УДК 004.6+004.94

Д.В. Легеза

СИСТЕМА МОНІТОРИНГУ ЗАХВОРЮВАНОСТІ НА ГОСТРІ РЕСПІРАТОРНІ ЗАХВОРЮВАННЯ В УКРАЇНІ

Based on the multivariate regression analysis, we construct a mathematical model which allows analyzing and predicting the ARD incidence on the territory of Ukraine. The obtained results indicate a high accuracy of our calculations and the model adequacy to real-life conditions.

Вступ

Щодня лікарі стикаються з проблемою передбачення кількості хворих на гострі респіраторні захворювання (ГРЗ). Від точності передбачення залежить, чи буде перевищений епідемічний поріг та чи слід готуватися до можливої епідемії. Наявність інформації про захворюваність на грип за кілька років [1] дає можливість якісно її проаналізувати та виділити певні закономірності в динаміці зміни кількості хворих на ГРЗ.

Для проблемної області, що розглядається, вже створено програмний комплекс EsculapPro, який хоч і розв'язує задачу індивідуальної діагностики конкретної особи, однак жодним чином не вирішує проблем, які запропоновано для аналізу. Крім цього програмного комплексу, створювалися й інші програми, але вони використовувалися лікарями лише з метою аналізу результатів випробовувань ліків і вакцин.

За наявності інформації про захворюваність по Україні за 9 років для її аналізу необхідно створити принципово новий підхід, якого не було у попередніх роботах, і з цієї точки зору робота є актуальною.

З метою розв'язання поставленої задачі необхідно створити математичну модель, на основі якої побудувати програмний комплекс, який дав би змогу оброблювати дані, аналізувати їх, виділяючи компоненти. На основі розділення даних можна виявити багато важливої інформації. Крім того, програмний комплекс дасть можливість виконувати короткотерміновий прогноз на один контрольний тиждень з метою передбачення епідеміологічної ситуації в країні.

Результати даної розробки можуть бути корисними Міністерству охорони здоров'я України та фахівцям як медичної, так і інформаційної галузей.

Постановка задачі

Метою роботи є створення математичної моделі, що якнайкраще описувала б експериментальні дані про захворюваність на ГРЗ. На основі вибраної математичної моделі потрібно створити програмний комплекс, який дасть змогу виконувати такі операції з обробки експериментальних даних:

1) довготерміновий аналіз даних – виділення основного тренду та інших компонент на основі даних за довгий період часу;

2) короткотермінове прогнозування – передбачення епідеміологічної ситуації на короткий термін (на 1 тиждень).

Проблема вибору способу для аналізу вихідних даних

У статті розглядаються дані про захворюваність на ГРЗ по 10 контрольним містам України, надані Міністерством охорони здоров'я України [1]. Оскільки ці дані за своїм змістом є числовими рядами, то існує кілька якісних методів їх обробки:

1) екстраполяційний аналіз;

2) факторний аналіз і аналіз числових рядів;

3) регресійно-кореляційний аналіз.

Основним критерієм вибору методу обробки було використання всіх наявних даних для їх подальшої обробки. На основі цього критерію було вибрано як базовий спосіб обробки даних регресійно-кореляційний аналіз.

Попередній аналіз вихідних практичних даних показав, що найкращим їх наближенням серед гладких функцій є відрізки тригонометричних рядів Фур'є вигляду $f(x) = A\sin(\omega x \pm \varphi) + C$ або $f(x) = A\sin(\omega x) + B\cos(\omega x) + C$. Періодичний характер вказаних даних можна прослідкувати на рис. 1.

