УДК 62.50

ФИЛЬТРАЦИЯ И ПРОГНОЗИРОВАНИЕ ДАННЫХ ДЕМОГРАФИЧЕСКИХ НАБЛЮДЕНИЙ

В. Г. Рудалев*, А. И. Кремер**

** Воронежский государственный университет*
*** Воронежский филиал Российского государственного социального университета*

Поступила в редакцию 20.04.2014 г.

Аннотация. Исследуются алгоритмы калмановской фильтрации в задачах текущего оценивания и прогнозирования возрастно-половой структуры популяций. Обсуждаются результаты построения специальной демографической модели с целью применения ее в условиях поступления неполных и неточных данных текущих мониторингов.
Ключевые слова: демографическая модель, фильтр Калмана, оценка состояния, сструктура популяций.
Annotation. Kalman filtering algorithms are investigated in the problems of current estimation and forecasting age-sex populations structure. It Is Discussed building to special demographic model for the reason using in condition of the arrival incomplete and inexact given current monitoring the population.
Keywords: demographic model, Kalman filtering, populations structure, state estimation.

ВВЕДЕНИЕ

Актуальными задачами демографической статистики и популяционной биологии является оценка и прогнозирование численности и возрастно-полового состава биологических популяций. Решение подобного рода задач обычно происходит в условиях неполных и неточных выборочных наблюдений. Современной тенденцией здесь является все более широкое применение методов системного анализа и теории управления.

Представляется перспективным применение математических моделей популяций в виде уравнений состояния и базирующихся на них методов теории оптимальной фильтрации Калмана-Бьюси [1]. Фильтры Калмана-Бьюси (ФКБ) специально разработаны для использования в условиях неполных и неточных наблюдений но, в силу ряда причин, чаще всего применяются в технических системах. Применение их в демографических областях сопряжено с рядом проблем, в том числе:

• трудности (или невозможность) проведения активных экспериментов над популяцией для более точного построения моделей;

• большое количество заранее неизвестных возмущающих факторов (экологических, социально экономических и прочих), нестационарность и стохастичность моделей;

• «плавность» изменений численности, затрудняющая решение задачи идентификации модели на ограниченном временном отрезке.

В работах [3, 4] проведено описание демографических систем в виде уравнений состояния и проанализированы возможности применения ФКБ на примере прогнозирования численности населения ряда регионов при наличии относительно точных текущих наблюдений.

Целью данной статьи является исследование эффективности использования ФКБ для решения задач текущего оценивания и прогнозирования численности и возрастно-полового состава демографических систем в условиях неточных наблюдений. С учетом перечисленных особенностей задачи, анализ проводится на смоделированной тестовой популяции, прототипом которой является

популяция мышей-полевок в регионах южного Урала [5].

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПОПУЛЯЦИИ

Для описания демографической модели разобьем популяцию на группы в соответствии с возрастом и полом. Например, если предположить для популяции грызунов, что величина временного интервала, охватывающего одну возрастную группу, равна 2 месяцам, а продолжительность жизни – 16, то популяция состоит из 8 возрастных групп: 0–2 месяца, 2–4, 4–6, …, 14 мес. и старше.

Динамика популяции описывается следующими дискретными уравнениями состояния, являющимися обобщением матричной модели Лэсли [2]:

$$X[k+1] = A[k]X[k] + U[k], \qquad (1)$$

$$Y[k] = CX[k], \qquad (2)$$

$$k = 0,1,2,\ldots$$

$$X[0] = X_0.$$

Здесь $X[k] = \begin{bmatrix} X^F[k] \\ ---- \\ X^M[k] \end{bmatrix}$ – вектор состояния, включающий два вектора – вектор состояния, описывающий динамику женского состава $X^F[k]$ и мужского $X^M[k]$ в зависимости от номера интервала текущего наблюдения k. Вектор $X^F[k]$ и, аналогично ему, $X^M[k]$, имеет вид

$$X^F[k] = \begin{bmatrix} x_1^F[k] \\ ..... \\ x_N^F[k] \end{bmatrix},$$

где N – число возрастных групп. Младшие индексы вектора соответствуют младшим возрастным группам.

