які могли бути спричинені технічними похибками; переведення зеначення температури в іншу шкалу (з Фаренгейт у шкалу Цельсія).

```
## Підготовка даних до роботи
data <- read.csv('city_temperature.csv')

# Обираємо дані про добову температуру міст України
data.ukraine <- data[data$Country == "Ukraine", ]
print(unique(data.ukraine$City))
# Обираємо дані про добову температуру в м. Київ
data.kiev <- data.ukraine[
  data.ukraine$City == "Kiev", c("Day", "Month", "Year", "AvgTemperature")
  ]

# Необхідно для переведення числового вектора в об'єкт часового ряду
# Крайні роки та місяці
min.year <- min(data.kiev$Year)
min.month <- min(data.kiev$Month)
max.year <- max(data.kiev$Year)
max.month <- max(data.kiev$Month)

# Будуємо графік середньої температури за добу в обраному місті
plot(data.kiev$AvgTemperature, type='l')
```

Далі візуально покажемо що з себе представляють початкові дані про температуру.

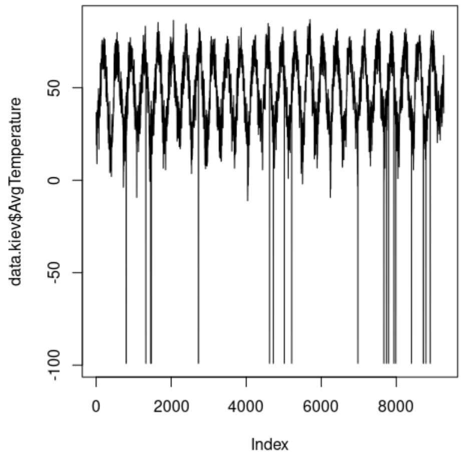


Рис. 1: Повні дані про середню температуру за добу в м. Київ. Період: 1997-2020 роки.

Добре видно, що траєкторія процесу має періодичну (сезонну) компоненту. Також на очі трапляються технічні викиди: на графіку вони виглядають як сильні випадання траєкторії донизу. Кожен з цих викидів має спільне значення рівне -100 градусам за Фарентгейт. Якщо перевести цю в шкалу Цельсія, то вихолить -74 градуси. За останній десяток перебування у цьому місті, автор роботи не помічав подібних кліматичних аномалій. Таким чином, треба ці технічні похибки виправити на більш реальні: це можна зробими заміщенням попереднього значення локальним середнім. Але це будемо робити після перетворення часового ряду до переварюваного вигляду.

```
## Перетворюємо вектор в об'єкт часового ряду, де час вимірювання -- 1 доба
data.kiev.ts <- ts(
  data=data.kiev$AvgTemperature,
  start=c(min.year, min.month),
  end=c(max.year, max.month),
  frequency=365
  )
# Переводимо з F у C
data.kiev.ts <- (data.kiev.ts - 32) * 5/9

sampled.ts <- window(data.kiev.ts,
                     start=c(2005, 1),
                     end=c(2010, 12),
                     frequency=365)
n <- length(sampled.ts)

plot(sampled.ts,
     ylab="Temperature",
     main="Щоденна середня температура у м.Київ, 2005-2010")
grid()
```

Зауважимо, що переведення в об'єкт часового ряду трохи з'їло значення справа, аби щорічна частота співпадала повністю по даним. Досліджувати часовий ряд будемо на проміжку з 2005 по 2010 роки, а прогнозувати на майбутні дати. Зобразимо отриману вибірку часового ряду:

Рис. 2: Вибіркові дані про середню температуру за добу в м. Київ. Період: 2005-2010 роки.

Як було сказано раніше, заміщення викидів робиться локальними усередненнями:

$$\hat{X}_h(t_k) = \frac{\sum_{j=\overline{1,n}, j \neq k} 1\{|t_j - t_k| < h\} X(t_j)}{\sum_{j=\overline{1,n}, j \neq k} 1\{|t_j - t_k| < h\}},\ h > 0$$

Ширину вікна h поклали рівним 7 – цього було достатньо для отримання більш адекватних значень у ті дні, де дані про температуру "викривлені".

```
> out.mask <- abs(sampled.ts - mean(sampled.ts)) > 4 * sd(sampled.ts)
> print(sampled.ts[out.mask])
[1] -72.77778 -72.77778 -72.77778 -72.77778
> # Заміна локальним середнім
> idx <- (1:n)
> eps <- 7
> for(j in idx[out.mask])
+ {
+     x0 <- sampled.ts[j]
+     eps.hood <- abs(idx[-j] - j) < eps
+     sampled.ts[j] <- sum(eps.hood * sampled.ts[-j]) / sum(eps.hood[-j])
+ }
> print(sampled.ts[out.mask])
[1] 22.0555556  0.9848485 11.2676768 11.0303030
```

Залишається відобразити остаточний вигляд вибірки часового ряду, з якою доведеться працювати.

