УДК 681.3.068

А.Ю. СОКОЛОВ, О.С. РАДИВОНЕНКО, Т.В. КОРЧАК

Национальный аэрокосмический университет им. Н. Е. Жуковского "ХАИ", Украина

МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ ВСПЫШЕК ЭПИДЕМИЙ ИНФЕКЦИОННЫХ ЗАБОЛЕВАНИЙ

Работа посвящена проблеме прогнозирования заболеваемости сальмонеллезом. В статье рассматривается прогнозирование временных рядов статистическими моделями. Обсуждаются методы, применяемые к обработке данных в режиме реального времени. Предлагаются методы экспоненциального сглаживания и адаптивного экспоненциального сглаживания, применяемые к аддитивным и мультипликативным моделям временных рядов. Производится анализ практических результатов, полученных при эксплуатации рассматриваемых моделей.

прогнозирование временных рядов, прогноз, адаптивное экспоненциальное сглаживание

Введение

Эпидемиологический надзор за инфекционными болезнями и предупреждение о вспышках отдельных инфекционных заболеваний является актуальной проблемой. Заболеваемость болезнями пищевого происхождения имеет тенденцию к росту как в развитых, так и в развивающихся странах, эта проблема быстро перерастает в глобальную по своему значению.

Такое развитие событий требует решения проблемы обработки и анализа информации, полученной в процессе деятельности санитарно-эпидемиологичес-кой службы. Это трудоемкая задача, облегчить которую могут современные информационные технологии.

В статье рассматриваются данные, предоставленные санитарно-эпидемиологической службой по заболеваемости сальмонеллезом жителей города Харькова. Данная инфекция по сравнению с другими инфекционными болезнями носит более тяжелый характер. Бактерии рода Salmonella имеют широкий спектр хозяев, могут легко распространяться среди большого числа домашних животных, имеют множество факторов выявления и слежения.

Для решения поставленной задачи рассматриваем её как задачу прогнозирования на основе временных рядов.

В работе [1] Аладьев В.З. и Харитонов В.Н. описывают факторы, влиянию которых подвержен временной ряд и выделяют четыре составляющие:

- тренд, который представляет непериодическое изменение в среднем на временном интервале, на котором определен временной ряд;
- сезонный фактор, определяющий действия, повторяющиеся в единицах дней, недель, месяцев, лет;
- циклические факторы, влияющие на ряд;
- случайный фактор.

В работах [1 – 3] рассматриваются модели временных рядов: аддитивная и мультипликативная, которые представляют зависимость факторов. Зная суть явления, описываемого временным рядом, не составит труда выбрать подходящую модель.

Для решения задач прогнозирования на основе временных рядов обычно используются следующие адаптивные методы: метод экспоненциального сглаживания, методы Хольта и Брауна, метод Винтера. Рассмотренные методы обладают рядом свойств [3]:

- применимость для широкого круга задач;
- прогнозирование базируется на интенсивном анализе информации, содержащейся в отдельных временных рядах;
- не требуется большого объема информации;
- ясность и простота математической формулировки.

Анализ и моделирование изучаемого объекта

Графическое представление и описание поведения временного ряда. Проведем анализ данных, предоставленных санитарно-эпидемиологической службой, по заболеваемости сальмонеллезом населения города Харькова в период с 2003 по 2005 год [4]. Информация представлена в табл. 1.

Таблица 1

Заболеваемость сальмонеллезом за 2003 – 2005 гг.

№	месяц	Заболеваемость	№	месяц	Заболеваемость
1	Январь	19	19	Июль	24
2	Февраль	40	20	Август	38
3	Март	36	21	Сентябрь	23
4	Апрель	28	22	Октябрь	15
5	Май	38	23	Ноябрь	27
6	Июнь	55	24	Декабрь	9
7	Июль	75	25	Январь	33
8	Август	58	26	Февраль	41
9	Сентябрь	56	27	Март	25
10	Октябрь	42	28	Апрель	40
11	Ноябрь	23	29	Май	30
12	Декабрь	25	30	Июнь	35
13	Январь	12	31	Июль	78
14	Февраль	23	32	Август	51
15	Март	8	33	Сентябрь	59
16	Апрель	10	34	Октябрь	34
17	Май	23	35	Ноябрь	18
18	Июнь	31	36	Декабрь	33

По данным табл. 1 построим график заболеваемости сальмонеллезом в 2003 – 2005 гг. (рис. 1), где на оси абсцисс представлены даты наблюдения, на оси ординат – количество заболевших.