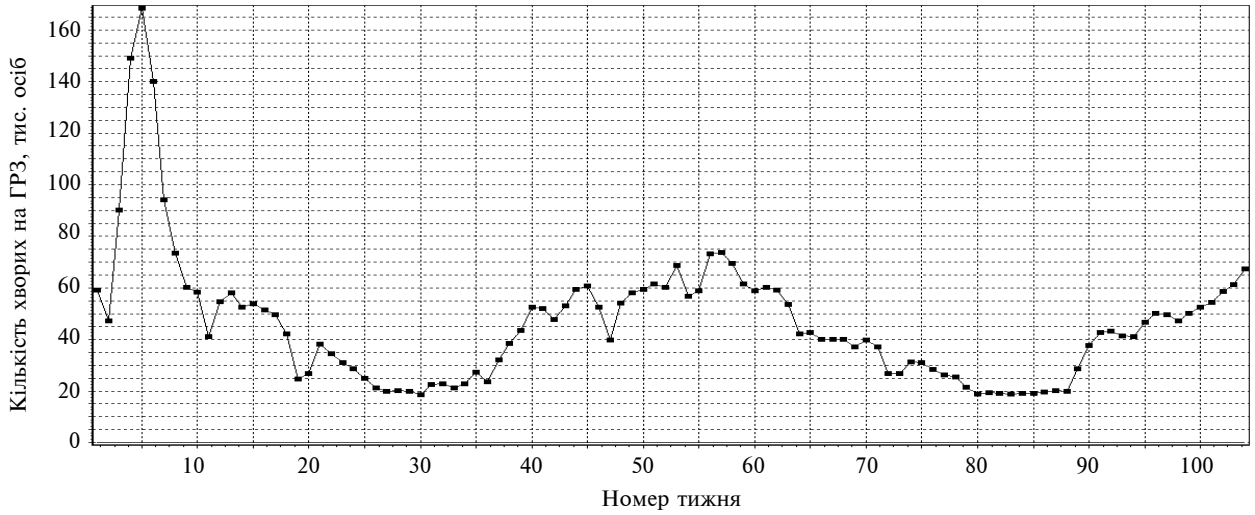

Рис. 1. Зведені дані про захворюваність на ГРЗ за 2 роки

Побудова математичної моделі для аналізу та прогнозування кількості хворих

В основі виконаного аналізу лежить багатофакторна побудова функцій, наближених до експериментальних даних у середньоквадратичному розумінні.

За основу для регресійного аналізу було вибрано функцію

$$f(x) = A_1 \sin(\omega x) + B_1 \cos(\omega x) + C_1, \qquad (1)$$

яка є відрізком ряду Фур'є [2, 3]. У цій функції невідомими параметрами є A_1, B_1 і C_1. При цьому ω є відомою сталою величиною.

Визначимо ω, виходячи з періодичності даних про захворюваність на ГРЗ. Період експериментальних даних, які розглядаються в роботі, дорівнює одному року, що слідує з графічного подання об'єднаних даних про захворюваність за два роки досліджень (рис. 1).

Отже, можемо визначити частоту ω:

$$\omega = 2\pi/T = 2\pi/52 = 0{,}1208. \qquad (2)$$

Період становить $T = 52$, оскільки в році 52 тижні.

В результаті підстановки (2) в (1) функція для регресійного аналізу набуває вигляду

$$f(x) = A_1 \sin(0{,}1208\,x) + B_1 \cos(0{,}1208\,x) + C_1. \quad (3)$$

У процедурі пошуку коефіцієнтів регресійного рівняння (3) було використано *гауссів метод Гаусса для формування нормальної системи рівнянь* [3]. Після виконання даної процедури коефіцієнти відшукуються на основі застосування *прямого метода Гаусса* до системи такого вигляду:

$$\begin{cases} n C_1 + A_1 \sum\limits_i x'_i + B_1 \sum\limits_i x''_i = \sum\limits_i y_i, \\ C_1 \sum\limits_i x'_i + A_1 \sum\limits_i (x'_i)^2 + B_1 \sum\limits_i x'_i\, x''_i = \sum\limits_i y_i\, x'_i, \\ C_1 \sum\limits_i x''_i + A_1 \sum\limits_i x'_i\, x''_i + B_1 \sum\limits_i (x''_i)^2 = \sum\limits_i y_i\, x''_i, \end{cases} \quad (4)$$

де $x'_i = \sin(\omega x_i)$, $x''_i = \cos(\omega x_i)$, y_i – емпіричне значення кількості хворих на x_i-й тиждень; A_1, B_1, C_1 – шукані значення коефіцієнтів функції (3); n – кількість експериментів.