Рис. 3: Вибіркові дані про середню температуру за добу в м. Київ. Період: 2005-2010 роки.

Тепер можна робити певну підгонку. Спочатку підбираємо параметри в адитивній моделі, далі – в мультиплікативній моделі ARIMA з сезонністю.

2.2 Розгляд адитивної моделі часового ряду.

Припустимо, що спостережуваний часовий ряд можна змоделювати наступним чином:

$$X(t) = a(t) + W(t),\ t = 2005, 2006, \dots, 2010$$

де $a(t)$ – невипадкова функція, яку потрібно підібрати; $W(t)$ – стаціонарний процес. Найскладніший етап на цьому етапі є вгадування тренду $a(t)$. Зробимо це з наступних візуальних міркувань.

На Рисунку (2.1) можна спостерігати періодичну поведінку, яка схожа у косинуса чи синуса. Тобто можна розглядати функції вигляду $C\sin(Ax + B)$, де сталі A, B, C можна оцінити методом найменших квадратів. Окрім періодичної компоненти, можна побачити підйом траєкторії, який можна було б підігнати деяким кубічним многочленом, параметри якого теж можна підібрати за МНК.

Спочатку зробимо підгонку поліноміальної частини тренду, використовуючи МНК.

```
poly.mean <- function(t, A, B, C, D)
{
  t0 <- t - 2005
  A + B * t0 + C * t0^2 + D * t0^3
}

sampled.values <- as.numeric(sampled.ts)
sampled.time <- time(sampled.ts)
data.ts <- data.frame(cbind(sampled.time, sampled.ts))
colnames(data.ts) <- c("X", "Y")

poly.ols <- lm(Y ~ I(X-2005) + I((X-2005)^2) + I((X-2005)^3), data=data.ts)
```

Подивимося на звіт з підгонки.

```
> summary(poly.ols)
Call:
lm(formula = Y ~ I(X - 2005) + I((X - 2005)^2) + I((X - 2005)^3),
    data = data.ts)

Residuals:
    Min      1Q  Median      3Q     Max
-32.601  -7.742   0.137   8.490  20.531

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.32100    0.90534   5.877 4.94e-09 ***
I(X - 2005)      4.37250    1.55912   2.804  0.00509 **
I((X - 2005)^2) -1.27241    0.72038  -1.766  0.07751 .
I((X - 2005)^3)  0.10924    0.09414   1.160  0.24603
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.721 on 1833 degrees of freedom
Multiple R-squared:  0.0129,        Adjusted R-squared:  0.01128
F-statistic: 7.985 on 3 and 1833 DF,  p-value: 2.741e-05
```

В принципі нам відомо, що часовий ряд не моделюється виключно поліномом, тому можна було очікувати паршивий коефіцієнт детермінації. Незважаючи на результати тесту Стьюдента на коефіцієнти, ми будемо використовувати всі коефіцієнти при кожному порядку. Втім, тест Фішера вбачає залежність по часу – але ця інформація не настільки важлива, бо ми самі знаємо емпірично про форму залежності. Віднімаємо поліноміальний тренд, отримуючи сталість поведінки траєкторії (не підіймається чи опускається з плином часу). Маємо наступну картинку:

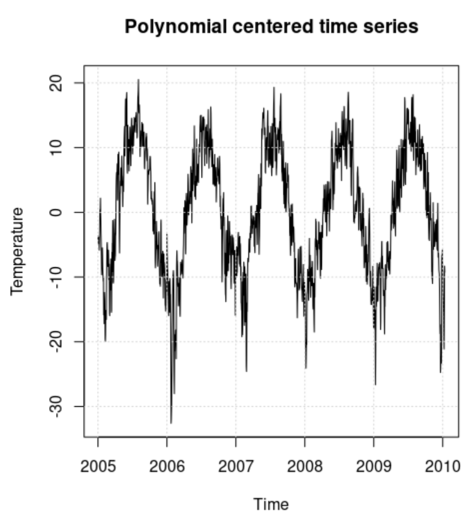


Рис. 4: Поліноміально центрований часовий ряд.