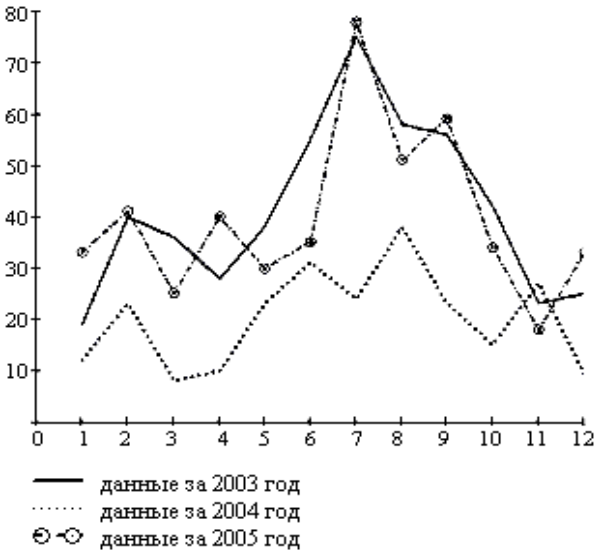

Рис. 1. Динамика заболеваемости сальмонеллезом отдельно по годам

Моделирование тенденции временного ряда и сезонных колебаний. Первой задачей исследования временного ряда является проверка гипотезы о существовании тенденции. Воспользовавшись методом Ф. Форестера и А. Стюарта [5] было установлено наличие тенденции, была выявлена неслучайная составляющая. Визуальный анализ показал, что в качестве неслучайной составляющей в анализируемом временном ряду имеет смысл выделять тренд и сезонную компоненту.

Для моделирования тенденции временного ряда (рис. 2) воспользуемся методом скользящих средних, который относится к механическим способам выравнивания [1, 2, 6]:

$$\frac{1}{5}\cdot\left(U_{n-2}+U_{n-1}+U_n+U_{n+1}+U_{n+2}\right)=$$

$$=\frac{1}{5}\cdot[5]\cdot U_n,$$

где $[5]\cdot U_n$ – оператор, характеризующий процесс суммирования пяти членов ряда;

$\frac{1}{5}\cdot[5]\cdot U_n$ – средняя из пяти членов.

При моделировании сезонных колебаний был выполнен расчет значений сезонной компоненты (рис. 2).

Отказ от анализа циклической компоненты был принят на основании того факта, что, во-первых, временной ряд имеет длину не более 6 лет, в то время как циклические явления проявляются с периодичностью, превышающей длину данного ряда. Во-вторых, при анализе сезонной компоненты можно одновременно учитывать и циклическую компоненту временного ряда.

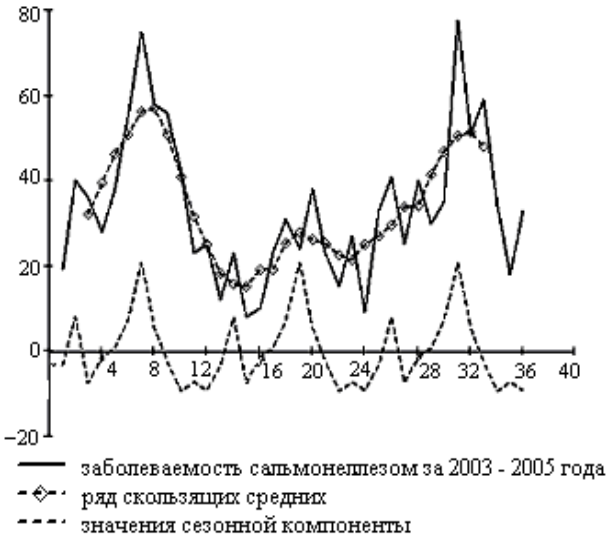

Рис 2. Моделирование тенденции временного ряда и сезонных колебаний

Описание прогнозирующих методов

Метод экспоненциального сглаживания. Особенность экспоненциального сглаживания состоит в том, что в процедуре выравнивания каждого наблюдения используются только значения предыдущих уровней ряда динамики, взятых с определенным весом. Этот метод рекомендуется для краткосрочного прогнозирования, для стационарных данных, или когда в данных есть медленный рост или наоборот, снижение во времени.

Экспоненциальное сглаживание ряда осуществляется по рекуррентной формуле:

$$F_t = \alpha \cdot X_t + \beta \cdot F_{t-1},$$

$$F_0 = \frac{\sum_{i=1}^{m} x_i}{m},$$

где F_t – значение экспоненциальной средней в момент t;

X_t – элемент временного ряда, $X = \{x_1, x_2, ...\}$;

α – параметр сглаживания, $0 < \alpha < 1$;

$\beta = 1 - \alpha$.