Після знаходження коефіцієнтів регресійного рівняння на основі розв'язання системи рівнянь (4) та отримання вигляду функції основного тренду обчислюються значення функції (3) в усіх точках абсциси (яка є номером тижня). Крім того, обчислені значення відображаються на графіку.

Після цього з вихідних даних виділяється основний тренд. Ця процедура є відніманням від значень експериментальних вихідних даних відповідних підрахованих даних основного тренду.

Наведена вище процедура виділення гармонік з вихідних даних повторюється кілька разів з різними частотами ω.

Для визначення незалежності виділених факторів з вихідних даних було використано коефіцієнти кореляції, розраховані для кожного набору отриманих даних на основі формул [4, 5]:

$$M[y_1] = \frac{\sum y_1}{n-1},\ M[y_2] = \frac{\sum y_2}{n-1};$$

$$D[y_1] = \frac{\sum (y_1 - M[y_1])^2}{n-1},\ D[y_2] = \frac{\sum (y_2 - M[y_2])^2}{n-1};$$

$$k_{y_1 y_2} = \frac{\sum (y_1 - M[y_1])(y_2 - M[y_2])}{n-1},$$

$$r_{y_1 y_2} = \frac{k_{y_1 y_2}}{\sqrt{D[y_1]}\sqrt{D[y_2]}}. \qquad (5)$$

Формула (5) відображає власне коефіцієнт кореляції та визначає силу зв'язку між розрахунковими даними й може змінюватися в межах [–1;1].

Побудова програмного комплексу на основі математичної моделі

На основі розробленої математичної моделі створено програмний комплекс, який дає змогу здійснювати обробку даних двома основними методами:

- довготерміновим аналізом даних;
- короткотерміновим прогнозуванням захворюваності.

Довготерміновий аналіз даних дає можливість виділяти основні компоненти з вихідних даних. Після виділення компонент з'являється можливість графічного подання результатів розрахунків і можливість здійснення кореляційного аналізу, який є розрахунком сили зв'язку між кожною виділеною компонентою.

Короткострокове прогнозування дає змогу здійснювати передбачення захворюваності на ГРЗ на один тиждень наперед.

Для короткотермінового прогнозування як базовий метод обробки інформації було взято регресійно-кореляційний аналіз. Функція вибрана з класу поліноміальних і має такий вигляд:

$$f(x) = Ax^2 + Bx + C,$$

де A, B, C – невідомі параметри.

До вибору точок, по яких будується регресійна функція, ставляться такі умови.

1. Якщо $y_i < y_{i+1} < \ldots < y_{i+k} > y_{i+k+1}$, то для аналізу вибирається інтервал $[x_i; x_{i+k}]$, тобто весь інтервал, де експериментальні дані мають тенденцію до зростання.

2. Якщо $y_i > y_{i+1} > \ldots > y_{i+k} < y_{i+k+1}$, то для аналізу вибирається інтервал $[x_i; x_{i+k}]$, тобто весь інтервал, де експериментальні дані мають тенденцію до спадання.

3) Якщо існує такий інтервал, що $y_i > y_{i+1} < < y_{i+2} > \ldots > y_{i+k-1} < y_{i+k} > y_{i+k+1}$, тобто він має характерний "пилкоподібний" характер, то для аналізу беруться всі точки, що входять до даної структури, тобто $[x_i; x_{i+k+1}]$.

Тут y_i – значення, що відповідає x_j, тобто, по суті, це кількість хворих у конкретний j-й тиждень.

На основі вказаних умов і з використанням регресійного апарату виконуються розрахунки.