Видно, що процес не починається з нуля, тому до моделі з синусом введемо адитивну константу, яку теж потрібно оцінити. Таким чином, робимо підгонку параметрів за МНК в функції зв'язку вигляду:

$$g(x) = A\sin(Bt + C) + D$$

```
trig.mean <- function(t, A, B, C, D)
{
  A * sin(B * t + C) + D
}

trig.ols <- nls(Y1 ~ A * sin(B * X + C) + D,
                data=data.ts,
                start=list(A=15, B=2*pi, C=4.5, D=-3))
coef.trig.ols <- coef(trig.ols)

trig.mean.ols <- function(t)
{
  trig.mean(t,
            A=coef.trig.ols["A"],
            B=coef.trig.ols["B"],
            C=coef.trig.ols["C"],
            D=coef.trig.ols["D"])
}

curve(trig.mean.ols(x), col="red", lty=2, lwd=2, add=T)
```

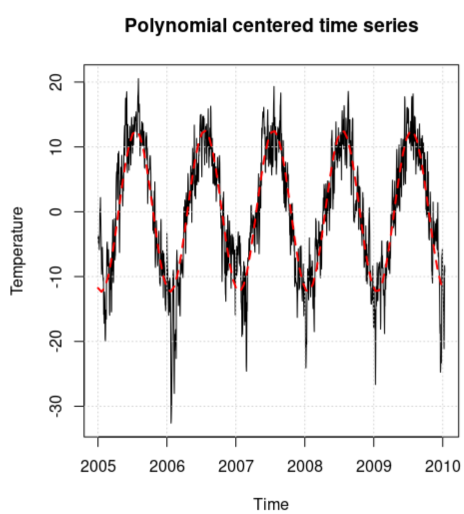


Рис. 5: Поліноміально центрований часовий ряд. Червона крива – підгонка періодичної компоненти.

Переглянемо звіт з підгонки моделі.

```
> summary(trig.ols)
Formula: Y1 ~ A * sin(B * X + C) + D

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 12.428512   0.134490  92.412   <2e-16 ***
B  6.286438   0.007636 823.232   <2e-16 ***
C -2.115185  15.329840  -0.138    0.890
D  0.085588   0.096163   0.890    0.374
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.074 on 1833 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 5.393e-07
```

Використовуючи отримані підгонки за двома моделями, комібнуємо результати та підганяємо параметри у спрощеній поліноміальній моделі тренду з періодичністю:

$$A \sin(B * t) + C + D * (t - 2005) + E * (t - 2005)^2$$

Результат підгонки за МНК має такий вигляд:

```
Formula: Y ~ A * sin(B * X) + C + D * (X - 2005) + E * (X - 2005)^2

Parameters:
    Estimate Std. Error    t value Pr(>|t|)
A -1.250e+01  1.324e-01 -9.442e+01  < 2e-16 ***
B  6.287e+00  5.358e-06  1.173e+06  < 2e-16 ***
C  7.753e+00  2.822e-01  2.747e+01  < 2e-16 ***
D  1.139e+00  2.585e-01  4.408e+00 1.11e-05 ***
E -1.742e-01  4.972e-02 -3.505e+00 0.000468 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.016 on 1832 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 2.9e-09
```

Підгонка виглядає непоганою. Варто дослідити залишки прогнозу моделі.