Константа α влияет на точность прогноза. Она также должна обеспечивать минимум ошибки прогноза. Если данные имеют существенные колебания или случайность, должно использоваться малое значение для α. С другой стороны, данные с малой случайностью или с четкой моделью, должны использовать большее значение для α.

Полиномиальные модели Брауна. Отличие между однократным экспоненциальным сглаживанием и моделями Брауна заключается в наличии добавочных формул для оценки тренда, таким образом, они могут быть использованы для прогноза нестационарных временных рядов.

Следующие формулы используются в адаптивной полиномиальной модели второго порядка:

$$S'_t = \alpha X_t + (1-\alpha) S'_{t-1};$$

$$S''_t = \alpha S'_t + (1-\alpha) S''_{t-1};$$

$$S'''_t = \alpha S''_t + (1-\alpha) S'''_{t-1},$$

где S'_t, S''_t, S''_t – экспоненциальная средняя первого, второго и третьего прядка.

Начальные условия:

$$S'_0 = a_t - \frac{\beta}{\alpha} b_t + \frac{\beta(2-\alpha)}{2\alpha^2} c_t;$$

$$S''_0 = a_t - \frac{2\beta}{\alpha} b_t + \frac{\beta(3-2\alpha)}{\alpha^2} c_t;$$

$$S'''_0 = a_t - \frac{3\beta}{\alpha} b_t + \frac{3\beta(4-3\alpha)}{2\alpha^2} c_t.$$

Оценка коэффициентов:

$$b_t = \alpha / 2\beta^2[(6-5\alpha)S'_t -$$

$$-(10-8\alpha)S''_t + (4-3\alpha)S'''_t];$$

$$a_t = 3S'_t - 3S''_t + S'''_{t-1};$$

$$c_t = \alpha / \beta^2[S'_t - 2S''_t + S'''_{t-1}].$$

Прогноз на m шагов вперед записывается простым квадратичным полиномом:

$$F_{t+m} = a_t + b_t m + 1/2 \cdot c_t m^2.$$

Адаптивная версия метода Хольта. Трехпараметрическая модель экспоненциального сглаживания Хольта подобна модели Брауна, так как оценивает тренд и использует его в прогнозировании. Уравнения записываются в следующей форме:

$$S_t = \alpha X_t + (1-\alpha)(S_{t-1} + T_{t-1} + R_{t-1}/2);$$

$$T_t = \beta dS_t + (1-\beta) \cdot T_{t-1};$$

$$R_t = \gamma d^2 S_t + (1-\gamma) \cdot R_{t-1},$$

где $dS_t = S_t - S_{t-1}$, $d^2 S_t = dS_t - dS_{t-1}$.

Окончательную формулу для прогноза можно записать как

$$F_{t+m} = S_t + T_t m + 1/2 \cdot R_t m^2,$$

где m – число шагов прогнозирования.

Переменные сглаживания α, β, γ подбираются с учетом минимизации ошибки прогнозирования.

Метод Хольта-Винтера. Проведя несколько модификаций над методом экспоненциального сглаживания, представим модель, учитывающую тенденцию и сезонность. Метод Хольта-Винтера учитывает данные факторы и аналитически записывается как:

$$T_t = C(\overline{X}_t - \overline{X}_{t-1}) + (1-C) \cdot T_{t-1},$$

$$S_t = B\frac{X_t}{\overline{X}_t} + (1-B) \cdot S_{t-L},$$

$$\overline{X}_t = A\frac{X_t}{S_{t-L}} + (1-A) \cdot (\overline{X}_{t-1} - T_{t-1}),$$

где T_t – фактор тренда от времени t;

S_t – фактор сезонности;

L – период сезонного цикла;

$\overline{X}_t$ – сглаженный ряд.

Оптимальные параметры A, B, C предлагается находить экспериментальным путем.

Схема составления прогноза на h шагов в соответствии с методом Хольта-Винтера выглядит следующим образом:

$$\widetilde{X}_n(h) = (\overline{X}_n + hT_n)S_{n-L+h},\ h = 1,2,...,L$$

$$\widetilde{X}_n(h) = (\overline{X}_n + hT_n)S_{n-2L+h},\ h = L+1, L+2,...,2L.$$

Модель Хольта-Винтера предполагает аддитивный тренд и мультипликативные факторы, но может быть модифицирован для работы с мультипликативным трендом и аддитивной сезонной компонентой.