Програмний комплекс має такі складники:

1) підпрограма "Довготерміновий аналіз даних";

2) підпрограма "Короткотермінове прогнозування";

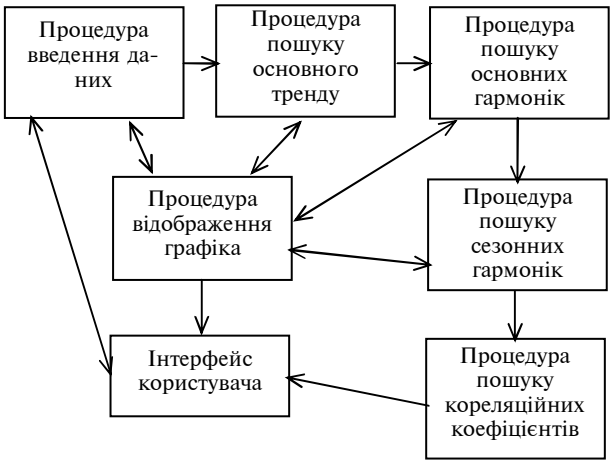

Рис. 2. Схема взаємодії складових частин підпрограми "Довготерміновий аналіз даних"

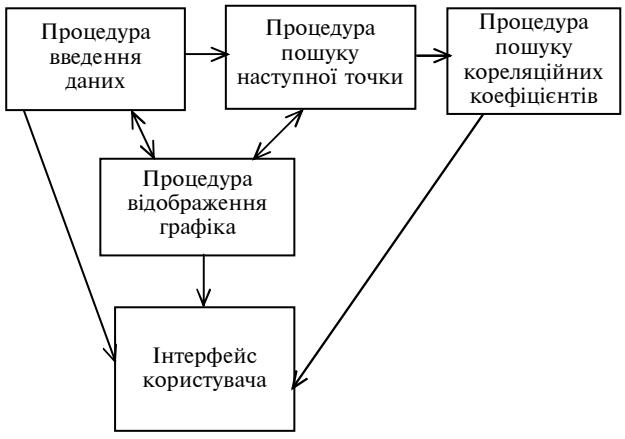

Рис. 3. Схема взаємодії складових частин підпрограми "Короткотермінове прогнозування "

Рис. 4. Структура компонентів програмного комплексу та їх взаємодії

3) інтерфейс користувача та підказки;

4) допоміжні модулі.

Схеми взаємодії складових частин підпрограми "Довготерміновий аналіз даних" і підпрограми "Короткотермінове прогнозування" подані на рис. 2 і 3 відповідно. На рис. 4 відображено структуру програмного комплексу та взаємодію його компонентів між собою.

Результати розрахунків

На основі вихідних даних було виконано розрахунки, результати яких наведено на рис. 5.

Програмний комплекс дає змогу маніпулювати результатами розрахунків таким чином, що можна окремо відобразити кожну компоненту на графіку, порівняти компоненти між собою, а також з вихідними даними, можна переглянути їх числові значення у вигляді розрахованих числових таблиць. Крім того, можна переглянути значення коефіцієнтів кореляції між компонентами.

Отримані результати чітко вказують на коректність гіпотези про вибір основної розрахункової функції у вигляді відрізку ряду Фур'є.

Крім того, показано, що відхилення розрахованих даних від вихідних не перевищує 1 %, що вказує на високу адекватність вибраної математичної моделі для обробки даних.

У процесі аналізу даних виділено три основні компоненти з вихідних даних: основний тренд, основні гармоніки і сезонні гармоніки.

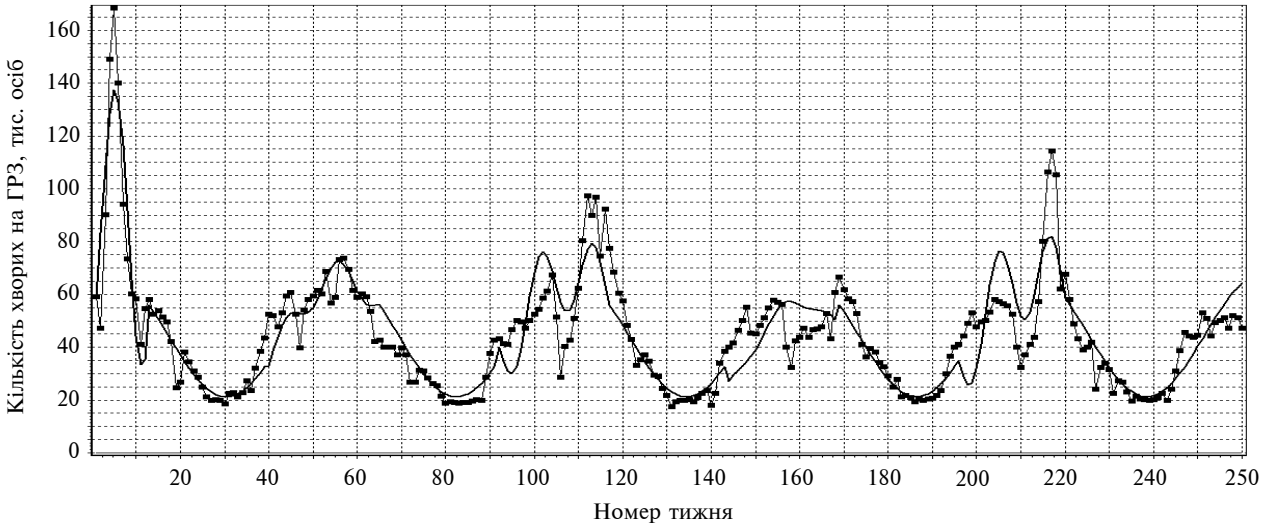

Рис. 5. Результати роботи алгоритму, що відображають вихідні та розраховані дані; ■ – дані реальної захворюваності ГРЗ; —— – розраховані дані на основі створеного алгоритму

Виділення сезонних гармонік зумовлене тим, що в інтервалі тиждень до кінця грудня–перший тиждень січня у багатьох людей вихідні й усі зайняті підготовкою до свят. Відповідно, кількість хворих у ці дні зменшується через те, що вони перебувають удома. Коли ж після свят усі виходять на роботу, кількість хворих стрімко зростає внаслідок ослабленості організму та численних контактів людей один з одним. Саме через це на графіках можна бачити різкий стрибок захворюваності на початку року.

Висновки

Аналіз оброблюваних даних встановив, що ці дані можуть бути досить точно описані *кількома членами ряду Фур'є*.

Виконані розрахунки підтверджують припущення про високу точність апроксимації вихідних експериментальних даних рядом Фур'є.

Під час аналізу даних було виділено три фактори захворюваності: *основний тренд, основні гармоніки і сезонні гармоніки*. Обгрунтовано коректність виділення цих факторів, оскільки один фактор незалежний, а два інші мають слабкий зв'язок. Про це свідчать отримані коефіцієнти кореляції: –0,66; 0,1185; –0,1149.

На основі аналізу отриманих даних можна стверджувати: основний тренд є *активністю бактерій і вірусів*, оскільки це випливає з часового інтервалу та вигляду функції основного тренду; основні гармоніки є *сприйнятливістю організму до бактерій і вірусів*; третя компонента – сезонні гармоніки – є *сезонною захворюваністю*.

Апроксимуюча крива є результатом впливу трьох зазначених вище факторів, що були вибрані як основні. Результати розрахунків довготермінового аналізу вихідних даних показали, що кореляція результату розрахунку з вихідними даними становить 99,9 %, а відносна похибка результату апроксимації вихідних даних – 1,09 %.

Подальші дослідження будуть спрямовані на пошук прихованих залежностей між знайденими факторами і пошук нових компонент у вихідних даних.

1. *Дані* про захворюваність на грип з офіційного сайту Міністерства охорони здоров'я України по 10 контрольним містам // МОЗУ. – 2011. – http://moz.gov.ua/ua/portal/mtop_ influenza/ (12.02.2011).
2. *Вапник В.Н., Глазкова Т.Г., Кощеев В.А. и др.* Алгоритмы и программы восстановления зависимостей. – М.: Наука, 1984. – 815 с.
3. *Шашков В.Б.* Прикладной регрессионный анализ. Многофакторная регрессия. – Оренбург: ГОУ ВПО ОГУ, 2003. – 364 с.
4. *Мостеллер Ф., Тьюки Дж.* Анализ данных и регрессия. – М.: Финансы и статистика, 1982. – 240 с.
5. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. – М.: Мир, 1989. – 540 с.

Рекомендована Радою
факультету прикладної математики
НТУУ "КПІ"

Надійшла до редакції
21 лютого 2011 року