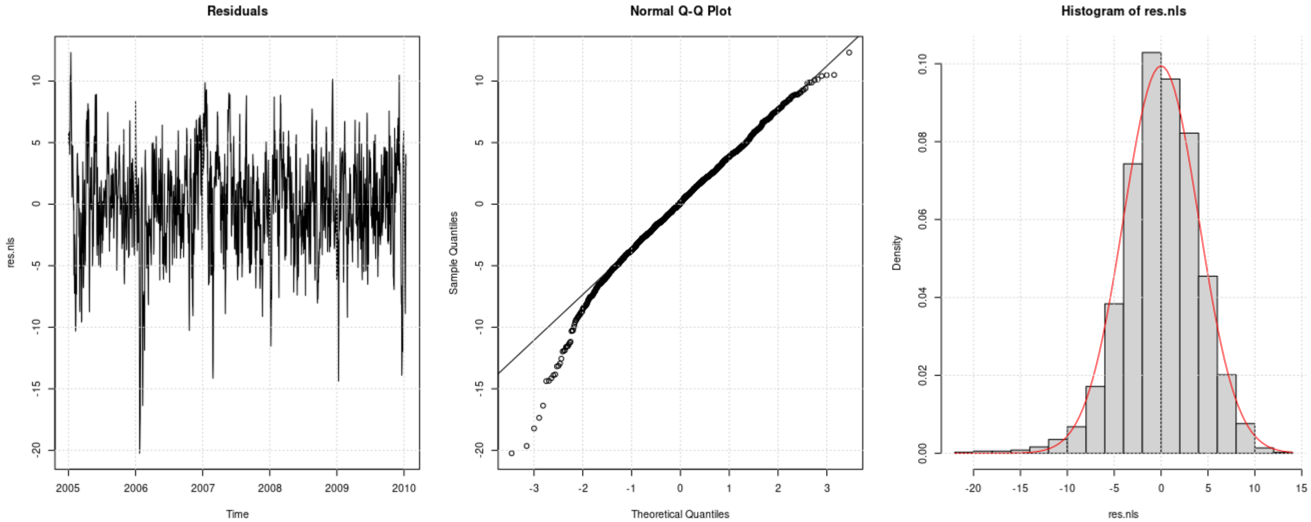


Рис. 6: Процес заликів: траєкторія, QQ-діаграма відносно $N(\mu, \sigma^2)$, та гістограма.

На рисунках можна побачити, що залишки прогнозу не мають приблизно гауссів розподіл завдяки великим відхилам донизу у деякі моменти часу. Відповідно емпіричний розподіл залишків мають важчий лівий хвіст у порівнянні з правим. Але, взагалі кажучи, форма розподілу нагадує гауссів розподіл.

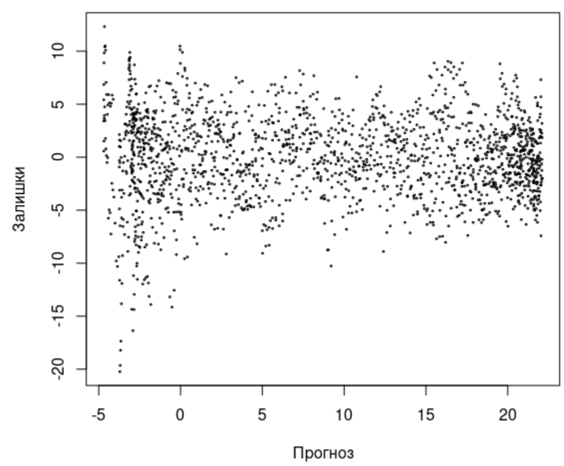


Рис. 7: Діаграма "прогноз-залишки".

Можна побачити більшу варіативність при прогнозуванні низьких температур, у порівнянні з прогнозом на інші тепліші температури, холодний прогноз виходить менш стійким. З іншого боку, як можна побачити на графіку траєкторії часового ряду, в 2006 році була рекордна температура взимку. Надалі модель добре "лягає" на інші сезони, в рамках спостережуваної вибірки.

Перевіримо, чи утворюють залишки стаціонарний процес. Побудуємо діаграми автокореляції та частинної авткореляції:

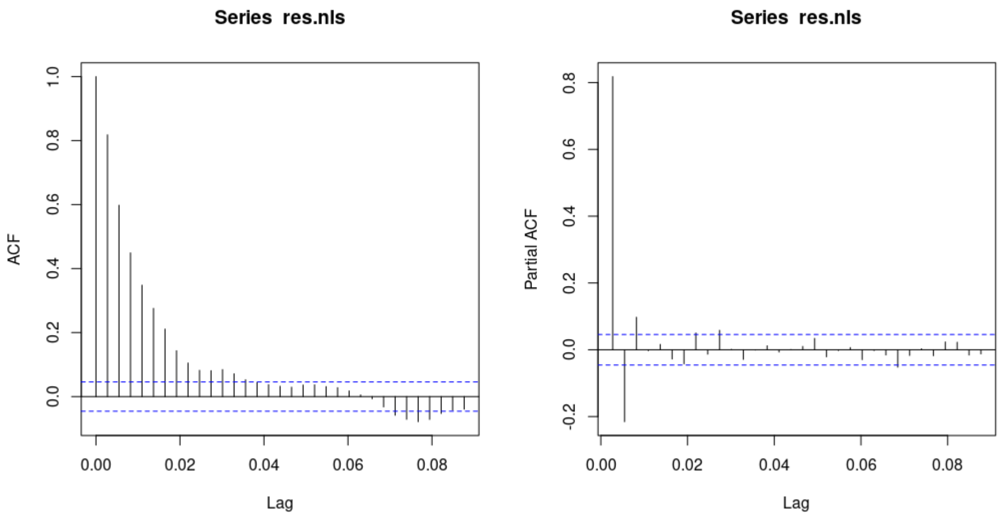


Рис. 8: Діаграма автокореляції та частинної автокореляції процесу залишків.

Зробимо підгонку параметрів в ARMA(p, q). Оберемо ті порядки p та q перебором. Максимальними будуть ті порядки, коли автокореляція та частинна автокореляція приблизно менша за 0,05. Наприклад, p обираємо за поведінкою автокореляції, відповідно $p_{max} = 14$ (ігноруючи нижнє коливання для деякого лага). Для q обираємо за поведінкою частинної автокореляції $q_{max} = 3$. Для пар (p, q), таких, що $0 \leq p + q \leq \min(p_{max}, q_{max})$ підганяємо параметри моделі ARMA та обчислюємо відповідний інформаційний критерій.

```
> p.max <- 14; q.max <- 3
> AIC.matrix <- matrix(nrow=p.max+1, ncol=q.max+1)
> for(p in 0:p.max)
+ {
+   for(q in 0:q.max)
+   {
+     if((0 <= p + q) & (p + q <= min(p.max, q.max)))
+     {
+       arma.pq <- arima(res.nls, order=c(p, 0, q))
+       AIC.matrix[p+1,q+1] <- AIC(arma.pq)
+     }
+   }
+ }
> AIC.matrix
           [,1]     [,2]     [,3]     [,4]
 [1,] 10319.913 8940.521 8479.442 8319.198
 [2,]  8285.921 8185.963 8187.005       NA
 [3,]  8201.050 8187.204       NA       NA
 [4,]  8185.780       NA       NA       NA
 [5,]        NA       NA       NA       NA
 [6,]       ...      ...      ...      ...
```

Найменше значення інформаційного критерія Акаїке досягається при $p = 3$ та $q = 0$.

Покажемо підігнані значення коефіцієнтів моделі:

```
> summary(res.nls.arima)

Call:
arima(x = res.nls, order = c(3, 0, 0))

Coefficients:
         ar1      ar2     ar3  intercept
      1.0154  -0.3113  0.0969     0.0140
s.e.  0.0232   0.0324  0.0233     0.2619

sigma^2 estimated as 5.013:  log likelihood = -4087.89,  aic = 8185.78
```

Побачимо, що обернені корені відповідного AR-многочлена лежать в одиничному колі:

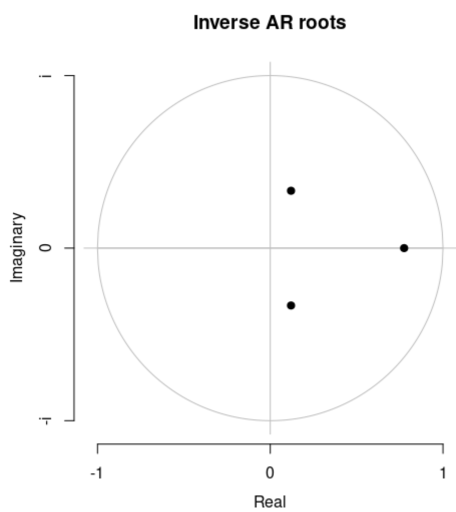


Рис. 9: Зображення обернених коренів оціненого AR-многочлена на комплексній площині.