Блочно-диагональная матрица $A[k]$ описывает динамику изменений возрастного состава населения. Ее структура зависит от величины временного интервала между $k+1$-м и k-м наблюдениями. Если упомянутый интервал равен продолжительности возрастной группы, то матрицу $A[k]$ можно представить в следующем виде:

$$A[k] = \left[\begin{array}{cccccc|cccccc} \alpha_1 & \alpha_2 & \alpha_3 & \ldots & \alpha_{N-1} & \alpha_N & & & & & & \\ \beta_1 & 0 & 0 & 0 & \ldots & 0 & & & & & & \\ 0 & \beta_2 & 0 & 0 & \ldots & 0 & & & 0 & & & \\ . & . & . & . & . & 0 & & & & & & \\ 0 & 0 & 0 & \ldots & \beta_{N-1} & \beta_N & & & & & & \\ \hline \gamma_1 & \gamma_2 & \gamma_3 & \ldots & \gamma_{N-1} & \gamma_N & 0 & 0 & . & . & \ldots & 0 \\ & & & & & & \delta_1 & 0 & . & . & . & 0 \\ & & & 0 & & & 0 & \delta_2 & . & . & . & 0 \\ & & & & & & . & . & . & . & 0 & 0 \\ & & & & & & 0 & 0 & \ldots & 0 & \delta_{N-1} & \delta_N \end{array}\right].$$

Здесь α_i (соответственно γ_i) – это среднее число родившихся и выживших особей женского пола (соответственно – мужского) на одну самку i-й возрастной группы за интервал наблюдения. Параметры β_i, δ_i (коэффициенты передвижки) равны:

$$\beta_i = 1 - \varepsilon_i^F, \ \delta_i = 1 - \varepsilon_i^M,$$

где ε_i^F, ε_i^M – коэффициенты смертности в i-й возрастной группе.

В векторе входных воздействий

$$U^T[k] = \left[u_1^{FT}[k], \ldots, u_N^{FT}[k], u_1^{MT}[k], \ldots, u_N^{MT}[k]\right]$$

i-й компонент описывает баланс миграции за пределы региона в i-й возрастной группе на k-м интервале наблюдения.

В общем случае вектор $U[k]$ является случайным процессом со средним значением $M\{U[k]\} = U^0$ и ковариационной матрицей

$$M\{(U[k] - U^0)^T (U[k] - U^0)\} = V_u.$$

На практике влияние миграции учитывается с помощью коэффициентов передвижки и U^0 можно положить нулю.

В уравнении (2) $Y[k]$ – вектор наблюдаемых значений структурного состава населения. Вид уравнения наблюдения (2) зависит от способа наблюдения. Если наблюдению и учету подлежит вся совокупность с учетом возрастного состава, т. е. вектор $X[k]$, то C является диагональной матрицей $C = diag[1, 1, ..., 1]$.

Неточность наблюдений моделируется введением в уравнение (2) шума наблюдений $v[k]$:

$$Y[k] = CX[k] + v[k]. \qquad (3)$$

На практике обычно наблюдения носят выборочный характер, поэтому можно пред-

положить, что $v[k]$ есть последовательность нормальных некоррелированных случайных векторов со средним значением $Mv[k]=0$ и ковариационной матрицей V_v.

Аргумент k у перечисленных выше параметров ($\beta_i[k]$, $\gamma_i[k]$ и др.) опущен лишь для получения более компактной записи. В действительности значения параметров матрицы A зависят от текущего момента времени k и меняются во времени случайным образом. Модель (1)–(2) можно классифицировать как нестационарную модель со случайными параметрами [1].

АЛГОРИТМЫ ОЦЕНКИ И ПРОГНОЗА СТРУКТУРНОГО СОСТАВА ПОПУЛЯЦИИ

Модель (1)–(2) является нестационарной и стохастичной, что затрудняет использование многих, ставших классическими подходов.

Непосредственное решение системы разностных уравнений (1)

$$\hat{X}[k+1]=A[k]\hat{X}[k]+U^0[k],$$
$$\hat{X}[0]=X[0], \qquad (4)$$

для расчета прогноза состояния $\hat{X}[k+1]$ на следующие моменты времени в большинстве случаев не приводит к успеху. Например, ошибки в определении начального состояния $X[0]$ (структурного состава популяции на базовый момент времени) или матриц A, U^0 оказывают негативное влияние на точность. Иногда при этом ошибки прогноза состояния могут даже неограниченно возрастать во времени из-за склонности демографических систем к неустойчивому поведению [2]. Но шумы наблюдения на точность здесь не влияют, так как в уравнении (4) используется только базовое значение оценки популяции $X[0]$, а не наблюдаемые значения.

Для прогноза только на один шаг вперед предпочтительна, на первый взгляд, следующая модификация уравнений (4)

$$\hat{X}[k+1]=A[k]Y[k]+U^0[k], \qquad (5)$$

где в качестве «базы» $Y[k]$ используется текущее наблюдение вектора $X[k]$ (в предположении, что вектор $X[k]$ наблюдается полностью). Ошибки в матрицах и в начальном состоянии не окажут существенного влияния, так как на следующем шаге прогноз будет скорректирован. Однако вектор $X[k]$ может наблюдаться с ошибками (см. формулу (3)), а фильтрацию наблюдений метод не осуществляет.

Проблема фильтрации решается с помощью ФКБ. Алгоритм ФКБ позволяет находить оценки структурного состава населения, описываемого вектором состояния $X[k]$, статистически оптимальным образом по критерию минимума среднеквадратической ошибки. Задача оценивания по Калману заключается в определении оценок неизвестных текущих значений вектора состояния по известным данным наблюдения $Y[k]$, а прогнозирование выдает оценки на один шаг вперед.

Основным преимуществом ФКБ является наличие обратной связи по наблюдаемым данным, что при отсутствии шумов наблюдения в (3) гарантирует быструю сходимость оценок к истинным значениям, а при наличии шумов – оптимальный баланс между скоростью сходимости и ошибками фильтрации. Однако значения матриц A, B, C и статистических характеристик шумов (или их оценки) должны быть известны на каждом шаге алгоритма.

Для использования в демографических системах предпочтителен вариант ФКБ, вычисляющий оценку не текущего, а будущего вектора состояния $X[k+1]$ на один шаг вперед [1]. Оценки состояния вырабатываются в реальном времени по мере поступления новых наблюдений $Y[k]$:

$$\hat{X}[k+1]=A\hat{X}[k]+\hat{U}[k]+$$
$$+\Gamma[k]\{Y[k]-C\hat{X}[k]\}, \qquad (6)$$
$$\hat{X}[0]=MX[0].$$

Здесь $\hat{X}[k]$ – искомая оценка вектора состояния, $\hat{U}[k]$ – оценка среднего значения баланса миграции населения, $\Gamma[k]$ – матрица коэффициентов усиления обратной связи фильтра, вычисляемая по следующим формулам

$$\Gamma[k]=A[k]P[k]C\{C^TP[k]C+V_v\}^{-1},$$

$$P[k+1] = A[k]P[k]A^T[k] + V_U - \\ -\Gamma[k]C^T P[k]A^T,$$
$$P[0] = P_0.$$

Здесь P_0 – ковариационная матрица ошибки оценки начального состояния $X[0]$, V_U – ковариационная матрица входных шумов в уравнении (1).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Исследование эффективности работы ФКБ выполнялось с помощью разработанной Windows-программы, моделирующей популяцию.

На первом этапе эксперимента была проанализирована работа ФКБ для прогнозирования структурного состава популяции мышей-полевок при известных показателях рождаемости, миграции и смертности, но при наличии ошибок наблюдения. Прогноз ФКБ сравнивался с «точными» данными, полученными в результате моделирования, и с прогнозом по упрощенному алгоритму (5). Для упрощения анализа U^0 и V_U полагались равным 0, а миграция учитывалась с помощью коэффициентов передвижки β_i, δ_i. Относительное среднеквадратическое отклонение шума наблюдения σ_{vi} нормировалось к численности i-й возрастной группы популяции. Матрица ковариаций шума, используемая в ФКБ, составлялась из оценочных значений дисперсий:

$$V_v = diag\left\{\sigma_{v1}^2\hat{x}_1^2, ..., \sigma_{v2N}^2\hat{x}_{2N}^2\right\}.$$

Начальное значение ковариационной матрицы ошибки $P_v[0] = V_v$.

На рис. 1 представлена оценка распределения самцов по возрастам на последнем шаге наблюдений для случая одинаковых дисперсий $\sigma_{vi} = \sigma_v$. Прогноз с фильтрацией (ФКБ) более точен, чем упрощенный прогноз (5).

Более наглядно процесс фильтрации для полной численности популяции (без учета полов и возрастов) представлен на рис. 2. Как следует из формулы (5), неотфильтрованный прогноз популяции точно повторяет зашумленные результаты наблюдений, а отфильтрованный – сходится к истинному значению. Так как в данном случае все параметры модели известны точно, а входные шумы отсут-

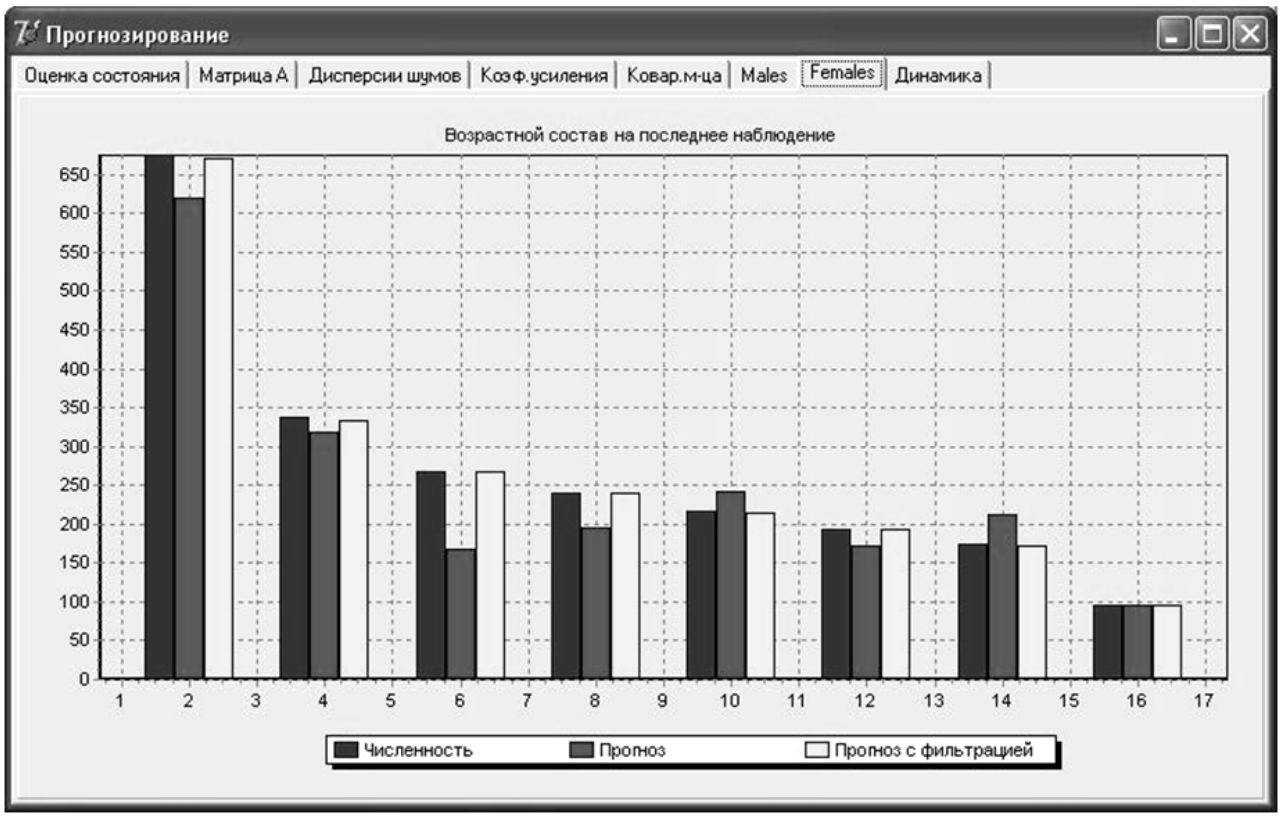

Рис. 1

ствуют ($V_U = 0$), то в силу известных свойств ФКБ дисперсия ошибок оценок стремится к нулю при росте числа наблюдений.

На втором этапе эксперимента коэффициенты рождаемости и передвижки в уравнениях ФКБ отличались от истинных значений, используемых в уравнении (1). Относительная ошибка в определении коэффициентов рождаемости составляла 50 %, а коэффициенты передвижки оценивались методом наименьших квадратов по наблюдениям возрастов популяции. Ошибка наблюдения полагалась равной нулю. Таким образом, задача сводилась только к прогнозу на один шаг вперед. (Данная упрощенная постановка может быть применима для изучения демографии населения при наличии относительно точной системы государственного учета.) Здесь алгоритмы прогноза (5) и ФКБ дают одинаковую точность. Большие ошибки прогноза (до 50 %) были отмечены в младшей возрастной группе и обусловлены ошибкой в определении коэффициентов рождаемости. При отсутствии возмущающих воздействий и ошибок наблюдения ковариационные матрицы входных и выходных шумов играют роль весовых матриц, выбором соотношения которых задаются динамические характеристики ФКБ. В данном примере для ускорения сходимости это соотношение было выбрано большим: $\|V_U\| / \|V_v\| = 10$.

Задачей *третьего этапа* было исследование алгоритмов при наличии как шумов наблюдения, так и случайных изменений параметров. Рассматривался случай «быстрых» изменений коэффициентов рождаемости, описываемых некоррелированными во времени нормальными случайными процессами $s_i[k] \sim N[\overline{s}_i, \sigma_{si}^2]$, где $i = 1, 2, \ldots, 2N$, $k = 0, 1, 2, \ldots$. Коэффициенты передвижки считаются квазипостоянными.

В данной, более реалистичной, ситуации предлагается следующая субоптимальная модификация алгоритма фильтрации. Оценки коэффициентов передвижки вычисляются по наблюдаемым значениям возрастных групп с помощью рекуррентного метода наименьших квадратов. В качестве оценок коэффициентов рождаемости используются известные средние значения, характерные для данного биологического вида. Ошибки, связанные с коэффициентами рождаемости, моделируются с помощью ковариационной матрицы вход-

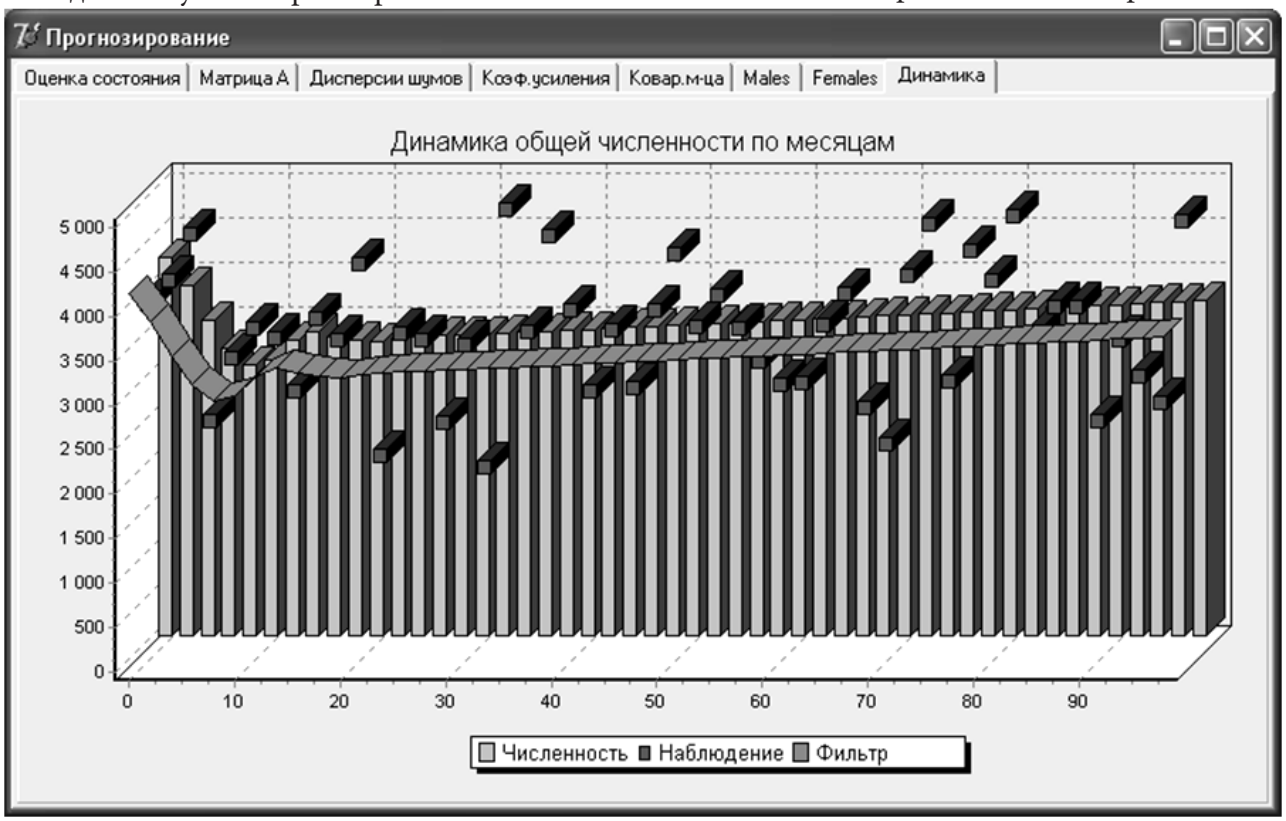

Рис. 2

ного шума. Для младшей возрастной группы самок дисперсия этих ошибок на k -й момент измерения (1-й диагональный элемент матрицы V_U) находится из уравнений (1) как

$$\sigma_F^2[k] = \sum_{i=1}^{N} (x_i^F[k]\bar{\alpha}_i \sigma_{si})^2, \qquad (7)$$

А для младшей возрастной группы самцов ($N+1$ -й диагональный элемент матрицы V_U)

$$\sigma_M^2[k] = \sum_{i=1}^{N} (x_i^M[k]\bar{\gamma}_{\bar{i}} \sigma_{si})^2. \qquad (8)$$

где $\bar{\alpha}_i$, $\bar{\gamma}_{\bar{i}}$ – средние значения коэффициентов рождаемости.

Остальные элементы матрицы V_U в данном примере можно положить нулю. В качестве значений численности возрастных групп используются их оценки, доступные на текущем шаге.

В качестве показателя эффективности использовалась относительная (к дисперсии шума) среднеквадратическая ошибка оценки полной численности популяции

$$\varepsilon_{отн}^2 = \left(\sum_{i=1}^{2N} X_i - \sum_{i=1}^{2N} \hat{X}_i\right)^2 / \left(\sum_{i=1}^{2N} X_i\right)^2 / \left(\left(\sum_{i=1}^{2N} \sigma_{vi}^2 X_i^2\right) / \left(\sum_{i=1}^{2N} X_i\right)\right)^2 = \qquad (9)$$

$$= \left(\sum_{i=1}^{2N} X_i - \sum_{i=1}^{2N} \hat{X}_i\right)^2 / \left(\sum_{i=1}^{2N} \sigma_{vi}^2 X_i^2\right),$$

дополнительно усредняемая по выборке и совокупности реализаций (аргумент k в (9) опущен).

Работоспособность алгоритма иллюстрируется рис. 3, полученным для нормированных среднеквадратических отклонений $\sigma_s = 0,5$ и $\sigma_v = 0,7$ (предполагается, что дисперсии для всех компонентов векторов одинаковы). Как видно из рис. 3, ФКБ отслеживает ту часть наблюдаемой динамики популяции (так называемые «волны»), которая обусловлена изменениями ее параметров, фильтруя ошибки наблюдения. Непосредственно ФКБ дает прогноз лишь на один временной шаг, равный длине возрастной группы. Однако отфильтрованная временная последовательность численности возрастных групп более удобна для дальнейшего использования классических методов регрессионного анализа и прогнозирования, чем исходные «неочищенные» от ошибок данные.

В табл. 1 приведены сводные значения усредненного показателя $\varepsilon_{отн}$ при различных сочетаниях СКО шумов наблюдений и отклонений параметров. Используется модифицированный с учетом изменений параметров по

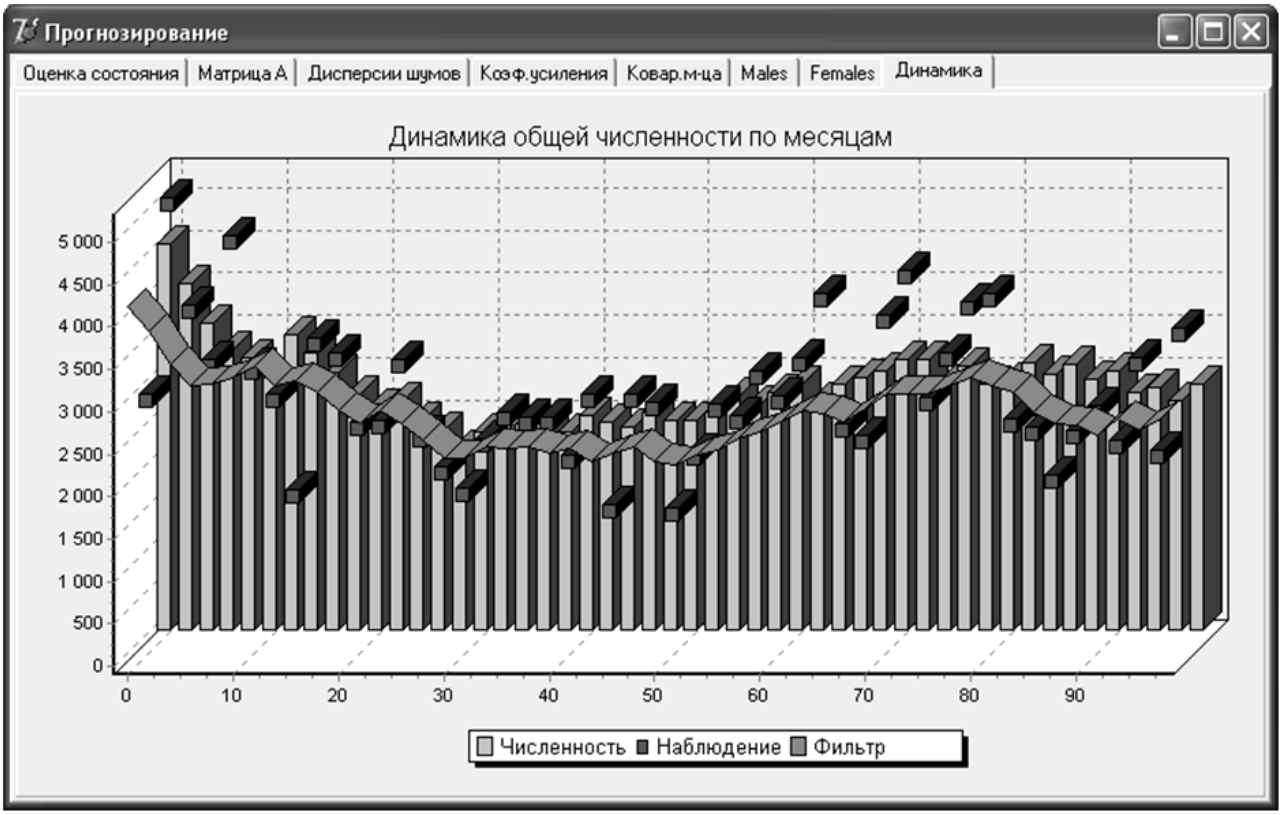

Рис. 3

формулам (7) и (8) алгоритм. Наибольший эффект ($\varepsilon_{отн} < 1$) достигаются, когда влияние ошибок наблюдения выше, чем отклонений параметров. Например, при $\sigma_v = 0,7$ и $\sigma_s = 0,3$ СКО при использовании модифицированного ФКБ уменьшается в 3 раза. В противном случае (см. первую колонку таблицы) эффективность ФКБ незначительна или отсутствует ($\varepsilon_{отн} \geq 1$), что объясняется отсутствием оценки параметров.

Таблица 1

СКО ошибки фильтрации $\varepsilon_{отн}$ при различных σ_s и σ_v

СКО ошибки рождаемости σ_s	СКО шума наблюдений σ_v			
	0,1	0,3	0,5	0,7
0,1	0,55	0,3	0,25	0,2
0,3	0,99	0,55	0,43	0,3
0,5	1,6	0,7	0,55	0,44
0,7	2,2	0,91	0,63	0,55

Отметим, для сравнения, что обычный алгоритм ФКБ (при $V_U = 0$) отслеживает только медленную составляющую в динамике популяции, что приводит к значительному ухудшению показателей. Положительный эффект $\varepsilon_{отн} < 1$ достигается лишь при малых отклонениях коэффициентов рождаемости $\sigma_s \leq 0,1$. Также заведомо неэффективен метод (5), не осуществляющий фильтрацию наблюдений.

ЗАКЛЮЧЕНИЕ

Результаты исследования алгоритмов оценки и прогноза структуры популяции при различных предположениях о характере динамики популяции и ошибках наблюдения свидетельствуют о перспективности применения моделей в форме уравнений состояния и базирующихся на них методов ФКБ. Предложенная методика фильтрации пригодна в условиях случайных изменений параметров популяции и может служить основой для построения более эффективных алгоритмов прогнозирования. Показано, что наибольший выигрыш в точности достигается, когда относительные погрешности наблюдений структурного состава популяции больше чем относительные погрешности в определении коэффициентов рождаемости.

СПИСОК ЛИТЕРАТУРЫ

1. *Синицын И.Н.* Фильтры Калмана и Пугачева / И.Н. Синицын. – М. : Логос, 2006. – 640 с.

2. *Keyfitz M.* Applied Mathematical Demography / M. Keyfitz, H. Caswell. –Berlin: Springer-Verlag, 2005. – 558 p.

3. Рудалев В.Г. Математическая модель динамики структурного состава населения / В.Г. Рудалев, А.И. Кремер // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – Воронеж, 2007. – № 1. – С. 109–113.

4. *Рудалев В.Г.* Анализ влияния ошибок идентификации в демографическом прогнозировании / В.Г. Рудалев, А.И. Кремер // Моделирование и управление в сложных системах: сб. научн. трудов Воронеж, 2010. – С. 14–20.

5. *Кшнясев И.А., Давыдова Ю.А.* Динамика плотности и структуры популяций лесных полёвок в южной тайге // Вестник Нижегородского ун-та им. Н.И. Лобачевского. Серия Биология. – Н. Новгород, 2005. – Вып. 1 (9). – С. 113–124.

Рудалев Валерий Геннадьевич – к.ф.-м.н., доцент кафедры технической кибернетики и автоматического регулирования ВГУ.
Тел.: 8-920-427-57-06. E-Mail: rud_wl@mail.ru

Кремер Александр Ильич – к.т.н., доцент каф. естественнонаучных дисциплин Воронежского филиала Российского государственного социального университета

Rudalev V.G. – Candidate of Physics-math. Sciences, Associate professor, the dept. of the Technical Cybernetics and Automation Control, VSU.
Phone: 8-920-427-57-06. E-Mail: rud_wl@mail.ru

Kremer A.I. – Candidate of tecnical.Sciences, Associate professor, Russian State Social University