Перед тем, как результаты прогнозирования могут быть получены, должна быть определена точность прогноза. Для этого можно использовать формулу расчета относительной ошибки прогнозных значений

$$\varepsilon_t = \frac{|\bar{x}(t) - x(t)|}{x(t)} \cdot 100\%,$$

где $\bar{x}(t)$ – прогнозное значение;

$x(t)$ – фактическое значение временного ряда.

Результаты численного эксперимента

Описанные алгоритмы были проверены на показателях заболеваемости. Прогноз осуществлялся на период последних трех месяцев 2005 г. На рис. 3 представлена динамика фактических значений показателей заболеваемости и процесс прогнозирования данных показателей с использованием метода Хольта-Винтера для аддитивной и мультипликативной модели. Значения полученных прогнозов приведены в табл. 2.

Ошибка между действительным значением и прогнозом в мультипликативной модели MSE=6,1386, в аддитивной модели MSE=8,96 .

Таблица 2

Оценка качества прогноза

Номер шага	Фактическое значение временного ряда	Прогнозные значения для аддитивной модели Хольта-Винтера	Прогнозные значения для мультипликативной модели Хольта-Винтера
1	34	34,7435	34,4183
2	18	29,0351	34,1294
3	33	12,1533	32,8704
ε_t		8,96	6,41386

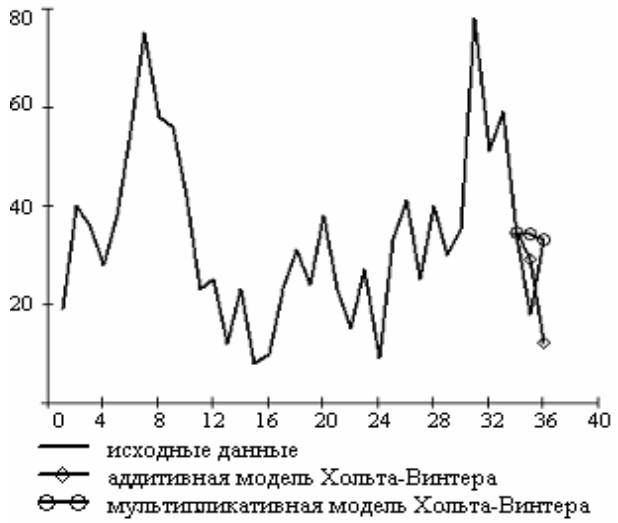

Рис. 3. Прогнозирование вспышек эпидемий методом Хольта-Винтера

Заключение

В данной статье был проведен анализ данных заболеваемости сальмонеллезом в г. Харькове. Рассматривались факторы, влияющие на временной ряд: тенденция, сезонная составляющая. Были рассмотрены методы прогнозирования, позволяющие провести анализ временного ряда.

Полученные результаты, представленные на рис. 3, показывают, что достоверность краткосрочного прогноза адаптивным методом Хольта является лучшей, чем по методу Хольта-Винтера. В этом случае нет необходимости использовать более сложные методы.

Точность методов зависит от параметров сглаживания. Результаты прогнозирования показали, что точность прогнозов выше при малых значениях сглаживающих констант.

Дальнейшими этапами работы являются:

- получение долгосрочного прогноза статистическими методами;
- построение прогноза нейросетевыми и нечеткими методами;
- сравнение результатов, получаемых нейросетевыми, нечеткими и статистическими методами на реальных данных, предоставленных санитарно-эпидемиологической службой;
- реализация информационной системы прогнозирования.

Литература

1. Aladjev V.Z., Haritonov V.N. General theory of statistics. – U.S.: Fultus Corporation, 2004. – 255 с.

2. Кобелев Н.Б. Практика применения экономико-математических методов и моделей. – М.: ЗАО «Финстатинформ», 2000. – 246 с.

3. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. – М.: Финансы и статистика, 2003. – 416 с.

4. Клещар Л.А., Радивоненко О.С., Корчак Т.В. Прогнозирование вспышек эпидемий сальмонеллеза с использованием методов анализа временных рядов // Эпидемиология. Гигиена. – 2006. – С. 209.

5. Ежеманская С.Н. Эконометрика. – Ростов: Феникс, 2003. – 160 с.

Поступила в редакцию 16.03.2007

Рецензент: д-р техн. наук, проф. М.Л. Угрюмов, Национальный аэрокосмический университет им. Н.Е. Жуковского "ХАИ", Харьков.