Але ж з теорії відомо, що обернені корені AR-многочлена є коренями характеристичного многочлена. Якщо характеристичний многочлен деякого різницевого рівняння має всі корені за модулем менші одиниці, тоді саме рівняння є асимптотично стійким, і зокрема існуватиме стаціонарний процес, який задовольнятиме рівняння ARMA. Але ж то оцінена форма AR-многочлена, для підкрілпення припущення про стаціонарність, можна скористатися KPSS тестом:

```
> kpss.test.res <- kpss.test(res.nls)
Warning message:
In kpss.test(res.nls) : p-value greater than printed p-value
> kpss.test.res
        KPSS Test for Level Stationarity
data:  res.nls
KPSS Level = 0.077456, Truncation lag parameter = 8, p-value = 0.1
```

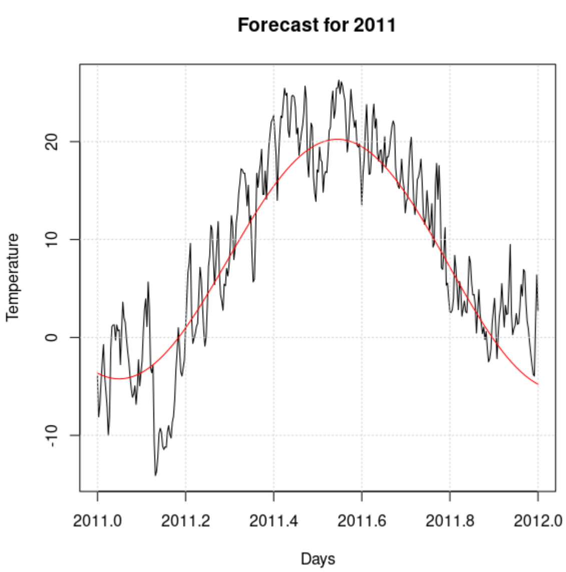
Тест має підстави для прийняття основної гіпотези про стаціонарність часового ряду.

2.3 Прогнозування.

Зробимо прогноз прогоди на один рік вперед, тобто на весь 2011 рік. Звісно це досить потужна заява, але якщо модель це дозолить?

```
## Прогноз погоди
to.forecast <- window(data.kiev.ts,
                      start=c(2011, 1),
                      end=c(2012, 1),
                      frequency=365)
times.for.forecast <- time(to.forecast)
data.ts.new <- data.frame(cbind(times.for.forecast, as.numeric(to.forecast)))
colnames(data.ts.new) <- c("X", "Y")

par(mfrow=c(1,1))
predicted.temp <- predict(full.nls, data.ts.new)
plot(times.for.forecast, as.numeric(to.forecast), type='l')
curve(full.func(x), col='red', add=T)
grid()
```



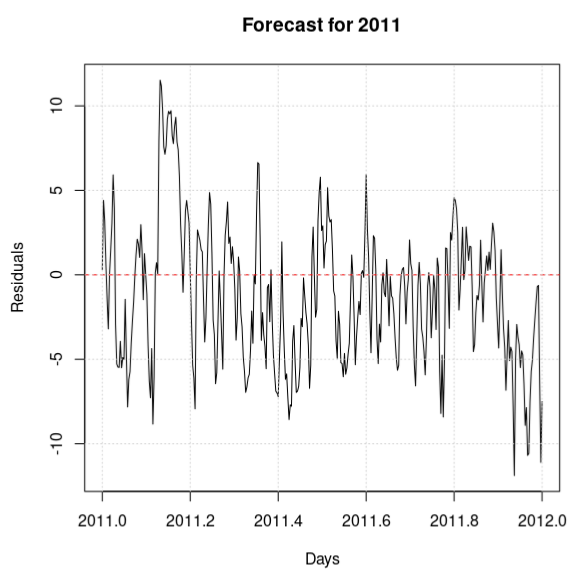


Рис. 10: Зліва: червона лінія – прогноз погоди протягом 2011 року. Справа – залишки.

Добре видно, що на 2011 рік, прогноз моделі вхоплює сезонний тренд, однак можна побачити, що на початок 2012 року помітне повільне зростання цього тренду.

3 Висновки.

Використання адитивної моделі часового ряду вийшло відносно адекватною ідеєю в розглянутому випадку. Запропоновано адитивну модель вигляду: $X(t) = a(t) + W(t)$, де $a(t) = A\sin(B*t) + C + D*(t-2005) + E*(t-2005)^2$ та $W(t)$ є процесом ARIMA$(3,0,0)$. Прогноз тренду на 2011 рік (тобто на один рік вперед) візуально вийшов непоганим. Дану модель можна розглядати як дешеву альтернативу мультиплікативної моделі ARIMA з сезонністю.

Література

[1] Дані про щоденну температуру у найбільших містах світу: https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities